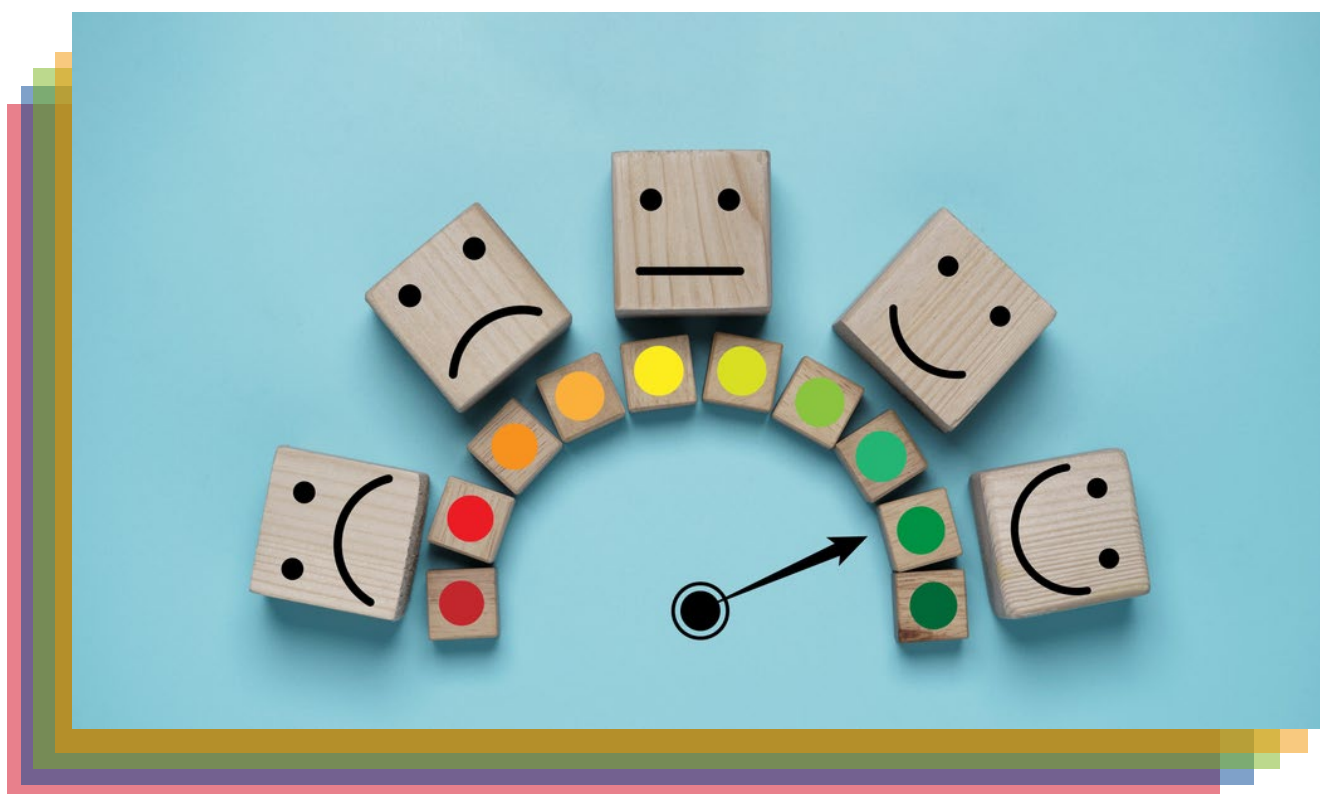


KCE PROCESS NOTE: RESEARCH METHODS FOR THE EVALUATION OF HEALTH SERVICES & POLICY INITIATIVES

FIRST EDITION



HELENA VAN DEYNSE, WENS CHRISTIAENS, JUSTIEN CORNELIS, KOENRAAD D'HOLLANDER,
STEPHAN DEVRIESE, PASCALE JONCKHEER, ROOS LEROY, CARINE VAN DE VOORDE, LEEN
VERLEYE, BÉNÉDICTE VOS, KOEN VAN DEN HEEDE



COLOPHON

Title:	KCE process note: Research methods for the evaluation of Health services & Policy initiatives
Authors:	Helena Van Deynse (KCE), Wens Christiaens (KCE), Justien Cornelis (KCE), Koenraad D'Hollander (KCE), Stephan Devriese (KCE), Pascale Jonckheer (KCE), Roos Leroy (KCE), Carine Van de Voorde (KCE), Leen Verleye (KCE), Bénédicte Vos (KCE), Koen Van den Heede (KCE)
Programme manager:	Imgard Vinck (KCE)
Scientific director:	Roos Leroy (KCE)
External experts:	Sibyl Anthierens (Universiteit Antwerpen), Wilfried Cools (MethoStats), Mieke Deschodt (KU Leuven), Randy Mellaerts (Remedi), Erik Schokkaert (KU Leuven), Walter Sermeus (KU Leuven), Theo Van Achterberg (KU Leuven), Steve Van den Bulck (KU Leuven, UHasselt)
Acknowledgements:	Irina Cleemput (KCE), Jens Detollenaere (KCE), Kayleigh De Meulemeester (KCE), Dorine Lenoir (KCE)
Reported interests KCE members:	KCE is a federal institution funded by RIZIV / INAMI, by the FOD Volksgezondheid, Veiligheid van de Voedselketen en Leefmilieu / SPF Santé publique, Sécurité de la Chaîne alimentaire et Environnement, and by the FOD Sociale Zekerheid / SPF Sécurité Sociale. KCE's mission is to advise policymakers on decisions relating to health care and health insurance on the basis of scientific and objective research. It is expected to identify and shed light on the best possible solutions, in the context of an accessible, high-quality health care system with due regard for growing demand and budgetary constraints. KCE has no interest in companies (commercial or non-commercial, i.e. hospitals and universities), associations (e.g. professional associations, unions), individuals or organisations (e.g. lobby groups) that could be positively or negatively affected (financially or in any other way) by the implementation of the recommendations. All experts involved in the writing of the report or the peer-review process completed a declaration of interest form. Information on potential conflicts of interest is published in the colophon of this report. All members of the KCE Expert Team make yearly declarations of interest and further details of these are available upon request.
Reported interests:	All experts and stakeholders consulted within this report were selected because of their involvement in the topic. Therefore, by definition, each of them might have an interest in the main topic of this report. An overview of all declarations of interests is available in Supplement 1 of this report.
Layout:	Ine Verhulst
Disclaimer:	• Several external experts (cf. supra) were consulted about a (preliminary) version of the scientific report. They were selected based on their expertise in the topic of this report; potential conflicts of interest were considered in this process. Their comments were discussed during meetings and/or via email. They did not co-author the scientific report and do not necessarily agree with its content. • This report has been approved by common assent by the Executive Board. Any (potential) persisting errors or omissions are the sole responsibility of KCE.
Publication date:	16 October 2026
Project number:	2025-70
Domain:	METHOD
MeSH:	Health Services Research; Research Design; Program Evaluation; Process Assessment, Health Care



NLM Classification:	W 84.3
Language:	English
Format:	Adobe® PDF™ (A4)
Legal depot:	D/2026/10.273/56
ISSN:	2466-6459
AI disclaimer:	This work was primarily human-created. AI was used to make stylistic edits, such as changes to structure, wording, and clarity. AI was prompted for its contributions, or AI assistance was enabled. AI-generated content was reviewed and approved. The following model(s) or application(s) were used: co-pilot.
Copyright:	KCE reports are published under a “by/nc/nd” Creative Commons Licence http://kce.fgov.be/content/about-copyrights-for-kce-publications. 
How to refer to this document?	H. Van Deynse, W. Christiaens, J. Cornelis, K. D'Hollander, S. Devriese, P. Jonckheer, R. Leroy, C. Van de Voorde, L. Verleye, B. Vos, K. Van den Heede. KCE process note: Research methods for the evaluation of Health services & Policy initiatives. Method. Brussels: Belgian Health Care Knowledge Centre (KCE). 2026. KCE Reports 426. https://doi.org/10.57598/R426C. This document is available on the website of the Belgian Health Care Knowledge Centre.



■ TABLE OF CONTENTS

LIST OF FIGURES	7
LIST OF TABLES	8
LIST OF ABBREVIATIONS.....	10
■ SCIENTIFIC REPORT	15
INTRODUCTION.....	15
1 PROGRAMME THEORY AND LOGIC MODEL	20
1.1 INTRODUCTION	20
1.2 GENERAL CONCEPTS EXPLAINED	21
1.3 ELEMENTS OF PROGRAMME THEORY AND LOGIC MODELS	23
1.4 DEVELOPMENT PROCESS	24
1.4.1 Preparation	25
1.4.2 Modelling phase	27
1.5 LIMITATIONS OF PROGRAMME THEORY AND LOGIC MODELS.....	29
2 EXPERIMENTAL STUDY DESIGNS	30
2.1 QUANTITATIVE CONSIDERATIONS FOR STUDY DESIGN	31
2.2 ROLE OF EXPERIMENTAL STUDY DESIGN	31
2.3 VALIDITY AND STATISTICAL PRECISION IN EXPERIMENTAL STUDIES	32
2.4 EXPLANATORY VS PRAGMATIC STUDIES	32
2.5 SOURCES OF BIAS	34
2.5.1 Selection and allocation bias.....	34
2.5.2 Performance and detection bias.....	34
2.5.3 Follow-up frequency bias	35
2.5.4 Attrition bias.....	35
2.5.5 Operational bias	35
2.5.6 Reporting bias	35
2.6 RANDOMISATION TECHNIQUES	36
2.6.1 Simple randomisation.....	37
2.6.2 Permuted block randomisation with stratification	37
2.6.3 Minimisation (dynamic allocation)	38
2.7 EXPERIMENTAL STUDY DESIGNS.....	39
2.7.1 Parallel randomised controlled trials	39
2.7.2 Factorial design	40
2.7.3 Crossover design.....	41
2.7.4 Cluster randomised trials.....	42



2.7.5	Stepped-wedge designs	44
2.7.6	Other study designs	45
3	QUASI-EXPERIMENTAL DESIGNS.....	47
3.1	INTRODUCTION TO QUASI-EXPERIMENTAL DESIGNS	47
3.1.1	From experimental to quasi-experimental designs	47
3.1.2	Outline of this chapter	49
3.2	DESIGNS FOR COMPARISON-BASED EFFECT IDENTIFICATION	50
3.2.1	Common types of bias.....	50
3.2.2	Basic designs	51
3.2.3	Strengthening internal validity	57
3.3	DESIGNS FOR BEFORE-AND-AFTER EFFECT IDENTIFICATION	59
3.3.1	Common types of bias.....	59
3.3.2	General designs for simple before-and-after effect identification.....	60
3.3.3	Strengthening internal validity for simple before-and-after effect identification .	62
3.3.4	General designs for longitudinal time-structure-based effect identification	63
3.3.5	Strengthening internal validity for longitudinal time-structure-based effect identification.....	65
3.4	CAUSALITY IN QUASI-EXPERIMENTS.....	68
3.4.1	From association to causation.....	68
3.4.2	The causal inference framework	69
3.4.3	Directed acyclic graphs	71
3.5	ANALYTICAL METHODS FOR CAUSAL INFERENCE.....	73
3.5.1	Propensity score methods.....	73
3.5.2	G-computation	76
3.5.3	Doubly robust estimators.....	78
3.5.4	Instrumental variables	79
3.6	A PRACTICAL GUIDE TO DEVELOPING A QUASI-EXPERIMENTAL DESIGN AND ANALYSIS PROTOCOL	81
3.6.1	Component 1: Research question and estimand	82
3.6.2	Component 2: Make causal assumptions explicit	82
3.6.3	Component 3: Data availability and structure	83
3.6.4	Component 4: Evaluation design choice	84
3.6.5	Component 5: Study protocol specification	85
4	STATISTICAL CONSIDERATIONS FOR EVALUATION STUDIES.....	88
4.1	DEFINING THE RESEARCH QUESTIONS	88
4.1.1	Outcomes	88



4.1.2	Hypotheses.....	91
4.2	TYPE II ERROR: POWER ANALYSIS	93
4.2.1	Statistical power	93
4.2.2	What do we need for power estimation?	94
4.2.3	Factors affecting sample size calculation.....	96
4.3	MANAGING TYPE I ERROR: MULTIPLICITY	98
4.3.1	Multiple testing.....	99
4.3.2	Interim analysis.....	101
4.4	CLUSTERED AND LONGITUDINAL DATA	102
4.5	DEFINING THE TREATMENT EFFECT OF INTEREST	105
4.5.1	Intention-to-treat and per-protocol analyses	105
4.5.2	Intercurrent events	106
4.6	EVALUATING THE ROBUSTNESS OF CONCLUSIONS	107
4.7	GUIDANCE WITH RESPECT TO PROTOCOL, STATISTICAL ANALYSIS PLAN AND REPORTING.....	108
5	PROCESS EVALUATION	109
5.1	INTRODUCTION	109
5.2	WHAT IS PROCESS EVALUATION?	109
5.2.1	Process evaluation: beyond effectiveness evaluation	109
5.2.2	Other roles of process evaluations.....	110
5.3	THREE MAIN DIMENSIONS: IMPLEMENTATION, MECHANISMS OF IMPACT, CONTEXT.....	110
5.3.1	Implementation	111
5.3.2	Mechanisms of impact.....	111
5.3.3	Context	112
5.3.4	When to do a process evaluation?	112
5.4	THEORIES, MODELS AND FRAMEWORKS (TMFS)	112
5.4.1	Why use TMFs?	112
5.4.2	Main characteristics of a selection of TMFs	113
5.4.3	Focus on the mechanisms of impact and outcome: the 'outcome- and system-oriented' TMFs.....	114
5.4.4	Focus on the context: 'context and system complexity' TMFs.....	116
5.4.5	Focus on implementation and context: 'determinant' TMFs	117
5.4.6	Focus on implementation: 'process' TMFs.....	118
5.4.7	Focus on mechanisms of impact: 'mechanism-oriented' TMFs	119
5.5	DATA COLLECTION METHODS IN PROCESS EVALUATIONS	119
5.6	CONDUCT OF A PROCESS EVALUATION	119



6	REALIST EVALUATION	121
6.1	INTRODUCTION	121
6.2	WHAT IS REALIST EVALUATION?	122
6.2.1	The core conceptual framework: Context-Mechanisms-Outcome Configurations	122
6.2.2	Applying realist evaluation principles across the MRC framework	124
6.3	HOW TO USE REALIST EVALUATION: THE REALIST EVALUATION RESEARCH CYCLE	125
6.3.1	Develop programme theory	127
6.3.2	Translating realist programme theory into CMOC hypotheses	128
6.3.3	Design the evaluation protocol	129
6.3.4	Purposive sampling	130
6.3.5	Data collection	130
6.3.6	Realist data analysis – Retroduction	132
6.3.7	Synthesis	132
6.3.8	Output of the realist evaluation	133
7	EVALUATION OF ECONOMIC IMPACT	134
7.1	INTRODUCTION	135
7.2	ECONOMIC EVALUATION WHEN EFFECTIVENESS CAN BE ASSESSED	137
7.2.1	Economic evaluation with non-health outcomes	137
7.2.2	Key principles of economic evaluation	139
7.2.3	Economic evaluation of complex interventions	141
7.3	BUDGET IMPACT ANALYSIS	141
7.4	COSTING	143
7.4.1	Purpose and role in economic analysis	144
7.4.2	Perspective, identification and measurement of resources	145
7.4.3	Valuation of resources	146
7.4.4	Implementation and delivery costs	147
7.4.5	Costing approaches	147
7.4.6	Types of costs	151
8	IMPLEMENTATION RESEARCH AND THE KCE EVALUATION PROGRAMME	153
8.1	INTRODUCTION TO IMPLEMENTATION STRATEGIES AND IMPLEMENTATION RESEARCH	154
8.1.1	Implementation strategies	154
8.1.2	The intervention or innovation	154
8.1.3	Implementation research	154
8.2	STAGES IN IMPLEMENTATION RESEARCH	155



8.2.1	'As is situation'	156
8.2.2	Identify facilitators and barriers	156
8.2.3	Development of an implementation strategy	158
8.2.4	Evaluation	160
8.3	CONCLUDING REMARKS REGARDING THE USE OF IMPLEMENTATION RESEARCH IN THE KCE EVALUATION PROGRAMME	161
9	QUALITY INDICATORS	162
9.1	INTRODUCTION	162
9.2	FROM QUALITY GOALS TO QUALITY MEASUREMENTS	164
9.2.1	Purposes of quality assessment	164
9.2.2	Advancing from quality standards to quality measurement	164
9.2.3	Components of quality indicators	165
9.3	TYPES OF QUALITY INDICATORS	166
9.3.1	Structure, process and outcome indicators	166
9.3.2	Strengths and weaknesses of structure, process and outcome indicators	167
9.4	OPERATIONALISING THE THEORETICAL CONCEPT OF QUALITY INTO A SET OF QUALITY INDICATORS	168
9.4.1	Development of quality indicators	168
9.4.2	Data for quality measurement and improvement	171
9.4.3	Quality requirements for QIs	174
9.4.4	Recommended set of indicators	174
9.4.5	Composite indicators	175
9.4.6	Sources of (validated) quality indicators	175
9.5	ADAPTATION AND RETIREMENT OF QIS	176
9.6	GRADE INITIATIVE ON THE INTEGRATION OF GUIDELINE AND QI DEVELOPMENT	176
10	SYSTEMS THINKING AND COMPLEXITY SCIENCE (STCS) INFORMED EVALUATION	177
10.1	INTRODUCTION	177
10.2	SYSTEMS THINKING AS A WAY TO APPROACH COMPLEXITY	177
10.3	THE ADDED VALUE OF SYSTEMS THINKING	178
10.3.1	Systems thinking as a way of looking	178
10.3.2	Systems thinking as a diagnostic tool	178
10.4	UNPACKING SYSTEMS THINKING AND COMPLEXITY SCIENCE	179
10.4.1	What is complexity?	179
10.4.2	What is a system?	179
10.4.3	Premises of systems thinking and complexity science	180



10.5	APPLYING SYSTEMS THINKING AND COMPLEXITY SCIENCE TO EVALUATION RESEARCH.....	181
10.5.1	How to recognise whether systems thinking is the best fit for evaluating an intervention?	181
10.5.2	How to apply systems thinking to evaluate (complex) interventions?	184
10.5.3	STCS-informed methodology	188
10.6	A BRIEF INTRODUCTION OF SYSTEMS THINKING APPROACHES	188
10.6.1	System dynamics	188
10.6.2	Viable System Model (VSM)	191
10.6.3	Soft systems methodology (SSM)	192
11	DATA REQUIREMENTS AND SOURCES	195
11.1	ROLE OF DATA IN THE EVALUATION PROGRAMME	196
11.2	DATA SOURCES	197
11.2.1	Administrative databases	198
11.2.2	Registries.....	204
11.2.3	Surveys.....	205
11.2.4	Other data sources	207
11.3	OVERVIEW OF RELEVANT DATA.....	207
11.3.1	Data for characterising the target population	207
11.3.2	Data for costing and tariffs data	210
11.3.3	Administrative data and registries for process evaluation and quality measurement.....	211
11.4	DEFINING STUDY VARIABLES FROM AVAILABLE DATA	212
11.4.1	Illustrative applications	212
11.4.2	Good-practice principles.....	214
11.5	PROCEDURES FOR DATA ACCESS	215
■	APPENDICES.....	218
■	REFERENCES	235



LIST OF FIGURES

Figure 1 – Evaluation Framework Complex Interventions	17
Figure 2 – Example of a simple logic model from left to right	22
Figure 3 – Preparing for modelling and modelling process	27
Figure 4 – PRECIS-2 tool	33
Figure 5 – Randomised controlled trial	39
Figure 6 – Factorial design	40
Figure 7 – Crossover design	41
Figure 8 – Cluster randomised trials	42
Figure 9 – Stepped-wedge design	44
Figure 10 – SMART design	46
Figure 11 – Illustration of the controlled post-test-only design	51
Figure 12 – Illustration of the controlled pre-post design	52
Figure 13 – Illustration of a difference-in-difference analysis	53
Figure 14 – Illustration of the regression discontinuity design	55
Figure 15 – Illustration of a sharp (left panel) and a fuzzy (right panel) cut-off	56
Figure 16 – Illustration of the single group pre-post design	60
Figure 17 – Illustration of before-after comparisons using repeated cross-sectional data	61
Figure 18 – Illustration of slope changes (left panel) and intercept changes (right panel)	63
Figure 19 – Illustration of the multiple baselines design	64
Figure 20 – Illustration of the Controlled Interrupted Time Series design	66
Figure 21 – Illustration of synthetic controls	66
Figure 22 – Propensity score matching and propensity score trimming	75
Figure 23 – The criteria for an instrumental variable	80
Figure 24 – Outline of the five components of quasi-experimental study design	81
Figure 25 – Decision flow chart to assist design choice	85
Figure 26 – Superiority, non-inferiority and equivalence studies	92
Figure 27 – Probability of at least one type I error increasing with the number of conducted tests	99
Figure 28 – Illustration of four different adjustment methods for multiple testing ($\alpha = 0.05$)	100
Figure 29 – Intraclass correlation coefficient (ICC)	103
Figure 30 – Illustration of cluster response to intervention	104
Figure 31 – Key functions of process evaluations and their interrelations	111
Figure 32 – The realist evaluation research cycle	126
Figure 33 – Stages of implementation change	155
Figure 34 – MRC model seen through an STCS lens	183



Figure 35 – How to integrate STCS in evaluation	185
Figure 36 – Example of a feedback loop	189
Figure 37 – System dynamics model for “Fixes that backfire” – example	190
Figure 38 – Schematic representation of the Viable System Model	191
Figure 39 – The SSM's learning cycle	193
Figure 40 – Structure of the information available to KCE in the TCT data.....	202
Figure 41 – Data processing life cycle	216

LIST OF TABLES

Table 1 – Examples of indicators	28
Table 2 – Properties of pragmatic vs explanatory studies	32
Table 3 – Sources of bias	35
Table 4 – Randomisation techniques: key points	37
Table 5 – Permuted block randomisation with stratification.....	38
Table 6 – Example of minimisation	39
Table 7 – Strengths and challenges of RCTs	40
Table 8 – Strengths and challenges of factorial designs	41
Table 9 – Strengths and challenges of crossover designs	42
Table 10 – Strengths and challenges of cluster randomised trials	44
Table 11 – Strengths and challenges of stepped-wedge trials	45
Table 12 – Indications and contra-indications to the use of the controlled post-test-only design ..	52
Table 13 – Indications and contra-indications to the use of the controlled pre-post design.....	54
Table 14 – Indications and contra-indications to the use of the regression discontinuity design ...	57
Table 15 – Overview of key design choices for individual matching.....	58
Table 16 – Indications and contra-indications to the use of the single group pre-post design.....	61
Table 17 – Indications and contra-indications to the use of repeated cross-sectional data	62
Table 18 – Indications and contra-indications to the use of the interrupted time series design	64
Table 19 – Indications and contra-indications to the use of the multiple baselines design	65
Table 20 – An example of potential outcomes	70
Table 21 – Most common relations and their adjustments	72
Table 22 – Overview of data structures and their potential applications	83
Table 23 – An indicative ranking of quasi-experimental designs.....	84
Table 24 – Types of outcomes.....	89
Table 25 – Two types of error	91



Table 26 – Additional required parameters for sample size calculation by outcome type.....	96
Table 27 – Main characteristics of a selection of TMFs.....	113
Table 28 – RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance.....	115
Table 29 – CICI: Context and Implementation of Complex Interventions model.....	116
Table 30 – CFIR: Consolidated Framework for Implementation Research	117
Table 31 – An example of a CMOC matrix	132
Table 32 – Framework to classify costing methods.....	148
Table 33 – Classification of types of costs in the Belgian guidelines	152
Table 34 – Examples of linking implementation strategies to determinants.....	158
Table 35 – Rephrasing recommendations into QIs.....	170
Table 36 – International databases and sources of existing QIs	176
Table 37 – Degrees of complexity.....	179



LIST OF ABBREVIATIONS

ABBREVIATION	DEFINITION
ABC	Activity-based costing
ABC	Achievable Benchmark of Care
ACSQHC	Australian Commission on Safety and Quality in Health Care
ADH – HJA	Anonymous Day-care Stays (<i>Anonieme DagHospitalisaties / Hospitalisations de jour anonymes</i>)
ADL	Activities of daily living
AGREE	Appraisal of Guidelines Research and Evaluation in Europe
AHRQ	Agency for Healthcare Research and Quality
AIRE	Appraisal of Indicators through Research and Evaluation
APR-DRG	All Patient Refined Diagnosis Related Group
AT	As-treated
ATC	Anatomical Therapeutic Chemical
ATE	Average treatment effect
ATT	Average treatment effect among the treated
AZV – SHA	Anonymous Hospital Stays (<i>Anonieme Ziekenhuisverblijven / Séjours hospitaliers anonymes</i>)
BANANA	Basel Approach for coNtextual ANALysis
BC	Breast cancer
BCR	Belgian Cancer Registry
BCS	Breast conservative surgery
BelRAI	Belgian Resident Assessment Instrument
BFM	Budget of Financial Means
BIA	Budget impact analysis
BMI	Body Mass Index
CAC	Cluster autocorrelation
CAS	Complex adaptive system
CATWOE	Customer, Actor, Transformation, Worldview, Owner, and Environment
CBA	Cost-benefit analysis
CCA	Cost-consequence analysis
CEA	Cost-effectiveness analysis
CFIR	Consolidated Framework for Implementation Research
CFIR-ERIC	Consolidated Framework for Implementation Research – Expert Recommendations for Implementing Change
CICI	Context and Implementation of Complex Interventions

ABBREVIATION	DEFINITION
CIMO	Context-Intervention-Mechanism-Outcome
CMA	Cost-minimisation analysis
CMO	Context-Mechanism-Outcome
CMOC	Context-Mechanism-Outcome Configuration
CNK	National Code (<i>Nationaal Kodenummer / Code National</i>)
COICOP	Classification of Individual Consumption According to Purpose
CRT	Cluster randomised trial
CUA	Cost-utility analysis
DAG	Directed Acyclic Graph
DEGURBA	Degree of Urbanisation
DiD	Difference-in-differences
DPA	Data processing agreement
DRG	Diagnosis-related group
DTA	Data transfer agreement
DTR	Dynamic treatment regimen
EBP	Evidence-based practice
EHIS	European Health Interview Survey
EPOC	Effective Practice and Organisation of Care Group
EPIS	Exploration, Preparation, Implementation, Sustainment
EPS	Permanent Sample (<i>Permanente Steekproef / Échantillon permanent</i>)
ER	Oestrogen receptor
ERIC	Expert Recommendations for Implementing Change
ESHRE	European Society of Human Reproduction and Embryology
EU-SILC	European Union Statistics on Income and Living Conditions
FDG-PET(/CT)	Fluorodeoxyglucose positron emission tomography (/computed tomography)
FRAME	Framework for Reporting Adaptations and Modifications to Evidence-based Implementation Strategies
FOD – SPF	Federal Public Service (<i>Federale Overheidsdienst / Service Public Fédérale</i>)
FTE	Full-time equivalent
GDC	Guideline development checklist
GIN	Guidelines International Network
GP	General practitioner
GPC	Generalized Pairwise Comparison

**ABBREVIATION****DEFINITION**

GRADE	Grading of Recommendations Assessment, Development and Evaluation
HBS	Household Budget Survey
HDA	Health Data Agency
HER2	Human epidermal growth factor receptor 2
HICP	Harmonised Index of Consumer Prices
HIS	Health Interview Survey
HNSCC	Head and neck squamous cell cancer
HSR	Health Services Research
IAC	Individual autocorrelation
IBC	Invasive breast cancer
ICAMO	Intervention-Context-Actor-Mechanism-Outcome
ICC	Intraclass correlation coefficient
ICD	International Classification of Diseases
ICE	Intercurrent event
ICHOM	International Consortium for Health Outcomes Measurement
ICPC	International Classification of Primary Care
ICT	Information and communication technology
IDMC	Independent Data Monitoring Committee
IMA – AIM	InterMutualistic Agency (<i>InterMutualistisch Agentschap / l'Agence InterMutualiste</i>)
IOM	Institute of Medicine
i-PARIHS	Integrated-Promoting Action on Research Implementation in Health Services
IQTIG	Institute for Quality Assurance and Transparency in Healthcare (<i>Institut für Qualitätssicherung und Transparenz im Gesundheitswesen</i>)
IPCW	Inverse probability of censoring weighting
IPTW	Inverse probability of treatment weighting
ISQua EEA	International Society for Quality in Health Care External Evaluation Association
ITS	Interrupted time series
ITT	Intention-to-treat
IV	Instrumental variables
IVC – CSI	Information Security Committee (<i>Informatieveiligheidscomité / Comité de sécurité de l'information</i>)
JRC	Joint Research Centre

ABBREVIATION	DEFINITION
JSOP	Joint system operating procedure
LATE	Local average treatment effect
LYG	Life-years gained
MBD	Multiple baseline design
M(C)ID	Minimum (clinically) important difference
MDT	Multidisciplinary team
MPG – RPM	Minimal Psychiatric Data (<i>Minimale Psychiatrische Gegevens / Résumé Psychiatrique Minimum</i>)
MRC	Medical Research Council
MRT	Middle Range Theory
MZG – RHM	Minimal Hospital Data (<i>Minimale Ziekenhuis Gegevens / Résumé Hospitalier Minimum</i>)
NAM	National Academy of Medicine
NASS	Non-adoption, Abandonment, Scale-up, Spread and Sustainability
NICE	National Institute for Health and Care Excellence
NPT	Normalization Process Theory
NQF	National Quality Forum
NSW	New South Wales
OECD	Organisation for Economic Co-operation and Development
PFM	Process flow mapping
PP	Per-protocol
PPP	Purchasing power parity
PREM	Patient-reported experience measure
PRISM	Practical, Robust Implementation and Sustainability Model
PROM	Patient-reported outcome measure
PSA	Probabilistic sensitivity analysis
QA	Quality assurance
QALY	Quality-adjusted life years
QCA	Qualitative comparative analysis
QI	Quality indicator
RCT	Randomised controlled trial
RD	Root definition
RDD	Regression discontinuity design
RE-AIM	Reach, Effectiveness, Adoption, Implementation, and Maintenance
RIZIV – INAMI	National Institute for Health and Disability Insurance (<i>Rijksinstituut</i>



ABBREVIATION	DEFINITION
	<i>voor Ziekte- en Invaliditeitsverzekering / Institut National d'Assurance Maladie-Invalidité)</i>
ROM	Risk of Mortality
RT	Radiotherapy
SAP	Statistical analysis plan
SD	System dynamics
SET	System evaluation theory
SETIG	Systems in Evaluation Topical Interest Group
SLNB	Sentinel lymph node biopsy
SMART	Sequential Multiple Assignment Randomised Trial
SME	Subject matter expert
SOI	Severity of Illness
SOP	System operating procedure
SSM	Soft systems methodology
STCS	Systems thinking and complexity science
S-W	Stepped-wedge
TBCM	Taxonomy of Behavior Change Methods
TCT	Technical Cell (<i>Technische Cel / Cellule Technique</i>)
TDABC	Time-driven activity-based costing
TDF	Theoretical Domains Framework
TICD	Tailored Implementation for Chronic Diseases
TMF	Theory, model and framework
ToC	Theory of Change
TTE	Target trial emulation
TTT	Theories and Techniques Tool
TwIC	Trial within Cohorts
UCLA	University of California, Los Angeles
VAS	Visual Analogue Scale
VG-MZG – DI-RHM	Nursing Minimal Data Set (<i>Verpleegkundige Gegevens van de Minimale Ziekenhuis Gegevens / Données Infirmières du Résumé Hospitalier Minimum</i>)
VSM	Viable systems model
WBI	Whole-breast irradiation
WHO	World Health Organization

■ SCIENTIFIC REPORT

INTRODUCTION

KEY POINTS

- This process note is developed to support KCE researchers in the **(predominantly) prospective evaluation of health policy and health services interventions**, complementing existing methodological guidance and responding to the growing demand for evaluations of policy interventions (e.g. new payment models, care trajectories, integrated care programmes).
- The structure follows the **Medical Research Council (MRC) framework for complex interventions**, with a particular focus on the evaluation phase and its key components. This includes effectiveness, process, economic and implementation evaluations. These components may be studied separately or in combination.
- This is a **living document**: chapters will be revised and expanded as methodological developments emerge. Additional chapters will be added to address current gaps and evolving needs within the KCE evaluation programme.

Evaluation of Health Services Research and health policy interventions: complementing retrospective with prospective research methods

Over the past two decades, KCE has played a key role in advancing evidence-informed policymaking by assessing existing policies and supplying evidence from literature reviews and international examples. Topics within the Health Services Research (HSR) domain were traditionally addressed using methods such as systematic reviews, international comparisons, retrospective analyses of administrative data, qualitative research, and cross-sectional surveys. However, these approaches are no longer sufficient in all cases.

In recent years, the nature of topics submitted for the KCE annual programme has changed. Federal bodies such as the National Institute for Health and Disability Insurance (RIZIV – INAMI), the Federal Public Service Public Health (FOD – SPF), and the Minister of Health's policy unit increasingly request **prospective evaluations of policy measures or pilot projects**. Current examples include the New Deal in general practice,¹ and the evaluation of an Article 56^a project introducing a new financing mechanism in home nursing.²

Although these studies still fall within the HSR domain and continue to draw on familiar methods, they also require tailored methodological guidance. A dedicated process note is therefore needed to complement existing process notes (e.g. on HSR, qualitative research, patient participation, stakeholder involvement).³ It also aims to address studies that do not fit within the traditional scope of the KCE clinical trials programme.

A process note for the evaluation of mostly predominantly policy interventions

The aim of this process note is to guide KCE researchers in developing protocols for studies involving (mainly) prospective evaluations of policy interventions. Topics will typically, but not always, concern complex policy interventions (i.e. multicomponent interventions with impact on different target groups, see description below). While much of the guidance has been developed with the evaluation of

^a Article 56 of the Health Insurance Act enables innovative care models and therapies to receive temporary and conditional funding from RIZIV – INAMI in Belgium.



complex interventions in mind, care has been taken to ensure that methodological approaches relevant to less complex interventions are also adequately reflected.

Such evaluations will not only consider effectiveness (does the intervention produce the intended outcome(s)?) but also a process evaluation (e.g. distinguishing between intervention failure and implementation failure in negative studies and understanding which components drive the effect in case of positive trials). It may also require an economic assessment (e.g. economic evaluation, budget impact, costing study) and an evaluation of acceptability or feasibility of the study methods and the intervention itself.⁴ Studies may combine several objectives (Is there an effect? What are the barriers during implementation of the intervention?) and might require hybrid designs⁵ (see Chapter 8). Yet, a KCE study may also focus on a single objective. This decision should be made during the scoping phase together with the commissioner, based on practical, methodological, and policy considerations. The existing level of evidence is a useful guide for defining the study objectives. When a robust evidence base exists, the primary focus can be on evaluating the implementation strategy. When the evidence supporting the intervention is limited, it is preferable to focus the evaluation on the effectiveness, combined with a process evaluation.

So, in some cases, for instance, when the intervention already has a robust evidence base on effectiveness, the evaluation can focus on the implementation process. However, while not the preferred scenario, even when effectiveness evidence is limited (e.g. a consensus-based care pathway already introduced in practice) or not possible (e.g. not possible to set up a design with a comparator), it may still be policy relevant to understand how the intervention was implemented and produced change. Yet, process evaluations are not restricted to implementation processes. They may also examine stakeholders' experiences with an intervention and identify which intervention components contribute most to observed outcomes.

This process note builds on existing handbooks (e.g. Kaehne and Feather (2025)⁶; Richards et al. (2015; 2026))^{7, 8} and methodological literature and provides in each chapter specific guidance on their application within KCE research.

What do we mean with 'a complex intervention' in this process note?

This section outlines the concept of complex interventions. The complexity of an intervention may arise from several features. These include the number and diversity of components (e.g. sub-interventions that form the complex intervention), the range of behaviours targeted, the skills required from those delivering and receiving the intervention, the number of groups or settings involved, the flexibility allowed in delivery, the interaction with contextual factors, and the multidimensional nature of the outcomes targeted.^{4, 9, 10}

The overall effect often depends not only on individual components, but also on how these components interact with each other and with the wider context in which the intervention operates.¹⁰⁻¹² As a result, the specific contribution of each component is not always clear.⁴

Although the core components of complex interventions are to be standardised, many interventions allow flexibility and local tailoring (e.g. a RIZIV – INAMI convention in which the composition of a multidisciplinary team, together with some additional requirements (e.g. multidisciplinary team meetings) is specified but the mode of implementation is left open). Standardisation can therefore also focus on underlying functions and processes rather than rigid prescriptions for each component.¹³ A theoretical decomposition of the intervention (and its components), combined with agreed boundaries on acceptable and prohibited variations, helps to structure its implementation. This approach enables contextual adaptation while safeguarding core components.¹⁰

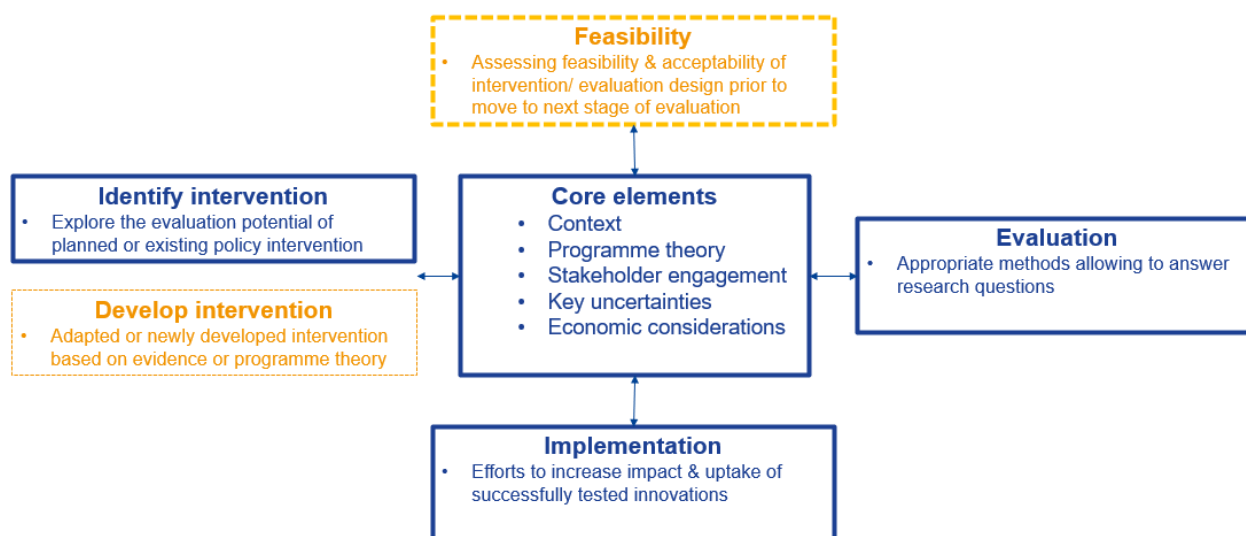
Viewing interventions as 'events in systems' underscores the need to consider mechanisms of change and contextual factors that shape effectiveness.^{10, 14, 15} It points to the importance of (the often evolving) context and of the conditions required for successful real-world implementation.⁴

EVALUATION FRAMEWORK USED TO DRAFT THIS PROCESS NOTE

The Medical Research Council (MRC), United Kingdom published a framework in 2000 to evaluate complex interventions. A checklist is provided to guide research based on the MRC framework.

The MRC framework was adapted and updated on several occasions.^{16, 17} In this document, we start from the recent publications of Skivington et al. (2021; 2024), see Figure 1.^{9, 10}

Figure 1 – Evaluation Framework Complex Interventions



Source: Adapted from Skivington et al. (2021; 2024)^{9, 10}; Blue: included in the current process note; orange: not part of the process note.

The framework includes four phases: development or identification of the intervention, feasibility, evaluation, and implementation.

A common set of **core elements** for each phase (Figure 1) includes:

- the context (physical, spatial, organisational, social, cultural, political, economic, etc.);¹¹
- a programme theory (a theory on how an intervention is expected to produce its effects, and under which conditions; see Chapter 1)¹⁸;
- stakeholder engagement¹⁰ at all stages of an evaluation (cf. standard KCE procedure);
- identification of key uncertainties (in the process of formulating the research questions); and
- economic considerations like a comparative analysis of alternative courses of action in terms of both costs (resource use) and consequences (outcomes, effects).

All these elements are relevant for KCE evaluations. Yet, the MRC framework also includes ‘intervention refinement’ as a core element. This implies that across the phases of complex intervention research, and particularly during transitions, interventions may require refinement based on data or further programme theory development.¹⁹ KCE evaluations will mainly be focused on one phase at a time and mostly the ‘evaluation phase’. Therefore, intervention refinement is less relevant since adaptations will most likely be proposed in the recommendations of the KCE evaluation study and will have to be implemented in follow-up studies or via policy measures. What’s more, in case of effectiveness studies, changes to the intervention will generally be rare as interventions are ideally kept stable throughout the study.



It is also possible that a KCE study focuses on a context analysis only. This can be relevant for policymakers to learn about the anticipated added value of the intervention, the willingness to try (e.g. survey about acceptability of the proposed policy intervention among healthcare professionals), etc. The KCE teams that have to perform such a context analysis are recommended to consult Chapter 10 (systems thinking and complexity science) and Chapter 5 (frameworks to analyse context). In addition, they can rely on the current KCE procedures like for instance conducting surveys, performing qualitative research, etc.

We briefly explain the four phases below:

- The **development phase** is a critical first step in complex intervention research.¹⁹ It includes identifying existing evidence (e.g. through systematic reviews of individual components, or the complex intervention itself in case the focus is on implementation: local adaptation and scaling up), careful modelling of how the intervention is expected to work, and the articulation of a programme theory.⁴ This programme theory outlines the key components of the intervention, their interactions, the mechanisms of change, and the contextual factors that may shape outcomes.^{4, 20} As KCE aims to maintain an independent role as evaluator, it does not develop interventions; this is the responsibility of the commissioner (e.g. RIZIV – INAMI or FOD – SPF). Policy-led interventions represent another common entry point into evaluation research, often being designed externally and implemented before researchers are involved. In such cases, the first task is to uncover the implicit theoretical assumptions underpinning the intervention and to construct an explicit programme theory (see Chapter 1).^{18, 21} Even when rollout is already underway, this step remains essential to identify mechanisms of change, contextual influences, and meaningful outcomes.¹⁰
- **Feasibility and pilot studies** are designed to resolve key uncertainties before a full evaluation. Feasibility studies focus on implementation and acceptability. Pilot studies are testing key trial procedures on a small scale. They assess predefined progression criteria related to the evaluation design (e.g. recruitment, retention, data collection) and the intervention itself (e.g. content, delivery, acceptability, adherence, cost-effectiveness).¹⁰ Skipping feasibility risks undermining both the design and implementation of subsequent evaluations, whereas incorporating feasibility findings can improve acceptability and strengthen the basis for piloting and full-scale evaluation.^{4, 10} While feasibility studies can be part of the KCE evaluation programme, they are not discussed further in this process note.
- The **evaluation phase** represents the core of complex intervention research and extends beyond the question of whether an intervention ‘works’. It is also the primary focus of this process note. Evaluation must consider multiple dimensions, including how an intervention interacts with its context, what wider impacts it produces, how it contributes to system change, and how the resulting evidence can inform decision-making in real-world settings.^{4, 10} Evaluation of complex interventions often include multiple outcomes across different domains, aligned with programme theory and stakeholder priorities. These may encompass both individual- and system-level changes, such as shifts in organisational practices or policies.¹⁰
- Chapter 2 introduces experimental study designs for evaluating the effectiveness of interventions. However, in the field of health services and policy interventions, practical, ethical, and contextual considerations may limit the feasibility of these designs. Chapter 3 therefore discusses quasi-experimental designs as an important alternative. Particular attention is given to the methodological requirements needed to draw valid conclusions and to address potential sources of bias. Specific consideration is given to designs using routinely collected data.²² Chapter 11 describes the use of (mainly) administrative data in the context of the KCE evaluation programme.
- Quantitative evidence on impact and effectiveness alone is rarely sufficient to evaluate complex interventions, where context and mechanisms of change are as important as effect estimates.^{4, 10} **Process evaluation** (see Chapter 5) therefore plays a vital role, providing insights into fidelity, implementation quality, mechanisms of impact, and contextual influences. By explaining why



interventions succeed or fail, and how they are adapted across settings, process evaluation strengthens programme theory and supports the transferability of findings.²³⁻²⁵

- A separate chapter is dedicated to the '**realist evaluation approach**' as this is an emerging approach which is often used in the context of the evaluation of complex interventions (see Chapter 6). Its primary aim is to identify the mechanisms underlying observed outcomes following an intervention and the contextual conditions that enable them. The focus is on explaining why, for whom, and under which circumstances an intervention works. It is a theory-driven approach and explicitly method-neutral, meaning that both quantitative and qualitative methods can be used to uncover the mechanisms that generate intervention outcomes. In the literature, a realist evaluation approach is often framed in terms of context (C) - mechanism (M) - outcome (O) configurations.^{26, 27}
- The evaluation of **economic impact** (see Chapter 7) is also within the scope of the KCE evaluation programme. While full cost-effectiveness analysis may be performed (cf. standard HTA procedures KCE), this process note will focus on aspects such as budget impact analysis and costing methods.
- Since policy interventions are embedded in larger systems, the use of '**Systems Thinking and Complexity Science**' in the context of the KCE evaluation programme is discussed in Chapter 10.
- While **implementation issues** should be anticipated during development, feasibility, and evaluation, phase 4 of complex intervention research addresses the long-term transferability, sustainability, and impact of interventions. The key concern is no longer whether an intervention works under study conditions, but whether it can be scaled, adapted, and maintained in diverse real-world settings.^{4, 8, 10} In the current process note, we describe the different steps to be taken during implementation research (see Chapter 8) and have a dedicated chapter to the development and monitoring of quality indicators (see Chapter 9).

The outline of the process note is as follows:

- Chapter 1: Programme theory
- Chapter 2: Experimental designs
- Chapter 3: Quasi-experimental designs
- Chapter 4: Analytical considerations
- Chapter 5: Process evaluation
- Chapter 6: Realist evaluation
- Chapter 7: Economic evaluation
- Chapter 8: Implementation research
- Chapter 9: Quality indicators
- Chapter 10: Systems thinking and complexity science
- Chapter 11: Use of administrative data.

This process note is inherently evolutionary in nature. This document represents a first version and starting point and will be updated as methodological insights continue to develop. This will include both updates of the current chapters as well as the introduction of additional chapters on topics that are currently not yet developed.



1 PROGRAMME THEORY AND LOGIC MODEL

KEY POINTS

Programme theory and logic model

- A programme theory is used to guide the evaluation of complex interventions as it helps to **open the 'black box'** of the intervention at study.
- A programme theory **articulates how an intervention is expected to bring about change**, the mechanisms through which this occurs, and the **contextual** conditions influencing these **mechanisms**.
- Making assumptions explicit enhances **transparency** and critical scrutiny, enabling assessment of whether causal pathways are evidence-informed, while also helping to anticipate on unintended consequences.
- A well-developed programme theory supports a **common understanding** among programme developers, implementers, policymakers, and evaluators, and helps surface gaps or inconsistencies in the underlying logic.
- **Logic models** provide a simplified visual representation of the programme theory, illustrating the sequence from inputs and activities to outputs and outcomes.
- These models **guide data collection**, inform indicator development, and support **communication** with commissioners and stakeholders.
- Within KCE evaluations, programme theories and logic models are important for designing rigorous, context-sensitive assessments of complex interventions. They should not be seen as end-product but used to support evaluation. Their application in specific projects might require some pragmatism: while systematic guidance offers structure, the level of detail and iteration required should be balanced against project timelines to ensure **timely development of a robust programme theory**.

1.1 Introduction

The development and evaluation of complex (policy) interventions (e.g. care pathways, integrated care programmes, or innovative funding models) require **a clear understanding of the assumptions** that underpin how these interventions are expected to bring about change. Such interventions can only be meaningfully assessed when the mechanisms through which they operate and the contextual conditions that influence them, are made explicit.

Programme theory enables such a **common** understanding and **language** among programme designers, implementers, policymakers, evaluators, and service users (e.g. healthcare professionals, patients, families).²⁸ Moreover, collaborative development of a programme theory surfaces implicit assumptions, identifies conceptual gaps, and enhances the acceptability and credibility of subsequent monitoring and evaluation decisions.

Programme theory also aims to support researchers **to specify the study protocol**, to decide about which data to collect (e.g. select validated instruments, use administrative data, develop interview guides), to decide on statistical testing, etc.

This chapter outlines how a programme theory can be used to articulate this underlying logic, and how a logic model offers a structured way to represent it visually. The aim of this chapter is to support KCE researchers who are setting up an evaluation of a complex intervention to use a programme theory and logic models.

1.2 General concepts explained

Programme theory: a way to open the 'black box' of an intervention

Every (complex) intervention will be based on explicit or implicit assumptions about the underlying problem and how an intervention works and creates positive changes in relation to this problem.²⁹

A theory about the programme (programme theory) provides the rationale behind the intervention. It offers a conceptual framework needed to understand how an intervention is expected to generate change, under which conditions, and through which mechanisms.³⁰⁻³³

Programme theory is, in this chapter, described as follows: *'it explains how an intervention is anticipated to generate its intended effects and the conditions under which these effects are likely to occur'*. It sets out:

- the **outcomes**;
- the essential **components of the intervention** and the ways in which they relate to one another;
- the **mechanisms** through which the intervention is expected to operate;
- the **contextual** features that may shape or modify those mechanisms; and the ways in which the activation of these mechanisms may, in turn, influence the surrounding context.^{9, 29}

A solid programme theory is structured, consistent, well-founded, realistic in its assumptions, and feasible to implement.^{34, 35}

This can be illustrated by an example.

In KCE report 405³⁶ a study protocol was developed to evaluate the impact of the introduction of a new payment system in home nursing. In this case, the **outcomes** were patient-related (e.g. improved self-management) and nurse-related (e.g. job satisfaction). The intervention included several **components** (e.g. hourly payment instead of the mixed payment system based on fee-for-service with lump sum per day payment; the introduction of BelRAI to assess the cognitive, physical, social and behavioural situation of the patient). One of the assumed **mechanisms** of the payment per hour is that nurses will experience 'more autonomy' as care will be less activity-driven. This is assumed to have a positive impact on job satisfaction. In addition, the payment per hour is expected, together with a better understanding of the functional status of the patient via BelRAI, to introduce a 'new way of working' focusing on the needs of the patient rather than on the activities that are financed. This is assumed to increase patient education resulting in more self-management behaviour. The **context** of a shortage of nurses was assumed to potentially interact with the new way of working (e.g. insufficient time to invest in self-management behaviour due to a shortage of staff).

INTERCHANGEABLY USED CONCEPTS

Many different concepts (e.g. programme theory, theory of change, theory of action, logic model) are used interchangeably and in a non-consistent way, although they differ in emphasis.^{19, 37-39} This might create confusion. The common **principle** prevails that all programmes embody a theory, i.e. an implicit or explicit proposition that *if* specific actions occur, *then* certain **mechanisms** will be activated, *leading* to specified outcomes.⁴⁰

A **programme theory** has two components:²⁹

- **theory of change (ToC)** provides an explicit, hypothesis-driven explanation of how and why a programme or (set of) intervention(s) is expected to produce change within a given context.^{31, 41, 42} This can be based on a formal theory, or it can be developed by combining practical experience, theoretical insights, and findings from the scientific literature, and expressed as hypotheses or assumptions about the mechanisms and causal reasoning. For instance, an eating disorder

prevention programme might theorise that improving body image and challenging unhealthy social norms around weight reduce disordered eating behaviours among adolescents.

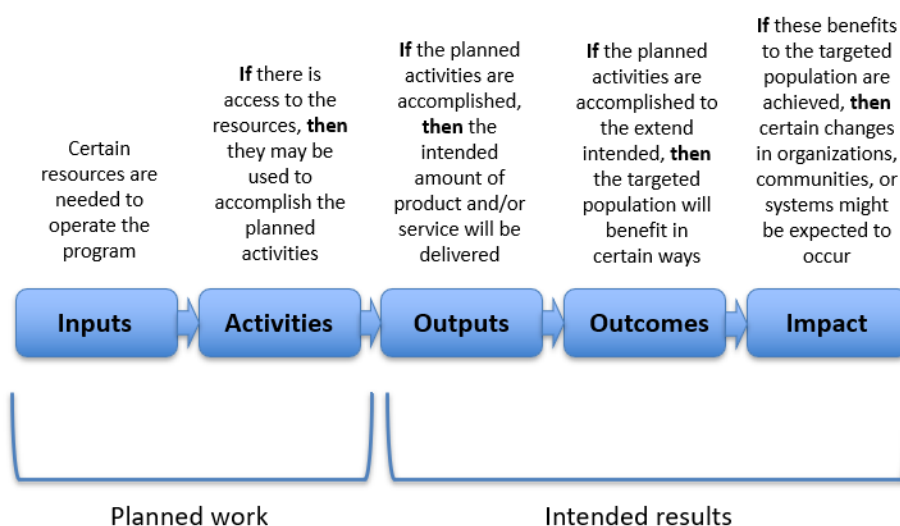
- **theory of action** sets out the activities that need to be carried out to bring about the intended changes.²⁹ Concrete activities that are assumed to lead to change for the eating disorder prevention programme are training school counsellors, delivering classroom-based sessions on media literacy and body image, and providing referral pathways to specialist support.

A programme theory can take various narrative or visual forms, but its main characteristic is that it provides a detailed causal explanation of how and why an intervention is expected to produce its intended outcomes. Logic models are frequently used to make a concise visual summary of this programme theory, depicting the key components and causal sequence from inputs to outcomes (see Figure 2).

- **A logic model** is a simplified, structured, visual representation of a programme theory that illustrates the sequential flow of a programme from inputs and activities through to outputs and outcomes.^{40, 43} It provides a typically linear view of how resources are used to deliver specific activities and how these activities are expected to deliver short-, medium- and long-term outcomes.

If we apply this to the example of 'home nursing' (KCE report 405):³⁶ one of the inputs (resources) was the 'creation of a dedicated project team'. One of the activities (an intervention component) was the 'development of a payment system per hour for home nurses' which is expected to result in an output: 'hourly pay of home nurses working in pilot practices'. This is expected to result in, among others, improvements in outcomes, 'an increased job satisfaction'. The long-term impact (not included in the logic model of KCE report 405) could, for instance, be an increased retention of home nurses and decreased nursing shortage.

Figure 2 – Example of a simple logic model from left to right



Source: Adapted from Kellogg Foundation (2004)⁴⁰

While a programme theory focuses on mechanisms and explanatory reasoning, a logic model emphasises structure, sequencing and operational logic. It answers the basic proposition: 'We plan to do X, which is expected to result in Y'.^{40, 44} Its visual clarity makes them particularly useful for programme planning, communication, and synthesis, providing a shared reference point that supports collaboration and alignment across the project team and stakeholders. Of note, the distinction between intervention components and implementation strategies in logic models should be clear. Intervention

components are the active ingredients of the programme, i.e. the resources, tools, or services offered to participants. Implementation strategies are the actions taken to deliver those components to the target population. Conflating the two risks treating delivery activities as mechanisms in the subsequent programme theory (frequent error in realist evaluation). It is therefore recommended to conduct this thinking exercise explicitly during logic model development: for each element, ask whether it is part of what is being offered (intervention component) or how it is being implemented (e.g. strategies to ensure its uptake). Visually distinguishing these, e.g. by numbering or colour-coding resources and activities separately in the logic model, makes implementation strategies explicitly visible and prevents them from being carried forward into the programme theory as mechanisms.

For KCE's purposes, the overarching programme theory (e.g. narrative form, vignettes, ...) can be visualised via a logic model. In the following sections, we outline the core elements that may make up both a programme theory and a logic model, and how the development process can unfold.

1.3 Elements of programme theory and logic models

This section outlines the common elements that make up a programme theory and a logic model (these elements are larger than the basic elements of a logic model presented in Figure 2) which are also used to describe the development process described in section 1.4.^{30, 40, 43, 45-47} A specific example and figure of a logic model can be found in KCE Report 405.³⁶

A programme theory starts with the articulation of the **specific problem** the programme aims to address. In most cases, this problem will already have been identified by the party submitting the topic to KCE, but it will need to be further explored and situated within the broader context. Next, the **goal** of the programme is formulated: what are the ultimate outcomes one is aiming for with the programme? Consequently, the **target population** can be specified. The following questions need to be clarified: who is/are the target group(s) of the programme (e.g. patients, healthcare professionals, volunteers), what are the demographics of these target groups (e.g. age, cultural background), who else is involved or affected by the programme/complex intervention (e.g. educators) and what is their role?

From there, 'a backwards mapping process'^b can be performed starting with identifying the outcomes.

- The **long-term outcomes** (also called distal outcomes) that the intervention can reasonably achieve basically resolve the issue identified in the problem statement.
- Note that the timepoint beyond which the programme cannot reasonably be held responsible for change is called the '**ceiling of accountability**'. It is usually situated between the (long-term) outcomes and the impact of the programme. It indicates the point where other elements, not related to the intervention, will interact with the outcomes of the intervention and influence (reinforce or inhibit) the ultimate result of the programme.
- After identification of the long-term outcomes, the **intermediate outcomes** (also called medium-term outcomes) that represent the necessary preconditions or sequential steps required to reach those long-term outcomes, are specified.
- **Short-term outcomes** (also called proximal outcomes) are those expected to arise immediately from the intervention. They are often changes in skills or knowledge and are the easiest to measure.

While this differentiation between short-, medium- and long-term outcomes should be made in theory, challenges might occur towards data collection to measure these, especially medium- and long-term outcomes, often due to limited resources or time. In the case of the home nursing example, no

^b Backwards mapping process refers to the fact that in developing a logic model, you do not start with step 1 'inputs', rather you start with defining (long-term) outcomes and impact and 'work backwards' towards the inputs. For detailed information see section 1.4.



distinction was made. Except that the outcomes (e.g. job satisfaction) were assumed to result from improvements in process outcomes such as 'increased job autonomy' and 'short-term' outcomes such as 'decreased workload'. As such, a timing (sequential) dimension was also included in that logic model. Yet, long-term outcomes, such as retention of nursing staff, were beyond the scope of the study timeline.

The next step is to specify the **activities** or strategies, which outline the various key components of the complex intervention, i.e. what will be done (e.g. developing an hourly pay for home nursing activities, providing telemonitoring to monitor symptoms of COVID-19 at home). What will be delivered by the intervention(s) are the expected **outputs**, i.e. the direct measurable results of the intervention(s) or activities done (e.g. number of educational trainings provided, number of drugs given). Many people struggle to distinguish between outputs and outcomes in programme theory. In simple terms: outputs are what you do; outcomes are the changes that occur because of what you do. The outputs are quantifiable by e.g. number, development (yes/no), amount and are often measured by a process indicator while the outcomes are measured by an outcome indicator.

To determine which resources are needed to carry out the activities, the required **inputs** have to be clearly defined: resources such as funding, facilities, equipment, staffing, expertise (note that if expertise is not available people need to be trained (activity)) and time have to be provided.

Further, **the assumptions**, the beliefs about how the programme will work are made explicit. This requires the formulation of testable hypotheses.²⁹ It is worth spending sufficient time on this section. It is good practice to involve stakeholders and to ask them to challenge and to complement the draft of identified assumptions to arrive to a full list of assumptions to be addressed. This also includes the formulation of the rationale (e.g. evidence, theory and/or consensus base) for selecting particular components and outcomes. These elements are then linked together in **a pathway of change** that visually maps the hypothesized mechanisms through which the intervention is expected to work (the 'how' and 'why' an activity is expected to result in desired outcomes). In this step, also indirect effects and feedback loops can be considered. Next to the desired consequences, it is also important to think in advance of **unintended consequences** of the intervention (e.g. will the payment per hour of home nurses, for instance, result in less patients treated per nurse as home visits take more time and consequently cause waiting lists?).²⁹

Also the **external factors** or **context** have/s to be taken into account. This element requires to consider the environment in which the programme is being delivered. Economic, political, cultural, historical and social contexts all impact the way the programme is delivered, and the outcomes that can be achieved. Likewise, the programme itself has potential to impact some of these factors too. There are specific frameworks to guide context analysis (see Chapter 5).

The result of this process is, in general, **a visual model (logic model) and an accompanying narrative** that explains the logic of the intervention. Based on the programme theory or logic model, **metrics and indicators** have to be developed to operationalise outcomes and enable an evaluation including an assessment of progress and effect.

Specific software programmes such as DoView, Microsoft Visio, and Atlas Ti can help to format the programme theory.^{48, 49} However, for the work at KCE we suggest to use a simple premade template document.⁴⁵

1.4 Development process

There are various ways to develop a programme theory or a logic model.^{19, 33, 49-58} What matters is that the 'black box' of the intervention is systematically unpacked and that whichever approach is chosen is applied consistently.

The development process includes a preparatory phase and a modelling phase (see Figure 2). The development of a programme theory or logic model should be aligned with the needs of the programme developers (e.g. RIZIV – INAMI, FOD – SPF).

Best practice is to formulate a programme theory based on insights from theories and existing evidence, **at the start of a project**, with active involvement of key stakeholders, and to refine it in subsequent phases. In case the programme is designed **without the involvement of the evaluation team**, it is good practice that the evaluation team first constructs a theoretical understanding of the intervention before to start with the evaluation.¹⁸ In these situations, programme theory is primarily used to unpack the 'black box' of a programme and particularly supporting the evaluation process.^{9, 10, 20, 21, 59, 60}

1.4.1 Preparation

The preparatory phase establishes the analytical foundation for developing a robust programme theory and logic model. Before starting the modelling, it is important to prepare it by synthesizing the empirical evidence related to the research question.^{7, 8, 61} Throughout this phase a mixed-methods approach can be applied. The KCE research team may select those methods that are most feasible (related to the topic) and expected to yield the most informative insights given the available resources for programme theory and logic model development.

SOURCES OF INFORMATION TO CONSULT

In this phase, a structured synthesis of all available information on why and how the proposed programme, with its intervention(s), is expected to work. Throughout this stage, **early engagement with the developers of the programme** is essential. Dialogue focuses on clarifying how the proposed interventions are expected to generate intended outcomes, if these assumptions are based on theory, empirical evidence or expert opinions, how these align with policy priorities, and what implementation constraints or enablers must be accommodated. This ensures that the emerging model reflects both scientific reasoning and real-world feasibility.

An important source to consult in this phase is the **submitter of the topic**. A template was developed to gather information (see Appendix 1). The template aims to collect information about the problem statement and the context of the intervention (e.g. *where will the intervention be installed? what are barriers and facilitators?*) as well as identification of best practices (supported by scientific evidence: a previous KCE report? A systematic review or RCT?). Consequently, information is gained on the different interventions (activities) of the programme (e.g. *what is the primary/secondary objective of the intervention? which data should be collected to measure these outcomes?*). Next the processes and mechanisms that are thought to be responsible to thrive change and the underlying assumptions have to be made explicit. Detailed information on the intervention delivery is also mentioned (e.g. content, sequence, dose, etc.) as well as the resources/inputs needed to deliver. Also specific information on planning and evaluation of the intervention is provided through this template.

At the same time, **technical reports** and other **grey literature** can be consulted by the KCE team, including implementation experience and previous evaluations (e.g. earlier evaluations conducted by university teams). This includes also consulting (previous) policy documents on the matter.

Doing a systematic **literature** study on complex interventions (or its components) is the standard when developing a programme. Yet, it is also possible to perform a (pragmatic) search for peer-reviewed literature to gain more information on the interventions' effectiveness and/or context. In that case, an option is to start with the identification of systematic reviews evaluating the effectiveness of the concerned intervention(s).

It is also possible to focus on specific **international practices and case studies** (e.g. via literature review and/or the consultation of experts), particularly to understand how similar complex interventions perform in other health systems. If an evaluation of a complex intervention in another health system exists, it can be used as a source of inspiration for the programme theory and logic model. The elements included (such as activities, outputs, outcomes, etc.) implemented can be checked and verified against the Belgian health system.



Field visits can be a very valuable source of information, not only to grasp the current way of work but also to estimate where contextual factors (e.g. funding system, educational level) of the foreseen interventions might occur. For the same reasons, (semi-)structured **interviews**, **focus groups** as well as **surveys** can be used to gather stakeholder information. These are not only important data sources to unravel the implicit assumptions made by the programme developers, it will also increase support from stakeholders for the evaluation of the intervention.

A **structured context analysis** can be performed, examining institutional, regulatory, financial, and sociocultural conditions.

WHAT SHOULD BE CLEAR AT THE END OF THE PREPARATORY PHASE?

Before any causal pathways are mapped, the KCE multidisciplinary research team may develop a precise and shared understanding of **the problem** the intervention seeks to address. The following questions can be helpful to understand the problem statement (7, 8, 61 and adapted from Shakman & Rodriguez, 2015⁶²):

- What is the core problem the intervention seeks to address?
- What are the immediate and underlying causes of the problem?
- Who is directly impacted by this problem?
- Who contributes to, shapes, or is otherwise involved in the problem?
- Who else (individuals, organisations, stakeholders) is already working on the issue, and who has an interest in seeing it solved?
- What evidence exists from research, evaluations, and field experience about the nature of the problem and the mechanisms that perpetuate it?
- What do we know about how the target group experiences this issue and what matters most from their perspective?

This ensures that the modelling process starts from a grounded understanding of the problem. Together with the problem statement, the **goal (aim)** of the programme and the **target population** should also be clarified. Also the context that will be in **scope of the study** has to be described. For example, will the evaluation solely focus on the health system, or is the welfare system also evaluated (e.g. in the home nursing example a shift of care might take place to the welfare system: family aid, etc.).

Once the problem statement, goal, target population and context/scope are sufficiently articulated, the multidisciplinary KCE team undertakes a structured synthesis of all available information on why and how the proposed programme, with its intervention(s), is expected to work. The purpose of this synthesis is to articulate an initial conceptualisation of the intervention's pathways of influence, expected outcomes, and critical assumptions. This becomes the **'first draft' of programme theory**, an early working model that captures preliminary causal reasoning without yet aiming for full precision. The preparation phase yields a well-informed starting point for formal modelling. It functions as due diligence: validating that the intervention concept is grounded in evidence, context-aware, and aligned with the needs and expectations of the commissioner (e.g. RIZIV – INAMI). The scope and depth of this phase will, however, depend on the project's timelines and available resources; for each study, the KCE team will assess how extensively this preparation can be undertaken and whether certain steps need to be approached in a more streamlined manner.

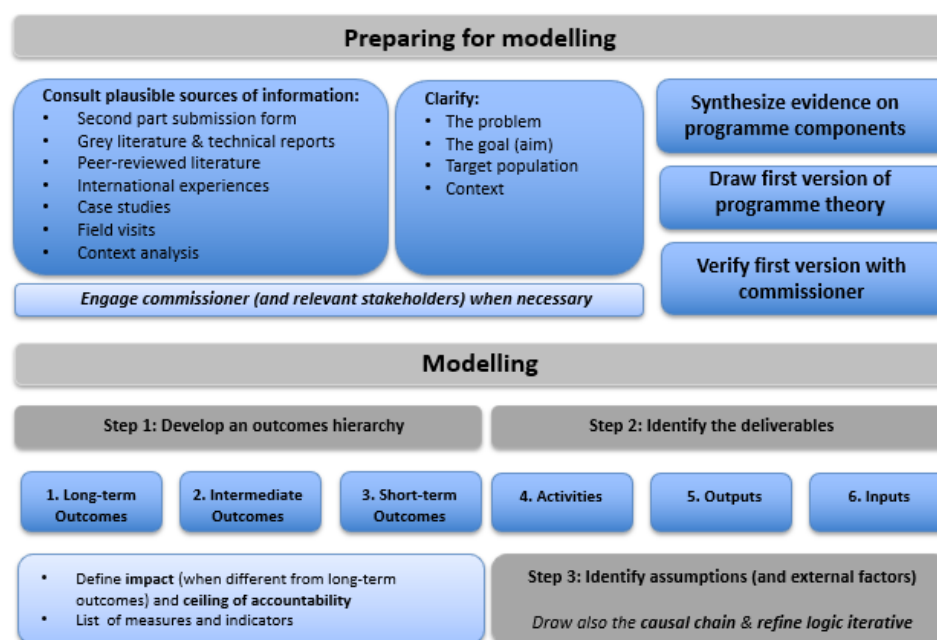
1.4.2 Modelling phase

The modelling phase translates the preparatory analysis into a structured programme theory. Drawing on the New South Wales (NSW) guide,^{43, 63} the modelling process consists of three interlinked steps:

- constructing the outcomes hierarchy,
- identifying deliverables, and
- articulating assumptions while reviewing programme logic.

This process is also called the **back casting process** or backwards mapping: i.e. starting from the outcomes and then listing the activities/outputs and resources required to achieve these outcomes. Take into account that the stepwise approach presented serves as a guiding framework and can be applied with flexibility by the KCE research team depending on the context and needs.

Figure 3 – Preparing for modelling and modelling process



Source: Adapted from NSW^{43, 63}

STEP 1: DEVELOP AN OUTCOMES HIERARCHY (OUTCOMES)

The **first step** in modelling the programme theory and logic model is to **develop an outcomes hierarchy**. The outcomes hierarchy is seen as the backbone of the programme theory and logic model. It represents what the intervention is expected to generate. Based on the information retrieved from the preparatory modelling phase, a comprehensive list of all candidate outcomes can be compiled. Outcomes can be clustered into logical or thematic groups, distinguishing, for example, behavioural, organisational, clinical, or system-level change. The outcomes should be sequenced in an if-then logic, showing how short-term changes (short-term outcomes) enable intermediate shifts (intermediate outcomes), ultimately contributing to long-term outcomes. For each causal link in the outcomes chain, the underlying rationale should be articulated, supported by evidence or theory that explains why the expected change is plausible.

For example, the long-term outcome of the programme may be *to reduce rates of overweight and obesity*. As the medium-term outcome has to be achieved before the long-term outcome, it may be



reduced consumption of energy-dense food. The short-term outcome can be *increased healthier choices in food outlet*. Verifying this by using the if-then logic provides the following sequential outcomes chain can be formulated:

- If there is an increase in healthier choices in food outlet, **then** there will be reduced consumption of energy-dense food, which will **then** lead to reduced rates of overweight and obesity.

It is important to translate the outcomes in metrics and indicators (see also Chapter 9 on Quality indicators). Indicators and measures are developed as a follow-up step after the construction of the logic model. It is advisable to include, within the study protocol, a table for each of the logic model's sub-domains, outlining how these are operationalised into metrics and indicators. In the above example on 'obesity', indicators could include those listed in Table 1. The selected measures are preferably, as much as possible, based on validated scales.

Table 1 – Examples of indicators

Outcome	Indicator
Short-term: increased healthier choices in food outlets	% food outlets offering a defined minimum proportion of healthier options (e.g. 30% of menu items meeting nutritional criteria) Change in sales volume of healthy vs. energy-dense items in participating outlets % food outlets implementing nutritional nudges (e.g. placement, signage)
Intermediate: to reduce consumption of energy-dense food	Average daily intake of energy-dense foods per target population (measured via food diary/survey) Change in household purchases of energy-dense products Self-reported reduction in consumption of high calorie snack foods
Long-term: to reduce rates of overweight and obesity	Prevalence of overweight and obesity in target population (BMI-based; mean BMI, waist circumference, body fat percentage) Incidence of obesity-related metabolic risk factors (e.g. impaired glucose tolerance, dyslipidemia) Trends in obesity-associated healthcare utilization or costs

Within this outcome exercise, also the **ceiling of accountability** can be identified by distinguishing outcomes directly attributable to the intervention from broader changes influenced by external factors, and which are considered beyond the ceiling of accountability.

Verification of the selection of outcomes and the outcomes hierarchy with the commissioners (and other stakeholders), may be necessary to ensure accuracy, feasibility, and shared understanding of anticipated change pathways.

STEP 2: IDENTIFY THE DELIVERABLES

Once the outcome structure is established, the modelling process specifies the concrete operational components that drive change. Therefore, **the second step is identifying the deliverables**, i.e. the activities that need to be undertaken to deliver the outputs and the outputs to be delivered.

Start with listing the **activities** (interventions) described in the programme and formulate at least one **output** (i.e. immediate deliverables, products, or services produced through these activities) for each activity. Certain activities for the specific example, described above (i.e. obesity), may be:

- *developing fact sheets, developing and promoting training for food outlet managers, and listing dietitian support.*

The outputs, for this example, can be:

- *(number of) sheets distributed, (number of) outlet managers attending the training, and dietitians providing pro- and reactive support.*

Clarifying deliverables ensures that the model does not remain purely conceptual but explicitly connects what the programme 'does' to what it 'hopes to achieve'. This step enhances evaluability by creating a clear line of sight between intervention design and observed outputs.

In this step, it is also important to **estimate what inputs (resources)** are needed to conduct the activities as well as the implementation strategies, i.e. the link between inputs and activities, also in case the programme is already developed. Inputs in the example described above can be:

- *staff, funding and partnerships.*

Ultimately, the pathway of change can be mapped, linking activities and outputs to specific outcomes and illustrating how actors, mechanisms, and contextual determinants interact along the pathway. Also potential feedback loops or bi-directional influences can be identified, acknowledging that health system interventions often operate in dynamic, non-linear environments.

STEP 3: FORMULATE ASSUMPTIONS AND REVIEW THE PROGRAMME THEORY

In a **third step**, the model requires a systematic articulation of **assumptions** which are the contextual conditions and behavioural expectations that must hold for activities and outputs to plausibly generate outcomes. Assumptions on each link (e.g. between inputs and activities, between activities and outputs) are to be made clear and whether or not they are supported by evidence. With assumptions made explicit, the full programme logic is then critically reviewed for internal coherence, causal plausibility, and completeness. This review is iterative and often requires structured feedback loops with stakeholders to refine the model, challenge implicit assumptions, and ensure feasibility. Within the context of each project, the team must also consider which iterations are genuinely necessary, as the 'schoolbook approach' needs to be balanced with pragmatic considerations in order to determine what is required to arrive at a robust programme theory in a timely manner.

WHAT SHOULD BE CLEAR AT THE END OF THE MODELLING PHASE?

The result is a structured and logical map of expected changes that anchors the remainder of the modelling exercise. The modelling phase concludes when the programme theory presents a logically robust and contextually grounded explanation of how the intervention is expected to work, why it should work under specified conditions, and what assumptions underpin this reasoning.

1.5 Limitations of programme theory and logic models

While programme theory or logic models offer valuable structure for conceptualising how interventions are expected to work, they also have important limitations that may be taken into account. Graphical logic models often simplify complex, non-linear systems and may inadequately capture dynamic interactions, feedback loops, or emergent behaviours.⁶⁴ Even a well-constructed programme theory cannot fully anticipate unintended effects or the influence of shifting contextual factors.^{35, 39, 65} It does not demonstrate causality; it proposes hypothesised relationships that require empirical validation through appropriate study designs. Developing and maintaining a programme theory and logic model are also resource-intensive, requiring iterative refinement and stakeholder engagement. Finally, because programmes operate within evolving systems, programme theories must be periodically updated to reflect implementation realities, new evidence, or policy changes.



2 EXPERIMENTAL STUDY DESIGNS

KEY POINTS

- **Experimental study designs are the most rigorous approach for causal inference**, as they rely on randomisation to compare an experimental with a control group and minimise bias.
- **A clearly pre-specified study design is essential** and must balance scientific rigor with constraints such as ethics, feasibility, budget, timelines, and (where applicable) regulatory requirements.
- **Study design choices are driven by a limited set of predefined objectives and endpoints**, typically centred on a single primary endpoint used for sample size calculation, power, and the main statistical comparison.
- **Validity of experimental studies depends on statistical precision, internal validity, and external validity**, where bias mainly threatens internal validity and lack of representativeness limits external validity (generalisability).
- **Explanatory and pragmatic trials represent a continuum**: explanatory trials maximise internal validity under ideal conditions, whereas pragmatic trials prioritise real-world relevance and external validity.
- **Bias is a major threat to experimental studies** and can arise in many ways, including selection, allocation, performance, detection, follow-up frequency, attrition, operational, and reporting bias.
- **Randomisation is the cornerstone of unbiased experimental research**, ensuring baseline comparability, eliminating selection bias, and forming the basis for valid statistical inference; it must always be paired with adequate allocation concealment.
- **Randomisation methods** involve a trade-off between unpredictability, balance, and practical complexity. The choice depends on study size, the number of prognostic factors to control and practical considerations.
- **Multiple experimental designs are available**, each with strengths and limitations, including parallel RCTs, factorial designs, crossover studies, cluster randomised trials (CRTs), and stepped-wedge designs.
- **Cluster randomised designs are particularly relevant for public health and health service interventions**, where individual randomisation is infeasible, control group contamination is likely, or ethical, political, and logistical considerations require group-level allocation.
- In **stepped-wedge designs** clusters cross over from control to experimental intervention in a staggered and randomised sequence. They are often seen as a pragmatic study design that balances rigorous scientific evaluation with logistical and ethical constraints. The major drawback is probably that they tend to take longer than parallel CRTs, involving a lot of repeated measurements, increasing the burden on participants and researchers. Some of the advantages of these designs may also be accommodated within parallel cluster randomised designs.

2.1 Quantitative considerations for study design

Research topics in health service and health policy interventions are often complex and often require mixed methods, which are characterised by the integration of a qualitative and quantitative approach.⁸ This chapter aims to describe methodological considerations, important in the design stage of a study, when considering one or more key quantitative endpoints.

In this chapter, we will describe methodological considerations in experimental study designs. In Chapter 3, quasi-experimental study designs are described.

2.2 Role of experimental study design

Experimental study designs are generally considered to be the most rigorous methodological approach for establishing causal relationships between an intervention and its outcomes or 'dependent' variables. Key to experimental study designs is the presence of an **intervention**, which can be described as the manipulation of a predictor – or 'independent' – variable imposed on the population of interest.

In the context of healthcare, the concept of an 'intervention' can be interpreted very broadly. The WHO states that 'Interventions include but are not restricted to drugs, cells and other biological products, surgical procedures, radiologic procedures, devices, behavioural treatments, process-of-care changes, preventive care, etc.'⁶⁶ More generally, this document pertains to all interventions related to the health system as defined by the WHO their 2000 World Health Report as 'comprising all the organisations, institutions and resources that are devoted to producing health actions. A health action is defined as any effort, whether in personal health care, public health services or through intersectoral initiatives, whose primary purpose is to improve health'.⁶⁷

Unlike quasi-experimental studies, these studies involve **randomisation** of at least two interventional arms (or episodes), of which an experimental and a **control arm**.

Another key feature is the presence of a clear **pre-specified design**. Indeed, a well-planned, pre-specified design, is one of the most important steps before rolling out an intervention of which the impact needs to be evaluated. This entails a compromise between different, sometimes competing considerations: the study objectives, statistical rigour, budget, timelines, ethics, political context and operational feasibility. Finally, if drugs or devices are involved, there might also be regulatory requirements to be taken into account.

Study design choices are driven by a limited number of pre-specified objectives and **endpoints**. In particular, usually a single primary endpoint is identified to guide sample size determination, study power, and the primary analytical comparison between study arms. This primary comparison typically evaluates the difference in a quantitative measurement such as the mean in each arm and is intended to inform a concrete decision such as a policy implementation.

Textbox 1 – Outcome vs endpoint

The term outcome typically refers to the measured variable itself, often at the level of the individual participant (for example, a score on a depression screening tool). In contrast, an endpoint refers to the specific way that outcome is defined for analysis, often incorporating an analysis metric, time point and comparison (for example, the mean change in depression score from baseline to 6 months in the experimental arm compared with usual care). However, the distinction is often not strictly maintained. For instance, 'overall survival' is measured by the time to death from any cause. This is often referred to both as 'outcome' or as 'endpoint'.⁶⁸



2.3 Validity and statistical precision in experimental studies

Results of one specific experimental study should be interpreted while taking into account the internal and external validity of the study, and the statistical precision of the results.

Internal validity refers to the extent to which the estimated effect accurately reflects the true effect of the intervention within the study population. Threats to internal validity arise from systematic errors, commonly referred to as bias.

External validity addresses the generalisability of study findings to the target population, beyond the study sample.⁶⁹

Statistical precision reflects the degree of uncertainty associated with an estimated effect. This uncertainty arises from random sampling variability and is commonly quantified through confidence intervals. Narrow confidence intervals indicate high precision and strengthen the conclusiveness of study findings.

In general, it is insufficient to conclude that an effect is simply 'statistically significant'. Reporting exact p-values and, more importantly, confidence intervals, provides a clearer understanding of both the precision of the effect estimates and the practical relevance of the findings. This is particularly important for complex interventions, where evidence across multiple outcomes and interventions must be considered jointly.⁷⁰

2.4 Explanatory vs pragmatic studies

Explanatory trials are conducted under highly controlled conditions and often maximise internal validity but may limit applicability to routine clinical practice. They are designed to test the scientific research hypothesis whether an intervention can work under ideal, highly controlled conditions. They typically employ strict eligibility criteria, standardised interventions, intensive monitoring, placebo controls, and surrogate or intermediate outcomes. Bias is considered highly detrimental in this context.

Conversely, **pragmatic trials** intentionally embrace the less controlled, messy context of routine clinical practice and are potentially better and quicker applicable to routine clinical practice. They are designed to inform real-world decision-making. They tend to include broader patient populations, integrate interventions into usual care, collect fewer data elements, and prioritise patient-centred outcomes. In pragmatic settings, certain sources of variability and bias may be tolerated as part of reflecting real-world conditions.

Table 2 displays some typical properties of these two study types. In reality, the distinction between these study types is often less binary.⁷¹⁻⁷⁴

Table 2 – Properties of pragmatic vs explanatory studies

Common elements	Pragmatic	Explanatory
Studied effect	Effectiveness: <i>Does it work, i.e. under usual conditions?</i>	Efficacy: <i>Can it work, i.e. under ideal conditions?</i>
Conditions	Real-life, routine practice	Well-defined, strictly controlled ideal conditions
Hypothesis	Choice between existing care options (<i>Decision</i>) → Answer of immediate practical importance	Cause-effect relationship (<i>Understanding</i>) → Answer might be of less practical importance
Participants	Broad eligibility criteria, Heterogenous	Strict eligibility criteria, As homogenous as possible
Experimental arm	Integrated into usual care, tailored to patient needs	✓ Standardised, requiring additional resources (extra nurses, special visits, tight monitoring, ...) ✓ Highly controlled adherence

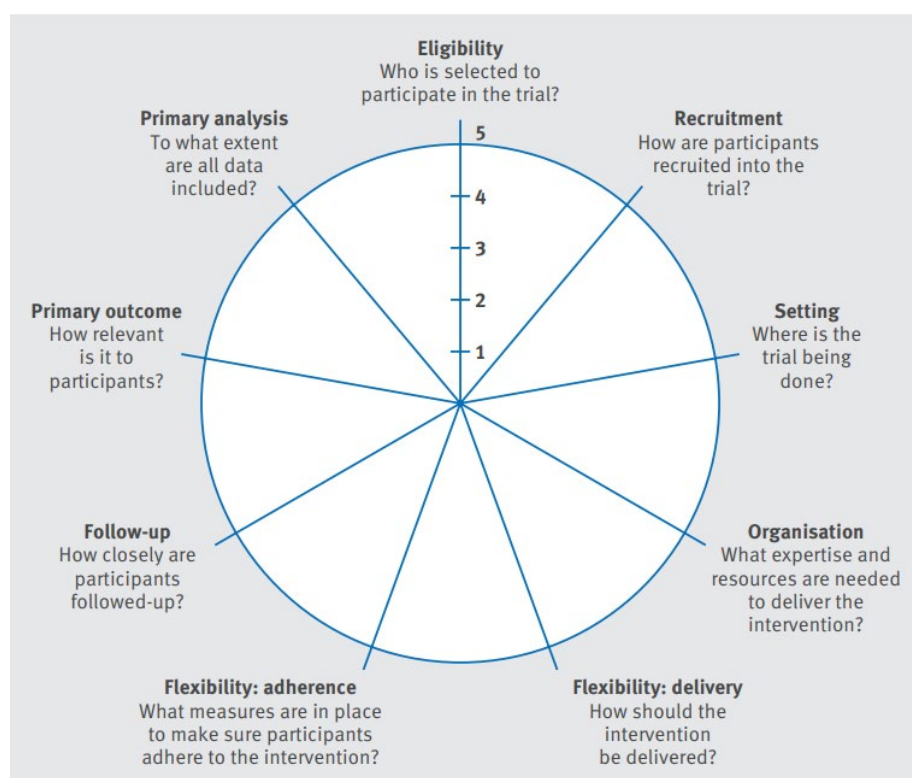
Control arm	Current standard/usual care (<i>Blinding not always possible</i>)	Placebo
Outcomes	Patient-centred	Often surrogate, intermediate
Data collected	Relatively few	Many
Follow-up	Usual care	More extensive
Validity	<i>Emphasis on external</i> Bias not necessarily detrimental, rather 'real-life'	<i>Emphasis on internal</i> Bias is very detrimental
Ethical	Equipoise?	

Turner et al. (2024)⁷⁵ highlight the complexity of carefully defining the 'usual care' control intervention. One of the main challenges is that heterogeneity among study sites may raise methodological and ethical issues.

In the context of this process note, a pragmatic study type will make sense in most cases. One major concern is that its design features may conflict with ethical standards for clinical trials, more specifically with the principle of '**equipoise**'.⁷⁶ This means that there must be genuine uncertainty within the scientific community about the preferred treatment or intervention, so that participants are not unduly harmed when their care is assigned randomly rather than directed by, for instance, their personal physician.⁷⁷ This equipoise is compromised when comparing to usual care, if prior evidence, such as findings from explanatory trials, already suggests it to be less efficacious.

The PRECIS-2 tool (Figure 4) is a validated framework that supports trial designers in aligning key design decisions with the overall objectives of a study. It encompasses nine domains that cover core elements of trial design, each rated on a scale from 1 (very explanatory) to 5 (very pragmatic).⁷⁸

Figure 4 – PRECIS-2 tool



Source: Loudon et al., 2015⁷⁸



2.5 Sources of bias

Bias is a methodological and statistical concept, meaning that the result one estimates from the data (such as a difference in means or a hazard ratio) is systematically, i.e. not explained by chance, different (too high or too low) from the real underlying value. In order to guarantee internal validity of the study results, bias should be avoided as much as possible. In clinical studies, different types of bias may arise.^{79, 80} We list the most important sources of bias, relevant in the design stage of a study (see Table 3), and techniques for controlling them. Note that bias definitions are not always consistent across various literature sources.

2.5.1 Selection and allocation bias

In the context of randomised trials, **selection bias** is typically understood as bias arising when knowledge of upcoming assignment influences the selection of particular subjects to receive a particular intervention. Either participant or healthcare provider could influence which intervention is received, for obvious reasons. **Allocation bias** is close to this definition of selection bias but refers primarily to imbalances in prognostic factors across study arms, and not in particular to 'preferential' allocation.

Mitigation strategies include robust randomisation procedures, allocation concealment, and adherence to **intention-to-treat** (ITT) principles. The former two will be discussed in the next section. The latter (ITT) means that each randomised patient is included in the analysis, and in the study arm as randomised. That means that if the 'actually received' intervention of the participant is different from the randomised, the patient is still included in the randomisation intervention for analysis.

The 'catalogue of bias' definition of selection bias is different, i.e. the bias occurring when the study population differs systematically from the population of interest, which is primarily a problem for external validity.⁷⁹ This might result in differences in participant characteristics (e.g. age, sex, severity of illness) between the trial sample and the target population and might impact the representativeness of the trial sample.

Both definitions, one relating to internal validity and the other to external validity, are important in the trial design.

2.5.2 Performance and detection bias

Performance and detection bias can occur when there is knowledge of intervention assignment among participants, researchers or care providers. **Performance bias** concerns different behaviour or performance of participants, researchers and care providers. There are differences in the care provided (other than the intervention itself; e.g. less intensive usual care given in the control arm; more advice given to patients in the experimental arm). **Detection bias** concerns differences in the outcome assessment between the experimental and control arm (e.g. researchers assess improvement more favourably in patients in the experimental arm; the physician actively looks for side effects more thoroughly in one study arm). **Blinding** participants, researchers, care providers and outcome assessors for the received intervention is an important measure against bias in clinical trials.

As noted earlier, blinding is often not feasible in pragmatic studies. This is particularly true for complex interventions where participants may play an active role, for example, by attending training sessions. However, it may still be possible to blind other parties involved, such as researchers, outcome collectors/assessors and data analysts.

When blinding cannot be achieved, it is especially important to select outcomes that can be assessed objectively. This helps reduce the risk of detection bias.

Finally, it should be noted that blinding is not always desirable, particularly in pragmatic trials, as it may reduce the extent to which study conditions reflect real-world practice.

2.5.3 Follow-up frequency bias

Differential **follow-up frequency** between arms constitutes a source of bias. For instance, one arm might have an outcome measurement each month and the other arm every two months. If time to event is measured (e.g. time to a certain drop in headache score) there will clearly be bias if the follow-up time is different. An evident solution is to use the **same follow-up procedures** for each study arm, as much as possible.

2.5.4 Attrition bias

Differential drop-out rates among study arms can disrupt randomisation. If participants with a certain risk profile in one study arm are more likely to withdraw, this might distort study results. For instance, in an oncological trial if patients drop out of the experimental arm because a drug is too toxic, the remaining cohort may look artificially successful. This kind of attrition might be foreseeable at the design stage. Careful planning and motivating participants to stay in the study is important. Also, it should be encouraged to continue collecting data, even when participants discontinue the intervention.

Sensitivity analyses, such as a 'tipping point analysis' can be necessary to plan in case differential dropout-out is observed at the analysis phase, together with a careful analysis of demographic and disease-related characteristics. A tipping point analysis evaluates the robustness of study findings by imputing missing data: i.e. estimating and filling in missing values under a range of stepwise scenarios of assumptions. For each scenario of assumptions, the imputed data are combined with the observed data and the analysis is repeated. The goal is to identify the 'tipping point', defined as the point at which the study conclusions about the intervention effect change, typically from statistically significant to non-significant. The plausibility of this tipping point helps determine whether the original study conclusions remain credible.⁸¹

2.5.5 Operational bias

Operational bias is introduced through deviations in trial conduct while the trial is running, often based on unblinded interim study results. When interim results may influence important study decisions, access to such information should preferably be restricted to an independent multidisciplinary group. This is further discussed in Chapter 4.

2.5.6 Reporting bias

This means that in the final study report measured outcomes are reported selectively. This is addressed through a clear, pre-defined prospectively registered transparent protocol and statistical analysis plan (SAP), and adherence to reporting standards.⁸² A first step in the process of KCE evaluation studies is therefore the publication of a detailed study protocol.

Table 3 – Sources of bias

Type of bias	Short description	Mitigation
Selection bias	<i>Preferential intervention allocation</i>	• Randomisation • Allocation concealment • Intention-to-treat analysis
Allocation bias	<i>Prognostic factor imbalance</i>	Randomisation techniques
Performance and detection bias	<i>Knowledge of intervention influences participants/investigators/outcome assessment</i>	Blinding
Follow-up frequency bias	<i>Differential follow-up frequency among arms</i>	Same follow-up procedures
Attrition bias	<i>Differential drop-out rates among study arms</i>	• Careful planning • Motivating participants • Sensitivity analysis



Operational bias	<i>Deviations in trial conduct due to unblinding study results</i>	• Access to unblinded interim results restricted to an independent multidisciplinary group
Reporting bias	<i>Selective outcome reporting</i>	• Study protocol • Statistical analysis plan • Adherence to reporting standards

2.6 Randomisation techniques

Randomisation is the most important methodological element to claim that a study is unbiased (see Table 4). In randomised controlled trials, allocation to either the control or the experimental arm(s) occurs randomly, ensuring that neither participants nor healthcare providers can influence which participant is allocated to which study arm. As a result, assignment is not driven by participants' characteristics, prognosis, preferences, or other underlying factors. Instead, randomisation aims to create groups that are comparable on both observed and unobserved characteristics. This reduces the risk that differences in outcomes are attributable to pre-existing differences between groups rather than to the intervention itself. This eliminates the possibility of conscious or unconscious selection bias.

This comparability is often demonstrated in the published results through a table of baseline characteristics. Also, randomisation makes sure overall balance is achieved according to the pre-specified target allocation ratio. Lastly, randomisation provides the basis for assumption-free statistical tests (randomisation and permutation tests), which can be very useful in some cases.

Ethically, randomisation is only warranted **when equipoise holds** (see above). There may also be other ethical or practical challenges associated with using randomised study designs. In such cases, quasi-experimental designs (see Chapter 3) provide an alternative when randomisation is not feasible or appropriate.

The most basic allocation ratio is 1:1, meaning that the number of participants is the same in both the experimental and control arm, in case of two arms. For three arms this would be 1:1:1, and so on. Sometimes there are good reasons to use an **unequal allocation ratio**, such as 1:2 with two times as many participants in the experimental arm. One might anticipate difficulties with recruitment or higher drop-out for one of the study arms. It might be useful to motivate people to participate if there is a higher probability of recruitment into the experimental arm. Lastly, some interventions may be cheaper than others. The downside of an unequal allocation ratio is an increase in required sample size. This is because different sample sizes in study arms increase the variance of the estimated effect, rendering the analysis less efficient. Finally, allocating more participants to one arm may raise concerns about equipoise and, consequently, about the ethical acceptability of the trial.

Randomisation alone is insufficient. **Allocation must also be adequately concealed**. Selection bias can still occur if randomisation is poorly implemented. For example, pre-printed lists of random numbers can be consulted, envelopes may be opened, and some 'static' allocation methods, such as permuted block randomisation with fixed block size (see further), can lead to predictable allocations.

The timing of the randomisation should be as late as possible, to avoid dropout. For instance, in an oncology study context, for randomisation between adjuvant and no adjuvant treatment post-surgery, it might be a good choice to randomise only after surgery and not before surgery.

We will discuss three types of randomisation: simple randomisation, permuted block randomisation with stratification and minimisation.^{77, 83-85}

Table 4 – Randomisation techniques: key points

Randomisation	Key statistical element to study design
Objectives	• Eliminate selection bias • Achieve target allocation ratio • Achieve prognostic balance • Basis for assumption-free statistical test
Ethically	Only when equipoise holds
Allocation ratio	• Equal (e.g. 1:1) vs unequal (e.g. 1:2) • Unequal: needs good motivation (equipoise?), increases sample size
Allocation concealment	Next intervention should be unpredictable
Timing	As late as possible, to avoid dropout
Types	• Simple randomisation • Permuted block randomisation with stratification • Minimisation

2.6.1 Simple randomisation

This method allocates the intervention **purely at random**, regardless of any participant baseline characteristics, basically corresponding with tossing a coin for every new candidate for the study.

This method is very easy and also completely unpredictable. However, there will inevitably be a deviation from the planned allocation ratio, i.e. an imbalance for the intervention variable. For instance, if 10 people toss a coin, the ratio of heads to tails will rarely be exactly 5:5. Also, accidental bias will occur due to imbalance in prognostic factors. For instance for sex, in a perfectly balanced trial, exactly half of the female participants would be allocated to each study arm and half to the other arm, and likewise for the male participants. The likelihood of important chance imbalances is low when the sample size is very high, like several thousand participants, but this likelihood is greater in studies of smaller size. If such an imbalance occurred for a variable with prognostic importance, this would confound the intervention effect and thus create an accidental bias, also decreasing the power of an unadjusted statistical test. **Adjusted statistical testing**, i.e. including unbalanced variables in the statistical model, can increase the power. However, in an experimental study design, one needs to pre-specify the adjustment for important baseline covariates, in order not to jeopardise the credibility of the study results.

It is usually desirable to avoid imbalances for important prognostic variables, rather than try to adjust for them by using more sophisticated analysis methods.^{84, 86}

2.6.2 Permuted block randomisation with stratification

Permuted block randomisation with stratification is a widely used method to reduce imbalances between study arms with respect to important prognostic variables.

Permuted blocks are **random sequences of intervention allocations that respect a pre-specified allocation ratio**, e.g. 1:1 for experimental (E) vs control (C) intervention. Within each block, the required balance between arms is guaranteed, while the order of allocations within the block is randomised.

In addition to these permuted blocks, **intervention allocation is stratified**, meaning that separate randomisation lists are created for different strata, which are defined by the cross-classification of the different pre-specified stratification factors. These are pre-specified categorical baseline variables that are expected to have a substantial impact on the outcome, such as type of practice, treatment centre, country, region, etc.

If there are **k prognostic factors** and **factor i has L_i levels**, the number of strata is given by $\prod_{i=1}^k L_i$. If all factors are binary ($L_i = 2$), the number of strata simplifies to 2^k . Using too many strata leads to over-stratification, in which blocks will not be completely filled, resulting in imbalances and a loss of statistical efficiency, manifested as reduced power.⁸⁷ Careful selection of a small number of prognostic



factors is therefore essential. In practice, usually only 2 or 3 factors are advisable, although in theory this is of course dependent on the sample size.⁸⁸ Stratification factors should be accounted for in the statistical analysis in order to preserve power.⁸⁶ In general, a stratified analysis results in a higher power.⁸⁹

The **block size** is the number of intervention allocations within each block, per stratum. The block size represents a trade-off between balance and predictability. Smaller block sizes achieve tighter balance but increase the risk that upcoming allocations can be guessed, potentially compromising allocation concealment. To mitigate predictability, several strategies may be used including larger block sizes, blinded block sizes and variable block sizes.

Suppose a trial with allocation ratio 1:1, stratification factors 'general practitioner' (GP) (e.g. GP1 and GP2) and 'sex' (male/female) and using a block size of 4. In this case 6 permutations are possible for each block: 'CEEC', 'CCEE', 'ECEC', 'CECE', 'ECCE' and 'EECC'. An example of randomisation lists by stratum is shown in Table 5. In GP1, the first male would then get intervention C, the second male intervention E (example highlighted in bold), and so on.

Table 5 – Permuted block randomisation with stratification

GP	Sex	List of permuted blocks			
		<i>block 1</i>	<i>block 2</i>	<i>block 3</i>	<i>block 4</i>
1	Male	CEEC	CCEE	CEEC	ECEC
1	Female	ECEC	CCEE	CEEC	CECE
2	Male	ECCE	CCEE	CCEE	EECC
2	Female	ECEC	CEEC	CECE	CEEC

A key advantage of permuted block randomisation with stratification is that **it ensures excellent balance**, both with respect to the intervention allocation ratio as to the selected prognostic baseline variables.

This approach has also limitations. Because randomisation lists are pre-generated, **the last intervention allocation in a block can be predictable**, particularly in open-label trials, i.e. where treatment allocation is not blinded. This can introduce selection bias if investigators are able to anticipate upcoming intervention allocations and selectively enrol participants. In double-blind trials, this risk is substantially reduced.

A second limitation is the **practical constraint on the number of stratification factors** that can be accommodated. Selecting the most relevant prognostic factors may be challenging and requires reliable prior scientific knowledge. Over-stratification should be avoided, as it compromises balance and statistical power.⁸⁴

2.6.3 Minimisation (dynamic allocation)

In contrast to permuted block randomisation, minimisation does not aim to achieve balance within strata, i.e. the cross-classification of prognostic factors. Instead, **balance is achieved across the (overlapping) marginal groups defined by each prognostic factor individually**. This distinction becomes clear from the example, shown in Table 6.

Suppose the next participant to be randomised is a female in GP1. The data of the participants enrolled so far are displayed in the table. Under minimisation, the next intervention allocation is chosen to minimise the overall imbalance across all prognostic factors considered. In this example, so far 9 participants have been randomised in GP1, and 16 women have been randomised. The totals in the fourth row of the table do not represent unique patients, because some participants are counted twice. The total imbalance is calculated by summing the imbalances within each factor. Here, the total imbalance equals +1, in favour of the experimental arm, indicating that allocation of this next female participant in GP1 to the control arm would minimise the overall imbalance.

Table 6 – Example of minimisation

	Control	Experimental	Imbalance
GP1	5	4	-1
Female	7	9	+2
Total	12	13	+1

In its pure form, this minimisation algorithm is deterministic, which is not desirable from a statistical standpoint, as it violates the randomness assumption for statistical testing. To avoid this, a random component is usually incorporated. Rather than deterministically selecting the intervention that minimises imbalance, allocation is made with a higher probability for that intervention.

The key advantage of this approach is that accidental bias can be mitigated for a large number of categorical prognostic factors, which is **especially useful in smaller studies**. Also, intervention allocation is largely unpredictable, even when the trial is open-label, because intervention allocation depends dynamically on the accumulated data.

However, minimisation is **more complex to implement than permuted block randomisation** and generally requires centralised randomisation. Furthermore, despite the addition of a random element, some statisticians remain concerned about the **partially deterministic nature** of the method. To increase the credibility of the results, a re-randomisation test may be performed, although adding some complexity to the statistical analysis.

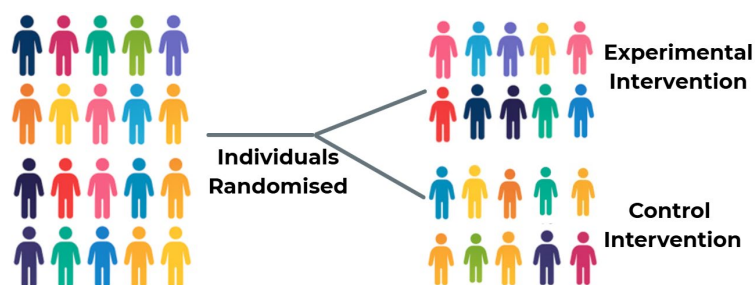
2.7 Experimental study designs

If an experimental design is feasible, it is often the best option to evaluate the effectiveness of an intervention. They include parallel arm RCTs, factorial RCTs, crossover studies, cluster RCTs and stepped-wedge study designs.

2.7.1 Parallel randomised controlled trials

In RCTs, randomisation of participants to at least two parallel arms is performed at the individual level before the intervention begins. As discussed earlier, randomisation is a key methodological feature for minimising bias in experimental studies, which is why RCTs are generally considered to provide the highest quality of evidence and to have high internal validity (see Table 7). However, in the context of this process note, **individual-level randomisation will not always be possible**.

Figure 5 – Randomised controlled trial



An important issue with individual randomisation is the risk of control group contamination. This occurs when control participants are aware of the experimental intervention and subsequently alter their behaviour, which can reduce the observable difference between study arms and dilute the estimated effect of the experimental intervention.⁶³

RCTs can be very costly and typically only feasible for well-funded projects. In addition, strict adherence to an RCT protocol may be impractical in real-world settings. For these reasons, a



pragmatic RCT (see section 2.4) rather than an explanatory one may be more desirable in many cases in this research field.

Table 7 – Strengths and challenges of RCTs

Strengths	Challenges
• Highest quality evidence • High Internal validity	• Costly → only well-funded projects • Only ethical when equipoise holds • Adherence to strict RCT-protocol, typical for explanatory RCTs, not practical • Control group contamination= information on experimental intervention leaks, influencing behaviour

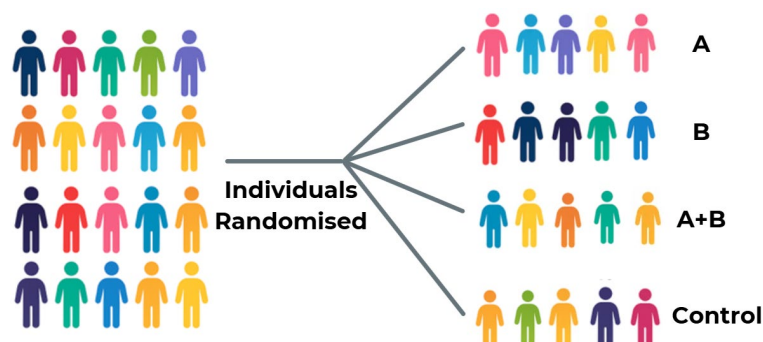
2.7.2 Factorial design

A factorial design can be used to investigate not just one, but two or more comparisons within a single trial, without the need for a higher sample size, thus optimising resource use efficiency. For example, instead of only *intervention A versus control*, a factorial design can simultaneously assess a second comparison, such as *intervention B versus control*.

When two interventions (A and B) are studied in a 2x2 factorial design, participants are randomised into four study arms:

1. Intervention A
2. Intervention B
3. Intervention A + B
4. Control

Figure 6 – Factorial design



As in a conventional RCT, each intervention – i.e. A and B – is allocated to half of the participants, with one quarter receiving the combination of both interventions. This design allows estimation of the main effect of each intervention by pooling arms as follows:

- Effect of intervention A: arms A and A+B versus arms B and control
- Effect of intervention B: arms B and A+B versus arms A and control

In this more typical application of factorial designs, the primary objective is to answer two independent research questions simultaneously: the effect of intervention A compared with control and the effect of intervention B compared with control. A key limitation here is the possibility of interaction (also called effect modification) between interventions A and B, resulting in power loss. Testing for interaction is statistically challenging as interaction tests generally have low power. Importantly, because the two main effects represent distinct primary comparisons, factorial designs do not create a multiple-

comparison problem, and adjustment for type I error inflation is not required (see further). In practice, such factorial designs are only suitable in specific circumstances and are often not feasible.^{90, 91}

Factorial designs can also be used when the primary objective is to investigate interaction effects between interventions. In this setting, an important advantage is that both the individual effects of each intervention and their combined effect can be estimated.⁹² However, since interaction tests generally have lower power, a larger sample size must be planned. The smaller the interaction effect, the larger the required sample size. For example, if the sample size is not increased, the interaction effect would need to be at least twice as large as the main effect to be detectable – an assumption that is often unrealistic in practice.⁹³ Because complex interventions often include multiple components (sub-interventions), factorial designs can be particularly useful and their use in studying complex interventions has been reported.^{94, 95}

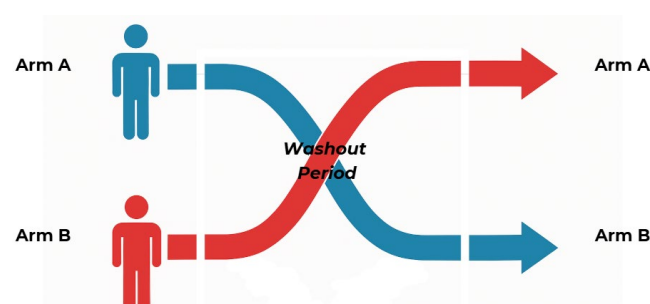
Table 8 – Strengths and challenges of factorial designs

Strengths	Challenges
Efficient use of resources	In case of two independent research questions - Requires strong hypothesis of no interaction - Power loss in case of interaction - Often not feasible In case objective is testing interaction: larger sample size needed

2.7.3 Crossover design

Another study design which is potentially more efficient than a classic RCT is the crossover design. **In a crossover study, each participant acts as his own control** (see Figure 7 & Table 9). As a result, between-patient variability is eliminated, resulting in a considerably lower sample size compared to a parallel-arm RCT. As each participant acts as his own control, confounding due to time-invariant baseline characteristics is eliminated. However, they remain vulnerable to time-related sources of bias. To address this, the sequence of interventions is randomised. The simplest design is the AB/BA sequence, in which participants are allocated to one of two randomised sequences (A followed by B, or B followed by A). This feature also points to a key limitation of this design: crossover studies are only suitable in specific situations where the condition or context is chronic and stable, outcomes are short-term and interventions can be withdrawn. Examples of this are short-acting analgesics in headache, bronchodilators in asthma, ...

Figure 7 – Crossover design



A major drawback is the risk of a 'carry-over' effect, whereby the effect of the period A intervention has a continued effect in period B. This concern extends to potential toxic or adverse effects, so participant safety requires particular attention in such designs. To minimise this risk of carry-over, a washout period is usually foreseen between two periods. This period must be long enough to allow any residual effects to wear off. When interventions involve behavioural changes in healthcare providers or patients, a crossover design is often not feasible, as such changes cannot realistically be reversed or 'unlearned' between study periods.



In the past, crossover trials were often analysed using a two-stage approach, in which a statistical test for carry-over effects was performed first. However, this strategy has been shown to produce incorrect p-values and is now considered misleading.⁹⁶

In addition to requiring a smaller sample size, crossover studies have the advantage that all participants receive the experimental intervention, which may be advantageous from an ethical, political or cultural perspective in certain contexts.

Nevertheless, crossover designs also have important challenges. As participants are exposed to multiple interventions, they may need to stay longer in the study, compared to a comparable parallel RCT, which may result in higher dropout rates. Also, withdrawing an intervention halfway, i.e. at the end of the first period, through the study may be impractical or unacceptable. Lastly, the statistical analysis can be somewhat more complex than that of parallel-group designs.⁹⁷

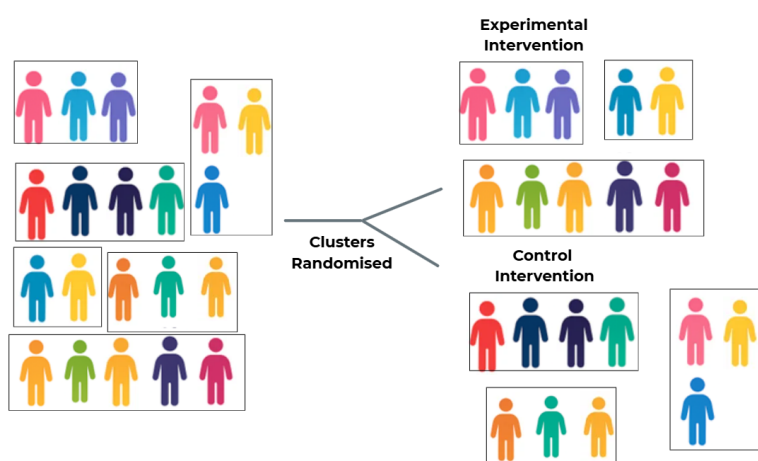
Table 9 – Strengths and challenges of crossover designs

Strengths	Challenges
• Elimination of between-patient variability eliminated allowing lower sample size • All participants receive the experimental intervention	• Potential risk of carry-over effects • Longer study duration and potentially higher dropout • Intervention withdrawn halfway through study • More complex statistical analysis • Only feasible in specific context (e.g. short-term outcomes)

2.7.4 Cluster randomised trials

In cluster randomised designs **randomisation occurs at the group level** rather than at the individual level. Examples of clusters are schools, general practitioner (GP) practices, hospitals, families or entire geographical regions. Interventions can be implemented at the cluster level, such as the introduction of an electronic clinical decision system in randomised hospitals, or at the individual level, such as a vaccination programme⁹⁸ (see Figure 8 and Table 10).

Figure 8 – Cluster randomised trials



Although clusters are the units of randomisation, outcomes are usually measured at the individual level, with participants nested within clusters. In some studies, each participant contributes only once at one single follow-up time point, i.e. the final time point, but often participants' outcomes are assessed at multiple time points throughout the trial conduct. In such cases, a longitudinal or repeated measures design is used. Based on the structure of outcome measurement, three main types of designs can be distinguished:

- In a **closed cohort study**, the same individuals are followed and – if a longitudinal study – measured repeatedly over time, producing longitudinal data. For example, in a GP

practice, a fixed group of patients with hypertension was followed within each practice and their blood pressure measured at regular intervals.⁹⁹

- In contrast, a **repeated cross-sectional study** assesses different individuals at each time point. For instance, the EPOCH trial evaluated the effectiveness of a national quality improvement program, including an integrated care pathway to reduce mortality among patients undergoing emergency laparotomy. Because patients are recruited as they present for surgery, the population of operated patients differs across time points.^{100, 101}
- Finally, some trials adopt an **open cohort design**, in which individuals may enter and leave the study over time. An example is the Depression management trial, where nursing homes were randomised to implement an intervention aimed at improving the diagnosis and management of depression. Residents were recruited at the start and additional residents were allowed to enter the study while the intervention was ongoing.^{101, 102}

Recruitment procedures may also differ between cluster trials. Clusters can be fixed at the start, which is for example typically the case for families. Alternatively, recruitment may be ongoing, with individuals enrolled continuously within clusters – for example, when regions are randomised to different vaccination programs and newborn infants enter the program over time.⁹⁸

There are many variations of cluster randomised trials (CRTs). The designs discussed earlier (parallel, factorial, crossover) can all be implemented at the cluster level. In addition, stepped-wedge designs are a specific and increasingly common form of cluster design, in which clusters crossover from control to experimental intervention in a staggered and randomised sequence.

Cluster designs are particularly well suited to addressing research questions targeting public health or health service improvement. CRTs are often used **when there is a high risk of control group contamination and when interventions are delivered at the organisational or community level**. In such cases, careful selection of the cluster unit is critical; for example, randomising nursing wards within the same hospital may still allow information spillover between groups, for instance in the cafeteria, during common training session, etc. In many contexts, individual randomisation is neither feasible nor desirable, making CRT a more appropriate choice. Practical and financial reasons may play a role, such as reduced costs when interventions are delivered by health professionals to entire hospitals or communities. Some interventions, including mass media campaigns or environmental modifications can only be implemented at the population level. Ethical, political or cultural considerations may also necessitate cluster allocation; for example, selectively providing vitamins to only some children within a village may be unacceptable.

Despite their strengths, CRTs are **more complex** than individually randomised trials. A key challenge is statistical. When outcomes are measured at the individual level, responses within the same cluster tend to be correlated, reducing the amount of independent information contributed by each participant. This loss of efficiency leads to reduced statistical power, necessitating a larger sample size. The degree of within-cluster correlation is quantified by the intraclass correlation coefficient (ICC) and will be discussed further. Complexity increases further as correlations arise both within clusters and within individuals across time in longitudinal designs. Another statistical concern is baseline imbalance at the individual level: because randomisation occurs at the cluster level, stratification cannot address imbalances at the individual level. Because of the loss of efficiency, an individually randomised RCT is preferable if the intervention can be delivered at the individual level and if the risk of contamination is low.^{63, 90, 103, 104}

Finally, CRTs pose **ethical challenges**. Because the units of randomisation and outcome measurement usually differ, this complicates the identification of research participants. Clusters are typically randomised prior to the identification and recruitment of individual participants. In such cases it is unclear who should provide consent. Recruiting participants after randomisation may also introduce selection bias, undermining the trial's validity. Additionally, when interventions are applied to clusters rather than individuals, cluster members may have little opportunity to avoid exposure, thereby weakening respect for individual autonomy and the right to refuse participation. Researchers



and ethicists have established guidelines for the ethical conduct of CRTs: the Ottawa Statement on the Ethical Design and Conduct of Cluster Randomised Trials.¹⁰⁵ A recent publication addresses some gaps in these guidelines and argues it is time for an update.¹⁰⁶

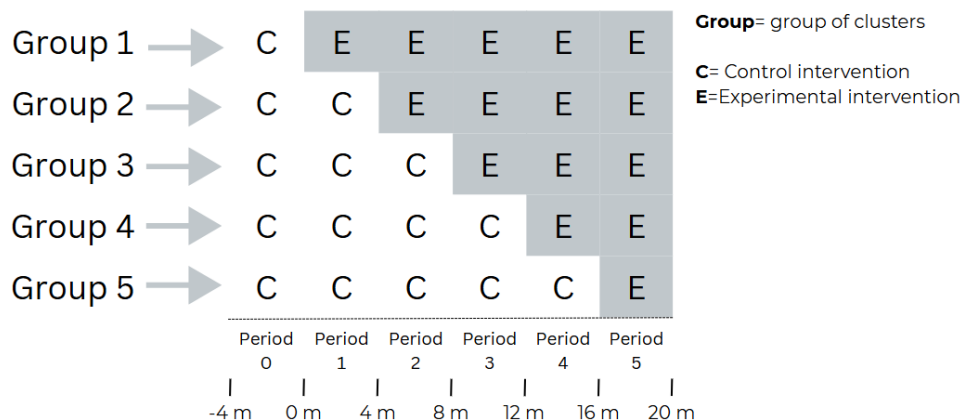
Table 10 – Strengths and challenges of cluster randomised trials

Strengths	Challenges
- Addresses research questions public health and health service interventions - Avoids control group contamination - Particular reasons - Financial - Practical - Ethical, political and cultural acceptability	- Within-cluster correlation → decrease statistical power - Baseline imbalance at the individual level - Ethical challenges

2.7.5 Stepped-wedge designs

A **stepped-wedge (S-W) trial** is a variation of a longitudinal cluster randomised trial (CRT). After an initial baseline period during which no cluster receives the experimental intervention, the intervention is sequentially rolled out to clusters. Clusters – or groups of clusters – are randomised to the timing of their switch from control to experimental. Once a cluster begins the experimental intervention, it continues to receive it until the end of the trial. Consequently, by the end of the study all clusters are exposed to the experimental intervention. Just as in a crossover study, there may be a washout or ‘transition’ period between the control and experimental interventions¹⁰¹ (see Figure 9 and Table 11).

Figure 9 – Stepped-wedge design



By design, S-W trials are longitudinal, with data collection throughout the study. Each cluster therefore contributes under both control and experimental interventions. Similar to a crossover design, each cluster acts as his own control. In addition, because at one given time point some clusters are exposed to the experimental intervention while others are still under the control intervention, the statistical analysis can combine within-cluster with between-cluster comparisons.

As a form of cluster design, S-W trials share many of the strengths and challenges of CRTs. They are particularly useful **when it is more feasible and efficient to implement an intervention gradually**, for example because of limited resources and logistical constraints. A S-W design may also be preferable when a crossover design is not feasible, more specifically **when intervention effects are irreversible after switching interventions**. In contrast to a parallel CRT, all clusters are guaranteed to receive the experimental intervention during the study period, just as in a crossover trial. This may be an incentive for participation. When there is strong belief in the experimental intervention's effectiveness – often based on earlier evidence – it may be ethically, politically or culturally desirable

that all clusters ultimately receive this intervention. S-W trials are therefore often seen as a pragmatic study design that balances rigorous scientific evaluation with logistical and ethical constraints.

From a statistical perspective, S-W designs are complex, both in terms of sample size calculation and analysis. Although the timing of switching is randomised, the intervention is still partially confounded with time, because all clusters follow the same directional sequence from control to experimental intervention, unlike in a true crossover design. Implementation can also be challenging, as clusters may not always be operationally ready to switch at their allocated time point. One of the key characteristics of S-W trials is that they tend to take longer than parallel CRTs, involving a lot of repeated measurements, increasing the burden on participants and researchers. These features may increase the risk of control group contamination, dropout, intercurrent events (see further) and compromised data quality.

As in a crossover study design, within-cluster pre-post comparisons are part of the statistical analysis. This results potentially in a lower sample size. Therefore, S-W designs can be an appropriate choice when there is only a limited number of clusters available.¹⁰⁴

It should be noted that some of the advantages commonly attributed to S-W designs, can also be accommodated within parallel CRTs. For example, clusters in a parallel CRT may be offered the experimental intervention after study completion, and sequential implementation of the experimental intervention is also possible without using a S-W design. Consequently, the choice between a S-W and parallel CRT requires careful consideration.^{63, 90, 101, 107-110}

Table 11 – Strengths and challenges of stepped-wedge trials

Strengths	Challenges
• Same as CRTs • More feasible and resource-efficient than CRT • Potentially lower sample size than parallel CRT, feasible when limited number of clusters available • All clusters receive the experimental intervention ○ May incentivise recruitment ○ Ethical, political and cultural acceptability • Alternative when crossover design not feasible	• Same as CRTs • Statistically complex ○ Sample size calculation ○ Analysis • Partial confounding of intervention effects with time • Operationally challenging to implement • Typically longer duration than parallel CRT ○ A lot of repeated measurements ○ Long wait for experimental intervention ○ Increased risk of dropout, intercurrent events and contamination

2.7.6 Other study designs

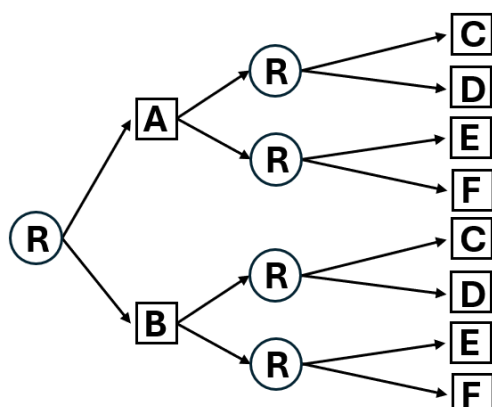
In addition to the study designs discussed above, many other designs exist. We briefly highlight some of these.

2.7.6.1 SMART

The Sequential Multiple Assignment Randomised Trial (SMART)¹¹¹ designs allow participants to be re-randomised, based on their response to an initial intervention, thereby enabling the evaluation of **dynamic treatment regimens (DTRs)**. In such designs, treatment decisions at later stages are guided by a prespecified set of decision rules that depend on intermediate outcomes. As a result, participants may undergo more than one randomisation. An example of such a SMART design is illustrated in Figure 10 where 'R' denotes randomisation.



Figure 10 – SMART design



In the example described by Piantadosi et al (2022),¹¹¹ a SMART design was used in the context of an intervention for alcoholism. In the first stage, patients were given the same treatment and randomised to one of two definitions of non-response: a more stringent definition (A) and a more lenient one (B). Depending on whether patients were classified as responders or non-responders to the treatment (in the example this was naltrexone + medical management) under these two different definitions, they were subsequently re-randomised to different treatment options in the second stage.

A common primary objective in SMART trials is to compare the effects of the initial intervention options (A vs B in Figure 10), averaged over the subsequent interventions. This is very similar to the main effect estimation in factorial study designs, with the difference that randomisation is sequential, as well as the possible restriction of intervention options according to response. This means that standard statistical methods can be used. A second objective is often the comparison of the second stage intervention options among responders (C vs D in Figure 10) or non-responders (E vs F in Figure 10), averaged over the initial intervention assignment. Again, standard statistical methods can be used. Other schemes are possible and other comparisons can be of interest.

2.7.6.2 Trial within Cohorts (TwICs)

The TwICs designs¹¹² aim to address limitations of standard RCTs, such as slow recruitment, limited generalisability, and dropout, by embedding trials within an existing cohort. Because eligible patients are identified and followed through this cohort, recruitment can be faster and patients are potentially more representative – i.e. if the cohort is representative for the target population – increasing external validity. Only those selected for the intervention are informed and asked to consent. However, this design also introduces challenges, including the timing of randomisation and treatment refusal in the intervention group, which affects interpretation of results. In particular, the treatment effect reflects the impact of offering the intervention rather than receiving it. Also, not all interventions are suitable for this design, especially if they involve high burden or safety concerns, as the non-compliance then may be too large.

3 QUASI-EXPERIMENTAL DESIGNS

KEY POINTS

- **Quasi-experimental studies evaluate the effects of interventions without random assignment.** They may use observational data or prospectively collected data following a non-randomised intervention.
- **Without randomisation, quasi-experimental studies are vulnerable to confounding and selection bias, making causal interpretation challenging.** Specific design and analytical strategies are therefore required to strengthen internal validity.
- **Different quasi-experimental designs rely on different assumptions, and their validity depends on how plausible these are.** The choice of design should be guided by the research question and the structure of the available data.
- **Time series-based approaches and regression discontinuity designs typically have the highest internal validity** (provided that their assumptions are met and they are appropriately applied), whereas single-group pre-post and controlled post-test only designs provide the weakest evidence.
- **The causal interpretation of quasi-experimental findings relies on counterfactual reasoning** about what would have happened in the absence of the intervention.
- **Under strict assumptions, including the absence of unmeasured confounding, causal conclusions may be drawn from quasi-experimental studies.** Developing a directed acyclic graph (DAG) is a key preparatory step to make these assumptions explicit and guide the analysis.

3.1 Introduction to quasi-experimental designs

3.1.1 From experimental to quasi-experimental designs

In experimental studies, researchers control both the intervention and the assignment of study units to intervention and control conditions. As discussed in the previous chapter, random assignment provides a clear basis for comparing outcomes across groups and for attributing observed differences to the intervention rather than to pre-existing group characteristics. This feature is key to the high internal validity of experimental designs.¹¹³

In practice, however, random assignment is not always feasible. In studies where researchers initiate an intervention, ethical constraints and practical or logistical barriers may limit their ability to randomise who receives it or when it is implemented.¹¹³⁻¹¹⁵ In the KCE evaluation programme, it will often not be possible to randomise subjects for one of the aforementioned reasons (which should be justified). Yet, causal evaluation may still be pursued using quasi-experimental designs (definition in Textbox 2), in which researchers evaluate interventions or exposure contrasts arising from routine practice or policy decisions.

Textbox 2 – Definition of quasi-experimental designs

A quasi-experimental design is defined as a research method used to evaluate interventions or establish causal relationships by manipulating an independent variable, while lacking random assignment of participants to groups.¹¹⁶



Without randomisation, intervention and comparison groups cannot automatically be assumed to be comparable. Patients receiving a certain treatment may differ from those who do not and policy interventions may be implemented preferentially in settings with particular needs or characteristics. These non-random allocation processes may introduce **selection bias**, whereby the intervention and comparison groups differ systematically because assignment is related to patient or setting characteristics. When these differences are also associated with the outcome of interest, they act as **confounders**, causing observed outcome differences to reflect both the intervention effect and underlying differences between groups. As a result, the **internal validity** of the study (i.e. the extent to which observed effects can be attributed to the intervention) may be threatened. Quasi-experimental methods are specifically designed to address these challenges by using design and analytical strategies to reduce bias and strengthen causal inference in the absence of randomisation.

Quasi-experimental methods also differ from randomised experiments in the **causal estimands** (i.e. the specific causal effect that a study aims to estimate) they estimate. In randomised experiments, the **average treatment effect (ATE)** and the **average treatment effect among the treated (ATT)** can be expected to coincide because randomisation ensures exchangeability between intervention and comparison groups. In quasi-experimental studies, however, the choice of estimand becomes more consequential, as different study designs may identify different causal effects. While some designs target population-wide effects such as the ATE or ATT, others estimate effects for specific subgroups of the population, in which case we talk about a **local average treatment effect (LATE)**. These concepts will be revisited later in the chapter in the context of causal inference (see section 3.4.2.3). While LATE estimates can have strong internal validity, their interpretation requires careful consideration, as they may not generalise beyond these design-specific subpopulations.¹¹⁷

Just like experimental studies, quasi-experimental studies can evaluate **researcher-initiated quasi-experiments**, while lacking random allocation. In the KCE evaluation programme, interventions are typically designed and implemented by the commissioning body before the evaluation, making randomisation often infeasible. Nevertheless, the KCE research team can influence the evaluation design and must carefully consider how the absence of randomisation affects internal validity. Non-random assignment to intervention and comparison groups may be driven by recruitment and implementation considerations, such as administrator selection based on professional judgement, or voluntary participation. For example, highly motivated individuals may be more likely to participate in a new programme, or administrators may preferentially select those considered most likely to benefit, potentially leading to an overestimation of the intervention effect. Conversely, interventions may be targeted towards individuals with the most severe needs, potentially resulting in an underestimation of the intervention effect.^{116, 117}

In contrast to experimental studies, the scope of quasi-experimental research extends beyond planned interventions and also includes **naturally occurring quasi-experiments**. These may involve **natural experiments**, referring to situations where an intervention occurs independently of the researchers, such as policy changes or the introduction of new programmes.¹¹⁸ In other cases, a concrete intervention may be absent, but contrasts in exposure levels can be interpreted as reflecting an implicit or hypothetical intervention, such as differences in policy intensity, eligibility thresholds, or implementation timing across settings. These designs are often practical and cost-efficient, particularly when based on routinely collected administrative or population-level data, and may provide pre-intervention measurements that enable baseline comparisons (see also Chapter 11 on the use of administrative data).¹¹⁹ By evaluating interventions as they occur in routine settings, they can generate evidence that reflects real-world public health and health system practice.¹²⁰ However, because researchers have limited control over implementation and assignment, these designs may be particularly vulnerable to selection bias and confounding.^{118, 120} Target trial emulation (elaborated upon in Textbox 3) provides one framework to strengthen causal inference for quasi-experiments using observational data.

Textbox 3 – Target trial emulation

Target trial emulation (TTE) is a framework for strengthening causal inference in situations where randomisation is not feasible.¹²¹ At its core, **TTE uses the protocol of a hypothetical randomised trial as a blueprint for the design and analysis of an observational study.**

Conceptually, TTE consists of two steps: (1) specifying the target trial that would ideally answer the research question, and (2) emulating this trial as closely as possible using observational data. In practice, an intermediate step is often required in which researchers define a feasible target trial that reflects the constraints of the available data and study setting before proceeding to the emulation stage.¹²² The target trial protocol typically specifies the study population and eligibility criteria, intervention and comparison conditions, assignment procedure, follow-up period, outcome, causal contrasts of interest (or estimand), and analysis plan.¹²³

The rationale for TTE is that observational studies are often designed around available data rather than around the causal question of interest. As a result, key design features such as eligibility criteria, intervention and comparison conditions, follow-up, outcomes and the causal effect of interest may remain implicit or be defined inconsistently. TTE addresses this by requiring researchers to **systematically consider, define and align these design elements before the analysis** is conducted. The structured approach imposes transparency and consistency on observational study design, preventing common design flaws resulting in avoidable biases (e.g. immortal time bias^c) and supporting clearer interpretation of causal effects.¹²⁴

Although initially developed for clinical and epidemiological research,¹²⁵ TTE is increasingly being applied in health policy evaluation, with recent methodological work providing guidance on its use in this context.¹²⁶ An example of such an application is the work of Andrew et al.¹²⁷ Importantly, TTE is intended to **complement rather than replace randomised controlled trials** (RCTs), which remain the gold standard for causal inference when feasible. TTE can be particularly valuable when RCTs are impractical, unethical, or expensive; when longer follow-up is required than is feasible within a trial; or when comparing interventions that have only been studied separately. In settings where RCT evidence exists, TTE findings can be benchmarked against trial results over the period of overlap, thereby increasing confidence in estimates derived from longer-term follow-up or real-world populations.

Several principles derived from TTE are incorporated into the practical guide for designing quasi-experimental studies presented in section 3.6.5.

3.1.2 Outline of this chapter

First, the upcoming two sections will discuss quasi-experimental designs, which can be divided into two main design families:

- **Designs for comparison-based effect identification** (section 3.2): Designs that estimate intervention effects by comparing outcomes between a group receiving the intervention and a group not receiving it. The counterfactual is approximated through differences between groups, using design or analytical strategies to address baseline differences.
- **Designs for before-and-after effect identification** (section 3.2): Designs using temporal aspects for effect identification, can be divided into two subgroups:

^c Immortal time bias arises when exposure status is assigned in a way that requires participants to survive long enough to receive the exposure, leading to an artificially inflated estimate of intervention benefit.⁷⁹ For example, defining exposure as receiving a care plan within 12 months after stroke guarantees that treated patients must survive long enough to receive the plan, creating an artificial survival advantage.



- **Simple before-and-after effect identification:** Designs that estimate intervention effects by comparing outcomes within the same population before and after the intervention. The counterfactual is approximated by assuming that pre-intervention outcomes reflect what would have happened without the intervention.
- **Time-structure-based effect identification:** Designs that estimate intervention effects by exploiting the timing of intervention implementation in relation to repeated outcome measurements. The counterfactual is approximated either by extending observed pre-intervention trends or by comparing changes across units that receive the intervention at different points in time.

For both design families, common sources of bias are discussed first to clarify the main threats to causal interpretation in that setting. Next, the general design types within each family are described, outlining how counterfactual comparisons are approximated in practice. Finally, strategies to strengthen internal validity are introduced, focusing on design features and analytical approaches that improve the plausibility of causal assumptions.

When all designs have been covered, section 3.4 addresses the transition **from association to causation**, the principles that support causal reasoning, the assumptions required for causal interpretation, and how to assess causal relationships and identifying sources of bias. This is followed by section 3.5 on analytical techniques within the causal inference framework.

Finally, section 3.6 presents a **practical guide** for developing quasi-experiments from research question to study protocol.

3.2 Designs for comparison-based effect identification

3.2.1 Common types of bias

Comparison group designs are widely used to estimate the effect of an intervention by contrasting an intervention group with a comparison group. Although this approach is stronger than single-group before-and-after designs, it is still vulnerable to several threats to internal validity. Differences between groups can arise for reasons unrelated to the intervention, and these can distort the estimated effect. Five common sources of bias are described below.

- **Selection bias**^{128, 129}

Groups may differ systematically at baseline in ways that influence the outcome. These differences can lead to biased conclusions about the intervention.

Example: Women who participate in breast cancer population screening are often healthier and more engaged with preventive healthcare than non-participants, making screening appear more effective than it truly is.

- **Attrition bias**^{128, 129}

Unequal dropout rates between groups can undermine comparability because differences emerge over time. Even when groups were equivalent at baseline, this can create an imbalance of characteristics between the intervention group and the control group.

Example: Participants may leave the control group because they feel neglected, creating an imbalance that did not exist at baseline.

- **Spillover or contamination bias**¹³⁰

Individuals in the control group may be unintentionally exposed to the intervention or its derived effects. This reduces the contrast between groups and weakens the ability to detect the true effect.

Example: Control group members adopt new behaviours after participants from the intervention group share tips or insights from a training programme that was part of the intervention.

- **Detection bias**¹²⁸

Outcome assessment may differ across groups due to researcher expectations or variations in how measurements are administered.

Example: Researchers might measure outcomes more thoroughly or with greater attention in the intervention group than in the control group.

- **Performance bias**¹²⁸

Groups may receive different levels of attention, support, or resources that are not part of the intervention but that still influence the outcome.

Example: Staff members may encourage or support participants in the intervention group more actively than those in the control group, which was not part of the intervention.

3.2.2 Basic designs

3.2.2.1 Controlled post-test-only design

Controlled post-test-only designs (illustrated in Figure 11) compare an intervention group with a control group without measuring the outcome prior to the intervention.¹¹³ Participants are assigned to the two groups through a non-random process, which means that **important baseline differences may exist but remain unobserved**. The effect of the intervention is estimated by comparing post-intervention outcomes, but this comparison relies heavily on the untested assumption that the groups were similar before the intervention.¹¹³ This assumption cannot be verified for unmeasured variables and can only be assessed using characteristics measured after the intervention, which means it can only be checked for characteristics that are not influenced by the intervention. Thus, **prior knowledge about the likelihood of comparability of the groups** with respect to the outcome is **required**.¹³¹ This may require both insight into the participant selection procedure as well as domain knowledge. Given their fundamental limitations in establishing causal relationships,¹¹³ controlled post-test-only designs should be reserved for situations where pre-intervention measurements cannot be obtained. Table 12 summarises indications and contra-indications related to the use of this design.

Figure 11 – Illustration of the controlled post-test-only design

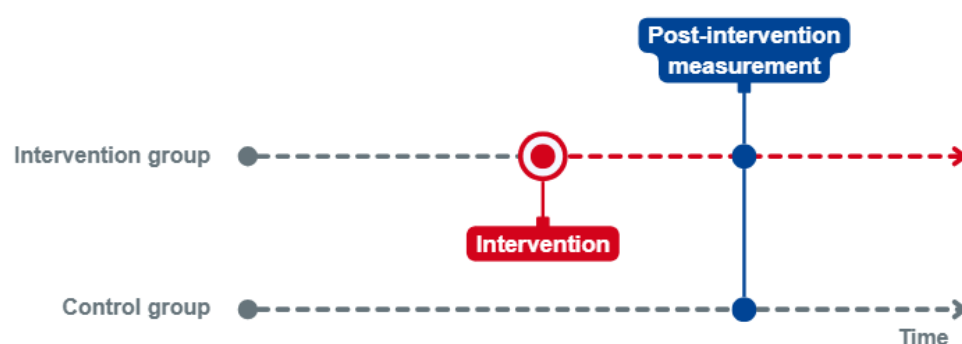
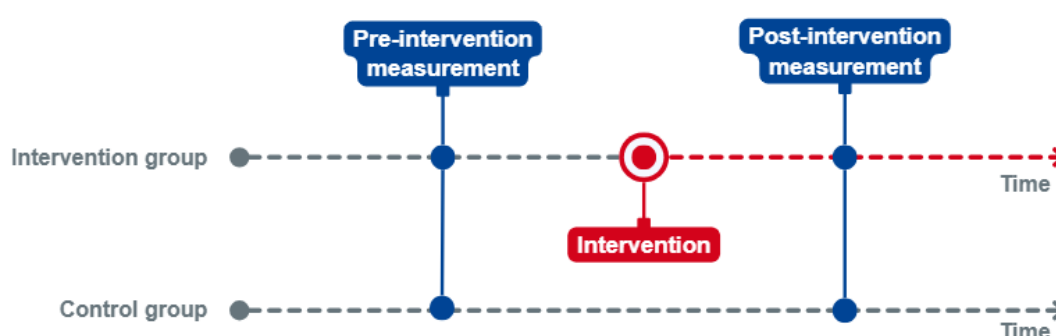


**Table 12 – Indications and contra-indications to the use of the controlled post-test-only design**

Indications	Contra-indications
- When pre-intervention measures are unavailable, such as in sudden or externally imposed interventions. - When there are good reasons to believe the groups to be comparable. - When group assignment is unlikely to be influenced by factors related to the outcome, and stable pre-exposure characteristics can be used to support group comparability (e.g. matching). - When the composition of the intervention and control groups remains stable over time, and the intervention is unlikely to influence outcomes in the control group through biases such as spillover.	- When baseline differences are expected or likely. - When pre-intervention measurements are available or can be reconstructed. - When outcomes are expected to fluctuate considerably over time. - When assignment to exposure depends on unmeasured factors that also influence the outcome. - When concurrent events or policies may differentially affect the groups.

3.2.2.2 Controlled pre-post design

A controlled pre-post design (depicted in Figure 12) is often described as the quasi-experimental approach that most closely resembles a randomised controlled trial, although it does not involve true random assignment. It estimates the impact of an intervention by comparing how outcomes change over time in an intervention group relative to a comparison group. This way, confounding due to temporal trends that would have occurred in the absence of the intervention is minimized.¹³² Therefore, the controlled pre-post design is superior to the post-test-only design. The measurement setup before and after the intervention in both the intervention and control groups (as illustrated in Figure 13) allows for a **difference-in-differences** (DiD) analysis.

Figure 12 – Illustration of the controlled pre-post design

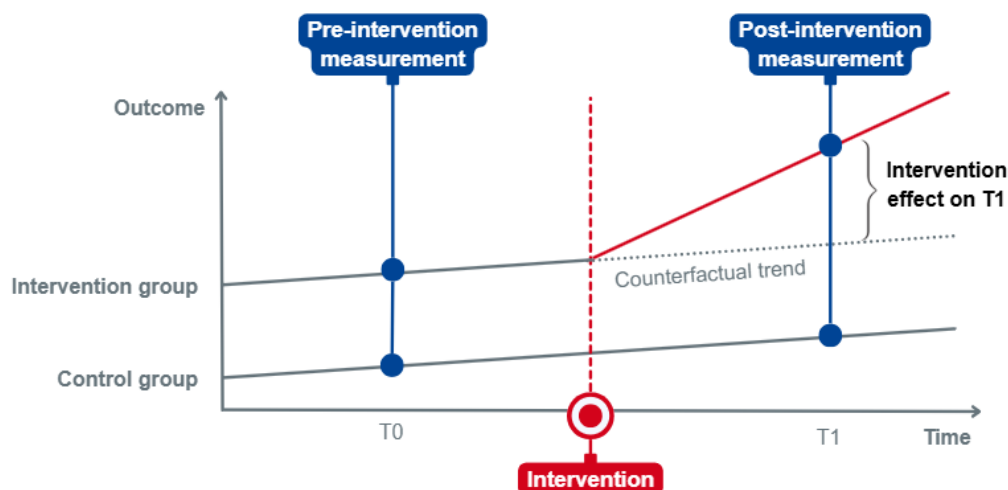
DiD is one of the most commonly used methods for evaluating the effects of policies.¹³³ The core identifying assumption of DiD analysis is that, in the absence of the intervention, both groups would have followed similar trends before the policy or programme was introduced, implying that common temporal trends and unexpected events would have affected the intervention and control groups equally.¹³⁴ When this parallel trends assumption is plausible, the difference in change between groups can be interpreted as the effect of the intervention. In its simplest form, the DiD impact can be calculated as a simple subtraction:¹¹⁵

$$\text{DiD} = (\text{Outcome T1} - \text{Outcome T0})_{\text{intervention}} - (\text{Outcome T1} - \text{Outcome T0})_{\text{control}}$$

Rather than relying on direct subtraction, the effect is more commonly estimated using a regression-based approach that includes an interaction term between time and intervention allocation.^{134, 135} This

method ensures a more flexible estimation of the intervention's impact, enabling adjustment for covariates, informing on the uncertainty of the estimation, and accommodating the complexity of the study design.

Figure 13 – Illustration of a difference-in-difference analysis



The key strength of DiD is that it **controls for unobserved confounding** arising from time-invariant differences between groups and from time-varying unobserved factors as long as the parallel trends assumption holds.¹¹⁸ Thus, DiD improves internal validity by accounting for baseline differences between groups and common temporal shocks.

Measured characteristics can be addressed using adjustment techniques discussed in section 3.2.3, but as the parallel trends assumption cannot be directly verified for unmeasured characteristics, internal validity ultimately rests on its plausibility rather than empirical proof.^{63, 115} When outcome data are available for multiple pre-implementation time points, outcome trends in the intervention and comparison groups can be examined directly to evaluate whether they evolve similarly prior to the intervention.^{118, 135} Alternatively, a placebo test using a group or outcome that should not be affected by the intervention should yield a zero effect, suggesting bias due to underlying differences in trends otherwise.¹¹⁷ Another approach is estimating the treatment effect using different control groups and verify that the results remain broadly consistent, which supports the equal trends assumption.¹¹⁷

The potential sources of bias discussed earlier in section 3.2.1, such as spillover, may undermine the validity of the DiD estimate.^{134, 136} It is important to note that the sources of bias due to repeated measurements over time (as will be discussed in section 3.3.1) are also relevant in difference-in-differences designs. However, to the extent that they affect the treated and control groups similarly over time in the absence of the intervention, their impact is accounted for under the parallel trends assumption. Finally, consideration should be given to the possibility of **anticipation effects**, which could undermine the design's core logic.^{135, 136} This refers to the situation where the pre-measurement is contaminated by the intervention itself, for example because participants adjust their behaviour because of information on an upcoming intervention.¹³⁶ Conducting pre-measurements before the time window of anticipation (e.g. before the public announcement date) is a potential way to handle this issue. An overview of indications and contra-indications for the use of the controlled pre-post design arising from these potential threats is presented in Table 13.

Finally, Dahabreh et al. (2026)¹³⁷ provide a useful overview of DiD, including several key elements to be reported for DiD studies.

**Table 13 – Indications and contra-indications to the use of the controlled pre-post design**

Indications	Contra-indications
- When pre- and post-measurements are available/possible in both groups. - When an intervention occurs at a clearly defined point in time, rather than being phased in gradually - When the composition of the intervention and control groups remains stable over time, and the intervention is unlikely to influence outcomes in the control group through biases such as spillover.	- When parallel trends between intervention group and control group are implausible, for example because external events or co-interventions differentially affect the groups. - When anticipation effects are likely to influence outcomes during the pre-intervention period in ways that cannot be clearly identified and accounted for.

3.2.2.3 *Difference-in-differences analysis using repeated cross-sectional data*

When repeated cross-sectional data are available for both an intervention group and a suitable comparison group across time, they can be analysed using a DiD framework.¹³⁸ In this setting, DiD exploits differences in outcome changes over time between exposed and unexposed populations to separate intervention effects from time-invariant differences between groups and from time trends affecting both groups similarly.¹³³

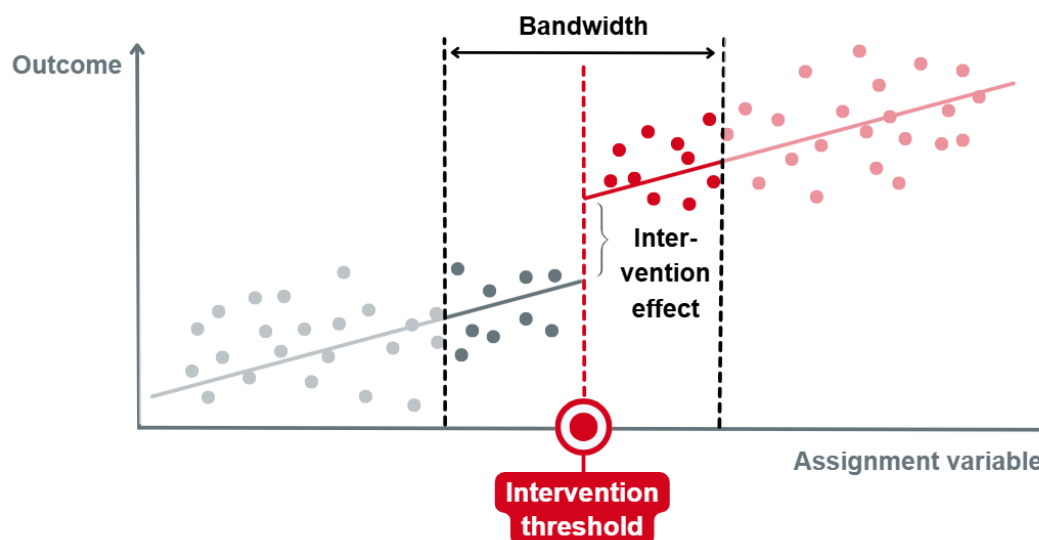
Although the same individuals are not followed over time, repeated cross-sectional DiD can support effect estimations at the population level, provided that the parallel trends assumption is plausible and that sampling and measurement are comparable across groups and time points.¹³⁸ This assumption is central, as differences in population composition across waves may otherwise confound estimated effects. To further address such compositional changes, methods that combine outcome modelling with weighting based on observed covariates may be used (see section 3.5.3).¹³⁹

When repeated cross-sectional data are used in a DiD framework, several of the limitations outlined in section 3.3.2.2 still apply, particularly those related to sampling variability, changes in population composition across waves, and measurement consistency. However, DiD accounts for time-invariant group differences and common time trends, reducing bias from baseline differences and shared external shocks. The **key remaining concern** is whether the intervention and control group would have followed **parallel trends in the absence of the intervention**, as violations of this assumption can bias results.

3.2.2.4 *Regression discontinuity design*

Regression discontinuity designs (RDD) (presented in Figure 14) are used when access to an intervention is determined by a **clearly defined threshold**, such as an income threshold for extra reimbursement, or a risk score that determines eligibility for intensive follow-up or case management. Individuals just above the threshold receive the intervention, while those just below do not. This creates two groups expected to be very similar in all respects except for the intervention assignment. Because intervention status changes sharply at the cut-off, while other characteristics are assumed to evolve smoothly, **individuals just above and just below the threshold are considered comparable except for the intervention status**. Thus, the effect of the intervention can be estimated as the difference in average outcomes between individuals just above and just below the cut-off.¹¹⁸

Figure 14 – Illustration of the regression discontinuity design



RDD is often applied in case of natural experiments and is particularly useful in health policy settings because many eligibility rules and clinical decision processes rely on formal thresholds.¹³² An important requirement is that this **cut-off is exclusive to the programme of interest**, so that crossing the threshold does not simultaneously affect eligibility for other interventions.¹¹⁷

While RDD is most often used in contexts where eligibility is determined by a threshold based on an individual characteristic, this is not its only possible application. In some settings, calendar time itself can serve as the assignment variable, with the intervention date acting as the cut-off.¹⁴⁰ In these cases, the effect is identified by comparing outcomes within a narrow time window immediately before and after implementation, under the assumption that individuals observed close to the cut-off are comparable and that no other factors change discontinuously at the same point in time. Because this comparability is only guaranteed locally, confidence that the estimated effect reflects the intervention decreases as the time window is extended further from the cut-off. In contrast to the time series methods discussed later in section 3.3.4, inference is therefore focused on the **effect at the moment of implementation** rather than on changes that evolve gradually over longer periods.

On a more practical level, the margin around the threshold is chosen by gradually widening a narrow window and checking the balance on observed characteristics at each step, stopping when meaningful differences appear.¹¹⁵ This balance check serves as an empirical diagnostic, with the key identifying assumption that individuals just above and below the threshold are comparable in both observed and unobserved characteristics, because their positions relative to the threshold are essentially determined by quasi-random variation. The choice of margin width involves a trade-off between statistical power (a large margin around the threshold for large samples) and avoiding confounding (a narrow margin around the threshold to ensure comparability).¹¹⁸ After determining the margin, the intervention effect is obtained by regressing the outcome on the assignment variable (the smooth relationship between the assignment variable and the outcome) and an indicator of intervention exposure (the discontinuous jump associated with receiving the intervention), with the coefficient on the indicator representing the intervention effect.¹¹⁸ This estimation is restricted to observations within the selected margin around the threshold, so that comparisons are made only among individuals who are locally comparable.

The key strengths and limitations of RDD arise directly from their reliance on a clearly defined threshold for group assignment. The main advantage of RDD is that comparisons close to the threshold are often described as mimicking randomisation, especially when there is some degree of random error involved.¹¹⁸ Thus, RDD addresses unobserved characteristics more convincingly than most other quasi-experimental approaches, because equivalence is assumed locally around the



threshold rather than constructed through adjustment.¹¹⁵ Therefore, RDD is widely regarded as **one of the strongest quasi-experimental approaches** in an observational context, as it can closely approximate randomised controlled trials under well-defined assumptions.¹⁴¹

The main limitation of RDD follows directly from the fact that it identifies a **local average treatment effect (LATE)** that is valid only for individuals located very close to the threshold. The effect for individuals farther from the cut-off may differ, which limits the generalisability of RDD findings to the broader population and can reduce their relevance for assessing overall programme effectiveness.^{115, 118, 140} In practice, comparisons between LATEs and ATEs have not shown large differences,¹¹⁵ but this remains an important conceptual limitation nonetheless.

RDD comes with **considerable data challenges**. The need for a large number of observations near the threshold on both sides to support reliable effect estimation can be an issue.¹¹⁵ Moreover, data on individuals who were not enrolled in the programme are often not systematically recorded.¹¹⁵ Because of these requirements, RDD is often used on large routinely collected (administrative) datasets.¹¹⁵ Finally, RDD requires a sharp cut-off (illustrated in Figure 15), while real-world decision-making sometimes entails exceptions or discretion. However, a fuzzy cut-off occurs when eligibility rules are not applied perfectly, so some individuals just below the threshold still receive the intervention (e.g. because the eligibility criteria are tampered with to provide additional support for patients) and some just above it do not (e.g. people who qualify for certain services may not be aware). This fuzziness can be handled analytically by using threshold-based assignment, meaning the intervention group a participant would belong to under perfect compliance, as an instrumental variable (instrumental variables will be covered in section 3.5.4).¹¹⁸ While limited noncompliance can be addressed in this way, the validity of the design still depends on the threshold's strength as a determinant of the intervention. If crossing the threshold barely affects receiving the intervention, the design effectively becomes weak and internal validity is compromised.¹⁴² These limitations are, amongst others, considered in Table 14).

Figure 15 – Illustration of a sharp (left panel) and a fuzzy (right panel) cut-off

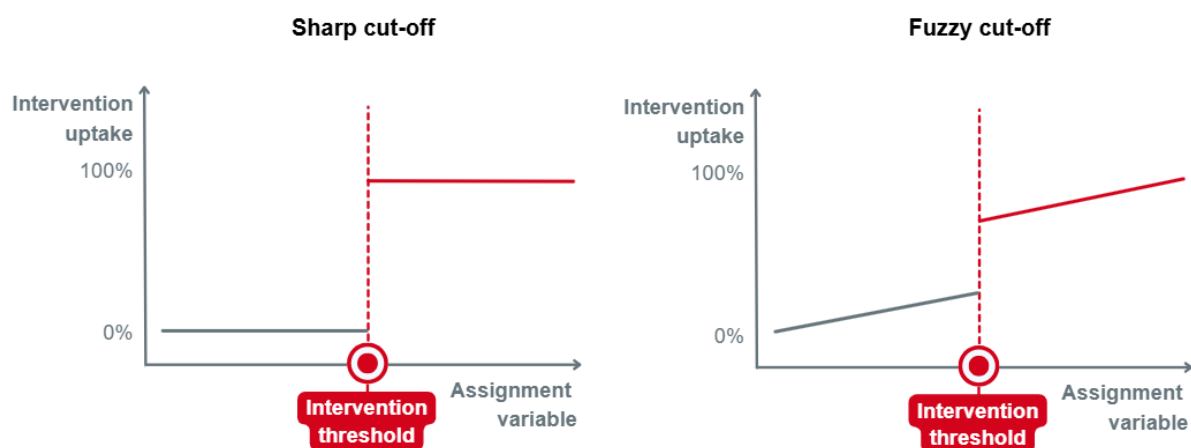


Table 14 – Indications and contra-indications to the use of the regression discontinuity design

Indications	Contra-indications
- When intervention assignment depends strictly on a cut-off through a policy eligibility rule and sufficient data exist close to the threshold. - When time can serve as a cut-off, provided that the policy change occurs at a clearly defined date and interest is limited to the immediate period around implementation. - When individuals close to the cut-off can be considered comparable at either side.	- When eligibility around the cut-off cannot be reliably determined (e.g. due to manipulation or imprecise measurement), particularly when crossing the threshold barely affects treatment. - In time-based applications, when analysing gradual, delayed, or long-term effects following a policy change, rather than the immediate discontinuity at implementation. - The cut-off is also used by other programmes than the one of interest.

3.2.3 Strengthening internal validity

3.2.3.1 Regression adjustment

In quasi-experimental studies, **regression-based covariate adjustment** can be used to control for confounding when key sources of bias, resulting from either selection bias or confounding, are captured by observed covariates and residual confounding from unmeasured factors is expected to be limited.¹¹⁸ Other fundamental design flaws (e.g. spillover) cannot be corrected with regression adjustment.

Regression adjustment is particularly appealing when the causal estimand is defined within a regression framework and when effect estimation already relies on modelling the outcome as a function of the intervention and covariates. It is important to note that correct model specification is essential for the validity of regression-based adjustment.¹⁴³

In practice, model misspecification can arise from common analytic choices, such as assuming linear effects when relationships are nonlinear, omitting important interaction effects, or adjusting for variables influenced by the intervention itself. Misspecification can lead to biased effect estimates even when all relevant confounders are measured.¹⁴³ To reduce sensitivity to model misspecification, doubly robust estimators can be considered (which will be covered in section 3.5.3).

3.2.3.2 Matching

Matching methods aim to **create an artificial comparison group that resembles the intervention group** as closely as possible based on observed baseline characteristics.¹¹⁷ Matching can be applied at different stages of a study. During the recruitment stage, researchers may actively select control participants who closely resemble individuals in the intervention group. During the analysis stage, matching can be used to select appropriate controls from a larger pool of observations. In either case, it is important to determine in advance which variables will be used for matching, as they must be measured consistently across all participants.

Matching is typically performed using baseline data collected prior to the intervention, ensuring that **matched characteristics are not affected by the intervention itself**. In some situations, matching can also be carried out using post-intervention data, provided that the variables used for matching are not plausibly influenced by the intervention and can reasonably be considered time-invariant.¹¹⁵

Matching approaches can be implemented at the group level or the individual level. Frequency or group matching aims to ensure that the overall distribution of key characteristics, such as age or sex, is similar between the intervention and comparison groups.¹⁴⁴ Individual matching constructs a comparison group by pairing each participant in the intervention group with one or more non-participants who have similar observed characteristics. Individual matching involves some key design choices, which are listed in Table 15.¹⁴⁵



Table 15 – Overview of key design choices for individual matching

Design aspect	Explanation
Matching ratio	Individual matching may be conducted one-to-one or one-to-many. The latter, where each individual from one group is matched to multiple individuals of the other group, is more common because the number of cases is usually smaller than the number of controls and because it increases statistical efficiency. ¹⁴⁵
With or without replacement^d	Matching with replacement is often used to reduce bias through closer matches, while matching without replacement offers greater precision by using more unique subjects. ^{146, 147}
Matching algorithm	Without replacement, matching can be implemented either via an optimal algorithm, often computationally demanding for large datasets, or via a sequential (greedy) approach ^{e, 147, 148} . With replacement, matching is typically implemented using a simple nearest neighbour procedure, as each treated unit can be matched independently.
Distance metric	Matching is based on a measure of similarity between individuals, with matches formed between individuals whose scores lie within a predefined distance (caliper) to ensure comparability. ^{115, 149} More complex alternatives include Mahalanobis Distance Matching and Coarsened Exact Matching, but these are not used as often.

Matching can substantially strengthen internal validity by reducing bias due to observed baseline differences between intervention and comparison groups. By constructing more comparable groups prior to outcome analysis, matching supports more credible comparisons by making it more plausible that observed outcome differences are attributable to the intervention rather than to pre-existing differences.¹⁴⁷

In addition, improved comparability may lead to **more precise effect estimates**, thereby increasing statistical power.^{147, 148} Matching is also transparent and intuitive, allowing researchers to directly assess whether comparable individuals have been identified and to evaluate the balance achieved in observed characteristics.

A practical **limitation** of individual matching is that participants for whom **no suitable match** can be found must be **excluded** from the matched sample. This loss of observations can reduce statistical power and external validity, as the resulting estimates apply only to the subset of participants for whom adequate matches are available. The problem becomes particularly pronounced when matching is based on many characteristics, as each additional variable makes the criteria for a suitable match more restrictive, quickly making it difficult or impossible to find comparable counterparts.¹⁴⁷ This issue is further exacerbated when intervention and comparison groups differ substantially at baseline, a phenomenon commonly referred to as the curse of dimensionality.^{115, 117} In practice, exact matching is therefore typically feasible only for a **small number of characteristics**, often no more than two or three, as suitable matches quickly become scarce. In section 3.5.1 of this chapter, propensity score matching is discussed, which proposes an alternative approach when matching on a larger number of variables is desired.

^d With replacement allows an individual to be matched multiple times, while without replacement restricts each individual to be used only once.

^e Optimal matching selects matches jointly to minimise overall distance across all pairs, whereas sequential (greedy) matching assigns matches one at a time based on the closest available match.

3.3 Designs for before-and-after effect identification

3.3.1 Common types of bias

Before-and-after designs rely on comparing the same individuals or population before and after an intervention. While straightforward and often feasible, these designs are vulnerable to several well-known threats to validity. Because there is no control group, any change over time may reflect influences *other* than the intervention itself. Five common sources of bias are described below. Throughout section 3.3, a distinction will be made between simple before-and-after designs and longitudinal time-structure-based designs. The latter build on simple before-after contrasts but strengthen internal validity by exploiting the alignment between outcome change and the timing of intervention introduction, making some of these biases less applicable.

- **Regression to the mean**^{129, 150}

When outcomes are observed at an unusually extreme point and subsequent change is interpreted as an intervention effect, later measurements tend to be closer to the average due to random variability rather than true change, a phenomenon often labelled a bias despite not arising from measurement error. In longitudinal time-structure-based designs, multiple pre-intervention observations reduce the risk of this type of bias.¹⁵¹

Example: An intervention introduced after an unusually high monthly hospital mortality rate may appear effective because mortality returns to typical levels in subsequent months, even in the absence of any real effect of the intervention.

- **Maturation bias**¹²⁹

Individuals naturally change over time, independent of any intervention. These changes can lead to improvements or deteriorations that may be mistaken for intervention effects. Multiple pre-intervention observations reduce the risk of this type of bias in longitudinal time-structure-based designs.¹⁵¹

Example: Children develop cognitively with age, patients recovering from an illness improve gradually, and behaviours fluctuate with seasonal cycles. In a pre-post design, such natural changes can easily be misattributed to the intervention.

- **History bias**¹²⁹

Events external to the intervention may occur between the pre-test and the post-test and influence the outcome. These events are unrelated to the programme being evaluated but can produce changes that appear, incorrectly, as if they were caused by the intervention. History bias may persist in longitudinal time-structure-based designs as external events coinciding with the intervention can still produce changes that are indistinguishable from the intervention effects. The extent to which this threat is mitigated depends on the specific timing-based design and how intervention timing is structured.¹⁵²

Example: A new policy, media campaign, or public health initiative is introduced during the study period, making it impossible to disentangle the intervention's effect from the broader context.

- **Testing effects bias**¹²⁹

The process of taking a pre-test can alter participants' behaviour, awareness, or responses at the post-test. This means the act of measurement itself influences the outcome. Testing effects are attenuated in longitudinal time-structure-based designs, because repeated baseline measurements make it less likely that one-off measurement reactivity is mistaken for an intervention effect.¹⁵³

Example: Completing a baseline questionnaire may draw attention to certain behaviours or symptoms, making participants more mindful or more practiced when completing the post-test.



- **Instrumentation bias**¹²⁹

Changes in the way measurements are taken between pre-test and post-test can introduce bias. Differences may arise from altered measurement tools, updated procedures, or variability among observers. Because of the longitudinal nature of time-structure-based design, instrumentation bias can threaten internal validity in such designs.

Example: Switching from paper surveys at baseline to digital surveys at follow-up or having different staff conducting observations with (slightly) different interpretations of the measurement protocol. Another example, specifically relevant for longitudinal time-structure-based designs, is changes to administrative procedures altering record-keeping practices.¹⁵⁴

- **Differential loss to follow-up bias**¹⁵⁵

Results can become biased when some participants drop out between the pre-test and post-test, especially when dropout is related to their characteristics or, more importantly, their outcomes (experimental mortality bias). If the people who remain differ systematically from those who leave, the post-test data no longer represent the original group. Longitudinal time-structure-based designs do not inherently protect against this threat. In designs that repeatedly follow the same individuals, longer follow-up periods may increase the opportunity for attrition and thereby increase the risk of bias.

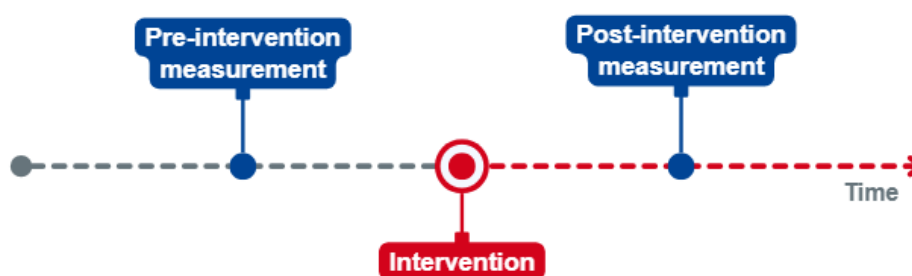
Example: Participants experiencing poorer outcomes or less favourable views of the intervention may be more likely to discontinue participation, inflating the apparent effectiveness of the programme.

3.3.2 General designs for simple before-and-after effect identification

3.3.2.1 Single-group pre-post design

In single-group pre-post designs, illustrated in Figure 16, all participants receive the intervention. Because there is no separate control group, the intervention's effect is inferred solely from changes within the same group by comparing pre-test measurements taken before the intervention with post-test measurements taken afterwards. Any improvement or deterioration in the outcome is attributed to the intervention.

Figure 16 – Illustration of the single group pre-post design



Single-group pre-post designs offer a practical alternative when the use of a control group is impractical or in case it raises ethical concerns, as all participants receive the intervention and no one is withheld from potentially beneficial treatment. They are also less costly, require fewer people and are easier to implement than designs requiring multiple groups.¹¹⁸ Because the same individuals are measured over time, these designs capture within-group change and avoid the selection bias that arises when comparing different groups. On the other hand, they are **highly vulnerable to other sources of bias**, since any observed change may stem from factors other than the intervention operating simultaneously.¹¹⁸ Without a control group, there is a risk of the biases outlined in the previous section, all of which can distort the apparent effect. As a result, **conclusions about the**

effects of the intervention drawn from single-group pre-post studies are inherently limited and must be interpreted with caution.⁶³ Nevertheless, a single-group pre-post design can still yield preliminary estimates of change and is often applied in pilot work to assess the likely effects of an innovation.¹⁵⁶

Table 16 – Indications and contra-indications to the use of the single group pre-post design

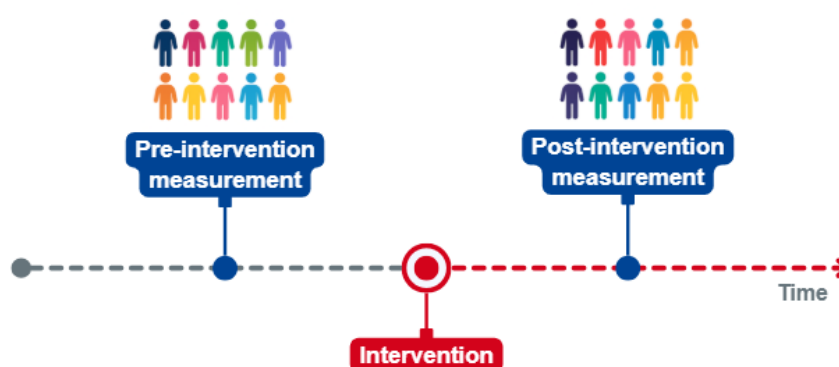
Indications	Contra-indications
- When the use of a control group is not feasible or ethical. - When time series data are not available. - When the study is an exploratory or pilot study.	- When more data are available (either with controls or time series) allowing more adequate design choices. - When influence of time, external events or other trends on the outcome variable is likely. - When the outcome measure is volatile over time.

3.3.2.2 Simple before-and-after comparison using repeated cross-sectional data

With repeated cross-sectional data, outcome data are collected at multiple time points from representative samples drawn from the same target population.⁶³ Each measurement relies on a newly selected cross-section, meaning the samples do not comprise the same individuals over time (e.g. a survey of nurses in hospitals at different time points without linking individual nurse responses across measurement periods). Consequently, **changes are estimated at the aggregate or population level** rather than at the individual level, enabling researchers to examine trends over time despite variations in sample composition.⁶³

Repeated cross-sectional data can be used to describe changes in population outcomes over time by comparing aggregate measures across pre- and post-intervention periods. These before-and-after comparisons operate at the population level and do not track individual change (see Figure 17). As a result, observed differences over time may reflect not only intervention effects but also broader temporal changes, including shifts in population composition, contextual factors, or other concurrent developments.

Figure 17 – Illustration of before-after comparisons using repeated cross-sectional data



Repeated cross-sectional studies are relatively simple to implement because each data collection wave samples new individuals rather than following the same participants over time. This avoids some logistical challenges of longitudinal studies, such as attrition, respondent fatigue, and the need for



longitudinal linkage. It also allows repeated assessment of population-level outcomes in settings where the underlying population may change over time.

Due to several important challenges, repeated cross-sectional data offer **limited ability to attribute observed changes to the intervention**, and trend interpretations must be made cautiously.¹⁵⁶ This design is susceptible to bias and confounding, as observed changes over time may reflect factors other than the intervention itself, given the potential biases outlined earlier. In addition, repeated cross-sectional designs carry a risk of selection bias if samples are not consistently representative of the population across waves,⁶³ as well as if the underlying population changes over time with regard to potential confounders of the outcome.¹³³ Ensuring representative and comparable samples across waves can help mitigate these risks.⁶³

Table 17 – Indications and contra-indications to the use of repeated cross-sectional data

Indications	Contra-indications
- When interventions are implemented at the population or system level. - When population-level trends before and after an intervention are needed. - When time series data (see below) are not available.	- When individual-level effects are of primary interest. - When the composition of samples or the underlying population is likely to change over time for potential confounders. - When strong causal conclusions are desired.

3.3.3 Strengthening internal validity for simple before-and-after effect identification

The methodological rigour of these designs can be strengthened by **increasing the number of data collection points**.⁶³ Adding a second pre-test before the intervention (AAB-design^f) strengthens evidence that helps rule out regression to the mean (in the single-group pre-post design) and other confounding factors as alternative explanations for any association observed between the intervention and the post-test results.^{114, 157}

The AAB-design can be extended by removing the intervention after the third measurement (AABA-design) and measuring a fourth time to assess whether this removed the outcome as well.¹⁵⁷ This withdrawal answers additional questions about what happens in the absence of the intervention.¹¹⁴ For the AABA design, it is important to determine in advance whether an effect is expected to persist and for how long, so that the final measurement is scheduled after any potential fade-out period.¹¹⁴ The final design adaptation (i.e. the AABAB-design) builds on the previous one by reintroducing the intervention again after the fourth measurement and conducting a fifth measurement where the effect is expected to reappear.¹¹⁴

Furthermore, a potential modification of the base design involves measuring a second dependent variable. Apart from the intervention, the second dependent variable should be closely related to the main outcome, influenced by similar factors and possible confounders, but not by the intervention.¹¹⁴ If the main outcome changes after the intervention, while this similar measure stays the same, it provides a strong indication that the observed effect is not simply the result of confounding. For example, a hospital introduces a hand hygiene programme to reduce healthcare-associated infections. Patient fall rates are measured as a second dependent variable because they are influenced by similar contextual factors (e.g. staffing levels, bed occupancy rate, patient case mix,...)

^f In the notation used here for educational purposes, each letter represents a measurement occasion, with A denoting measurement in the absence of the intervention and B denoting measurement in the presence of the intervention. This differs from the more common phase-based notation, where each letter represents a study phase rather than a single observation (e.g. AABA here corresponds broadly to an ABA design in phase-based notation).

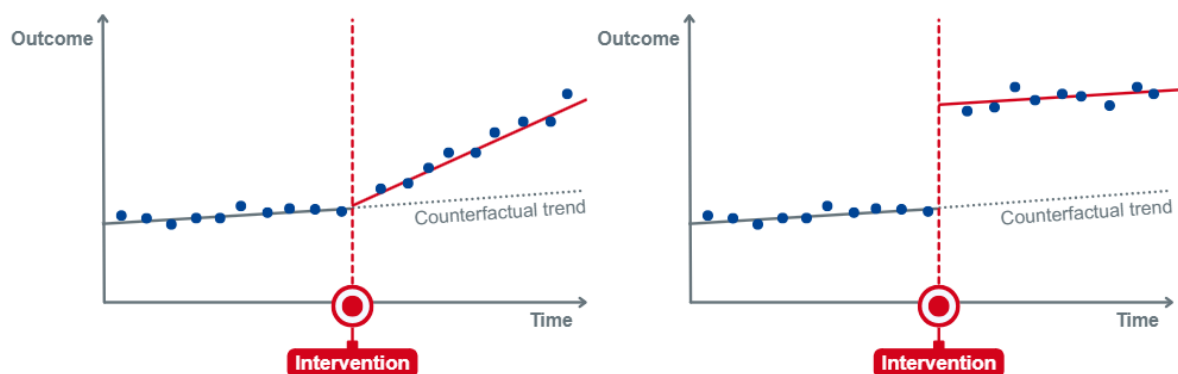
but are not expected to be affected by the intervention. If infection rates decrease while fall rates remain stable, this supports the effectiveness of the intervention.

3.3.4 General designs for longitudinal time-structure-based effect identification

3.3.4.1 Interrupted time series

Interrupted time series (ITS) designs **compare a group or population to itself over time using a large number of repeated measurements, with the pre-intervention period serving as the control comparison.**⁷ They are most commonly used in natural experiment settings,⁷ particularly in retrospective cohort studies based on administrative or routinely collected data, where long and frequent time series are readily available. The effect of the intervention is assessed by examining whether there is a change in the trend/slope (see left panel of Figure 18) or level/intercept (see right panel of Figure 18) of the outcome following the intervention compared with the pattern observed beforehand.^{63, 114, 157} ITS is best suited to **short-term outcomes** that are expected to respond shortly after implementation or following a well-defined lag.¹⁵¹

Figure 18 – Illustration of slope changes (left panel) and intercept changes (right panel)



The use of ITS designs has several important strengths and some limitations. ITS designs are relatively straightforward and cost-effective to implement, as they require data from only one group and do not involve treatment assignment.^{63, 158} ITS allows researchers to assess not only whether an outcome shifted following an intervention but also how the pattern of change evolves over time (level changes vs. slope changes). Nonetheless, it is important to consider in advance whether the expected effects are likely to be detectable within an ITS framework, as **effects often do not follow a fixed or immediate timeline in real-world settings** and may emerge at different times across populations and periods.¹⁵⁴

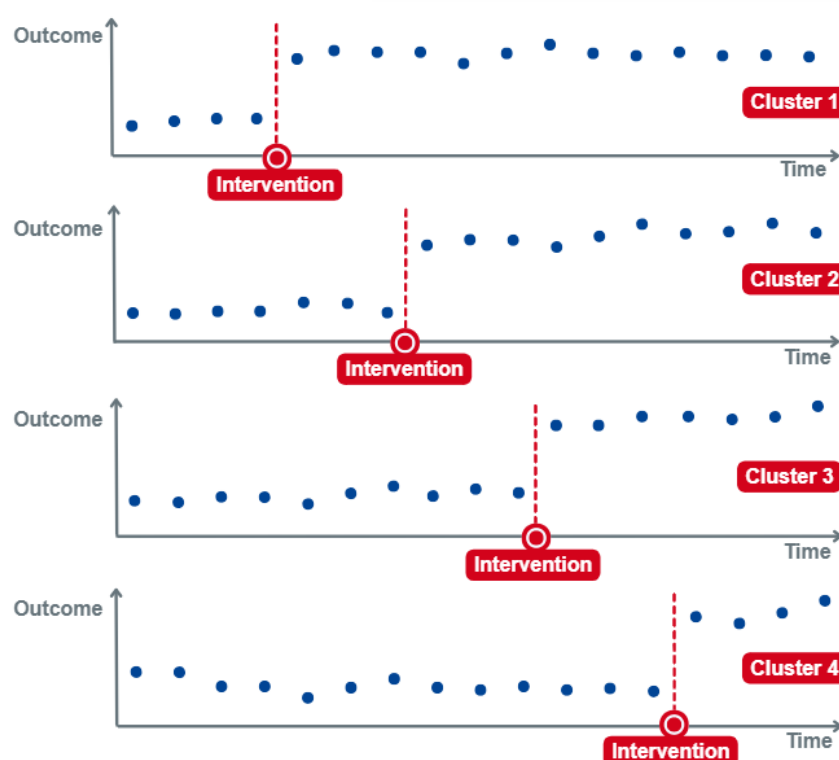
Because ITS explicitly models time trends, it provides stronger internal validity than simpler pre-post comparisons and can effectively control for time-based confounders, such as seasonal patterns or gradual background trends.¹⁵⁷ Nevertheless, this design remains vulnerable to the bias introduced by such trends.⁶³ Internal validity strengthens when many pre-intervention data points are available, enabling a stable, credible baseline before the intervention. At the same time, the need for a **substantial number of repeated measurements over a longer period** can also be a drawback because it limits the practical applicability of this design.⁷ Recommendations suggest collecting at least eight observations before the intervention and at least eight after it to support reliable estimation of trends and changes.^{154, 159} This minimum supports basic ITS modelling and interpretation. However, a much larger number of observations may be required for more advanced statistical analyses, especially when modelling complex autocorrelation structures, with some authors recommending substantially more data points (about 100).¹⁵⁴ In general, data analysis in the context of ITS is complex due to autocorrelation and requires advanced statistics.¹²⁹

**Table 18 – Indications and contra-indications to the use of the interrupted time series design**

Indications	Contra-indications
- When long time series exist, including sufficient baseline data. - When the intervention timing is sharp and well defined.	- When other concurrent events are expected to be implemented around the time of the intervention of interest. - When pre-intervention data are highly volatile, limiting ability to detect changes in level or trend.

3.3.4.2 Multiple baseline designs

Multiple baseline designs (MBDs) are **longitudinal designs in which an intervention is introduced at different time points (calendar time) across several individuals or clusters of individuals**, with repeated outcome measurements throughout, as shown in Figure 19.¹⁵² Evidence for an intervention effect is provided when outcome changes occur only after the intervention is introduced and at different times across the various baselines, thereby reducing the plausibility that observed changes are attributable to other events occurring at a single point in time.^{160, 161} By staggering intervention onsets, MBDs protect internal validity by guarding against history bias from a single external event, such as a concurrent policy change or media campaign, that might otherwise explain observed changes in outcomes.¹⁶¹

Figure 19 – Illustration of the multiple baselines design

MBDs are **closely related to interrupted time series designs**, as both rely on within-series comparisons over time, and to stepped-wedge designs, as both involve staggered implementation of an intervention across units.⁶³ Unlike stepped-wedge designs, however, MBDs do not require that all individuals or clusters ultimately receive the intervention, nor do they require randomisation of intervention timing, although randomisation of intervention order can be incorporated to strengthen the findings.^{152, 161} In applied and public health research, MBDs are often analysed as self-controlled time-series studies rather than as cluster randomised trials.¹⁵²

Two main subtypes of MBD can be distinguished. In **concurrent MBDs**, data collection across individuals or clusters occurs simultaneously, whereas in **non-concurrent MBDs**, the timing of measurements and the introduction of interventions differ across series.¹⁶⁰ Concurrent designs are generally considered stronger, as they more effectively control for shared historical influences. While non-concurrent designs may be more feasible when logistical constraints prevent simultaneous recruitment or data collection, they remain vulnerable to confounding by time-varying external factors.¹⁶¹

MBDs offer a flexible and ethically attractive alternative when withholding the intervention is unethical. In addition, this approach is especially useful when simultaneous rollout of interventions is impractical. At the same time, MBDs typically require a relatively large number of repeated measurements to establish stable baselines and to detect intervention effects, which may increase data collection burden and cost when routine data systems are not available.¹⁶¹ MBDs are best suited to interventions expected to produce relatively rapid effects, as interventions with long latency periods (e.g. in case behaviour change of healthcare professionals is part of the intervention) make it difficult to separate baseline and intervention phases clearly. Finally, the timing of intervention starts affects internal validity. If intervention onsets are too closely spaced, it becomes difficult to distinguish intervention effects from contemporaneous external influences.¹⁶¹

Table 19 – Indications and contra-indications to the use of the multiple baselines design

Indications	Contra-indications
- When intervention implementation can be staggered across individuals, groups, or settings. - When sufficient baseline data can be collected to establish stable or interpretable pre-intervention patterns. - When history bias from external events happening at the same time as the intervention is a real concern.	- When intervention start times cannot be meaningfully staggered or must be implemented simultaneously across all individuals or clusters. - When intervention effects are expected to emerge only after a long delay or with substantial variability in timing.

3.3.5 *Strengthening internal validity for longitudinal time-structure-based effect identification*

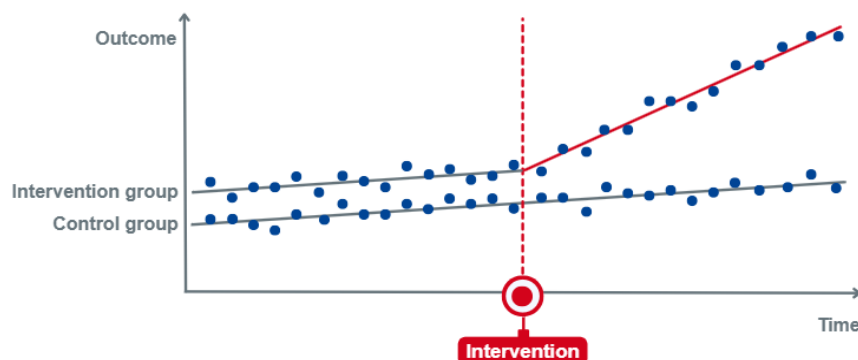
In general, the withdrawal and re-introduction of the intervention and the use of a non-equivalent second dependent variable, both discussed in section 3.3.3 are valid in this context as well.¹⁵⁴ In addition, the upcoming sections elaborate on two methods related to comparison with control trends.

3.3.5.1 *Controlled Interrupted Time Series*

Although ITS already offers stronger internal validity than simple pre-post comparisons, its robustness can be further enhanced by adding an unexposed comparison group as a reference for what would have occurred in the absence of the intervention.^{63, 132, 151} In a controlled ITS (shown in Figure 20), outcomes are measured repeatedly in both the intervention series and a comparison series that is assumed to have followed a similar trajectory in the absence of the intervention. By modelling differences in level and slope between these series before and after the intervention, researchers can better account for concurrent events and gradual changes unrelated to the intervention itself. **This approach helps reduce several common internal validity issues**, including history bias, maturation bias, and testing effects.¹²⁹ Its validity nevertheless depends on careful selection of the comparison group. When these conditions are reasonably met, ITS with a comparison series is often considered to be among the strongest quasi-experimental design short of randomisation.^{129, 154}



Figure 20 – Illustration of the Controlled Interrupted Time Series design

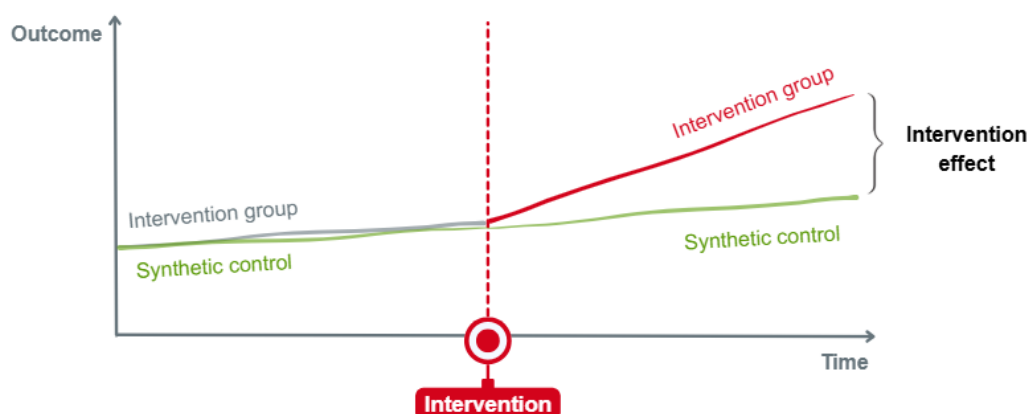


3.3.5.2 Synthetic controls

While controlled ITS designs improve internal validity by introducing a comparison group, further refinement is possible **when no single comparator adequately reflects the pre-intervention trajectory**. Synthetic control methods extend this idea by using a weighted combination of comparison entities to construct a more credible counterfactual,¹¹⁸ as illustrated in Figure 21.

Synthetic control methods are designed for situations where an intervention is received by **only one or a small number of cases** (such as individuals, hospitals or countries), with many comparable entities remaining unexposed to the intervention.¹⁶² This method is often used for situations with a single treated unit, but it can also be applied to multiple treated units.^{162, 163} The approach constructs an explicit counterfactual by combining information from these comparison entities into a single synthetic comparison. This synthetic control is **a weighted average chosen by a mathematical optimisation procedure to closely mirror the case of interest in observed characteristics and outcome trends before the intervention**.¹⁶²⁻¹⁶⁴ The process of assigning these weights is completely data-driven, with comparison cases that are more similar to the case of interest receiving higher weights.¹⁶³ Differences that emerge afterwards are interpreted as the effect of the policy relative to this counterfactual scenario.¹⁶² As a result, the approach is well-suited to macro- or meso-level analyses that do not rely on individual-level data or one-to-one matching between individuals.¹⁶³

Figure 21 – Illustration of synthetic controls



The synthetic control method may be viewed as an extension of DiD analysis, while being intuitively close to matching approaches aiming to similar construct a control group by balancing relevant characteristics.¹¹⁸ Synthetic control methods are particularly useful **when the parallel trends assumption** required for DiD analyses is **likely to be violated or difficult to verify**.¹⁶³ Rather than relying on this assumption or on explicit matching of observed covariates, the method reconstructs the

full pre-intervention outcome trajectory of the case of interest using a weighted combination of comparison cases. As a result, the approach can account for both observed and unobserved time-varying confounding, provided these dynamics are reflected before the intervention.¹⁶² This is illustrated in studies comparing synthetic control approaches to DiD analysis, which found different conclusions about programme effectiveness in contexts where the parallel trends assumption may not hold.^{162, 165}

Some challenges to the use of synthetic controls must be considered. Synthetic control methods are **data-intensive**, requiring a reasonably long pre-intervention time series to construct a credible counterfactual and a large, carefully selected donor pool. The validity of the method depends critically on the quality of the donor pool. Comparison entities must be similar to the case of interest, must not be affected by the intervention, and must not be subject to other shocks during the study period.¹⁶³ However, there is substantial subjectivity in establishing this similarity, with regards to the choice of the donor pool (e.g. all countries vs. high-income countries vs. EU countries), predictor variables considered and related inclusion/exclusion criteria.¹⁶³ Furthermore, there is no clear consensus on how to define a 'good' pre-intervention fit or how to formally verify the absence of shocks in comparison entities.¹⁶³ Despite this, the credibility of the results relies heavily on these concepts. Finally, applying synthetic control methods to settings with many intervention cases can be computationally demanding as well as being conceptually awkward, as the interpretation is most straightforward when focusing on a small number of cases.¹⁶² A potential approach to handle this can be to aggregate multiple units who received the intervention.¹⁶⁶

Textbox 4 – An illustrative example on synthetic controls

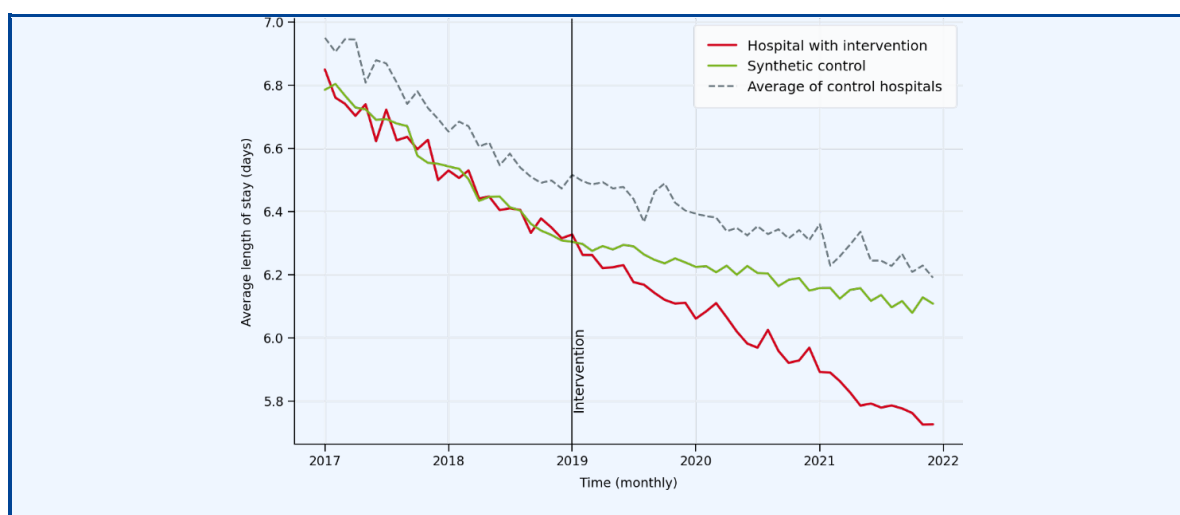
Imagine an evaluation of a quality improvement programme introduced in Hospital A in 2019 to reduce average length of stay. While this programme is unique to hospital A, several similar hospitals are available to serve as potential control groups.

A simple comparison with one hospital (e.g. Hospital B) may be misleading if no single hospital closely matches Hospital A's pre-intervention trend. Instead, a synthetic control is constructed by combining several hospitals into a weighted average that best reproduces Hospital A before 2019.

In this example, the donor pool consists of four hospitals:

<i>Hospital</i>	Similarity in pre-2019 trends	Weight in synthetic control
<i>Hospital B</i>	Moderate	0.25
<i>Hospital C</i>	High	0.40
<i>Hospital D</i>	Low	0.10
<i>Hospital E</i>	High	0.25
<i>Total</i>		1.00

The weights are chosen so that the weighted combination of Hospitals B–E closely matches Hospital A's length of stay before the intervention:



3.4 Causality in quasi-experiments

3.4.1 From association to causation

Health services and policy researchers are often interested in questions that are inherently causal, seeking to determine whether an intervention produced a particular effect. Although the concept of causality is intuitive, establishing a causal relationship is often challenging because causal effects cannot be observed directly. Importantly, observing an association between an intervention and an outcome does not necessarily imply that the intervention caused the outcome.¹¹⁷

Over the years, several frameworks have been developed to guide causal reasoning. One of the most influential is Bradford Hill's set of considerations for causal inference¹⁶⁷ (Textbox 5), which emphasises that causal claims should not be based solely on statistical associations but should also be evaluated in light of factors such as temporality, consistency, plausibility and experimental evidence. While these considerations do not provide definitive proof of causality, they offer a useful framework for assessing whether a causal interpretation of an observed relationship is credible.

Textbox 5 – Bradford Hill's considerations for causal inference

- **Strength:** Strong associations are generally more suggestive of causality than weak associations.
- **Consistency:** Similar findings are observed across different studies, populations and settings.
- **Specificity:** The exposure is linked to a particular outcome (i.e. one cause has one effect).
- **Temporality:** The cause must precede the effect. This is the only criterion that is universally necessary for causal inference.
- **Biological gradient:** Greater exposure is associated with greater effect (a dose-response relationship).
- **Plausibility:** A credible mechanism exists through which the exposure could affect the outcome.
- **Coherence:** The proposed causal relationship is compatible with existing knowledge and evidence.
- **Experiment:** Experimental or quasi-experimental evidence supports the relationship.
- **Analogy:** Similar causal mechanisms have been observed in related contexts.

Source: Bradford Hill (1965)¹⁶⁷

A useful framework for conceptualising the progression from describing relationships to answering causal questions is provided by Pearl's ladder of causation,¹⁶⁸ which describes three levels of reasoning:

- **Association ('seeing'):** The lowest level of the ladder concerns what can be learned from observed data and addresses questions such as whether two variables are related. Research at this level focuses on identifying and quantifying associations observationally, without necessarily making causal claims.
- **Intervention ('doing'):** The middle level considers the consequences of actions and asks what would happen if a particular intervention were implemented. Questions at this level are inherently causal because they concern the effects of policies, programmes or treatments. Experimental studies and most quasi-experimental designs are intended to answer questions at this rung.
- **Counterfactuals ('imagining'):** The highest level of the ladder of causation considers what would have happened under alternative circumstances. For example, what would the outcome have been if an intervention had not been implemented? This level underpins modern causal inference and requires reasoning about outcomes that cannot be directly observed.

Quasi-experimental designs can primarily be situated at the intervention level. Yet, an intervention can only be considered causal if any alternative explanations for the observed effect are not plausible. This requires the approximation of the relevant counterfactual outcomes, which cannot be directly observed. Consequently, modern approaches to causal inference increasingly view intervention and counterfactual reasoning as closely intertwined, with the latter providing the theoretical foundation for estimating causal effects.

3.4.2 The causal inference framework

3.4.2.1 Counterfactuals

A central idea in establishing causality is the counterfactual, as introduced in Pearl's highest level of the ladder, which captures what the outcome variable would have been in the absence of the intervention and is essential for determining whether the intervention changed an outcome.^{117, 168} Because a counterfactual cannot be directly observed for the same individuals, the counterfactual must be approximated.^{117, 169} Thus, constructing a credible counterfactual is a core task in any impact evaluation. In classic experimental designs, random assignment creates a control group that credibly estimates the counterfactual by balancing both observed and unobserved characteristics across groups. This makes alternative explanations for observed differences less plausible and strengthens causal inference.^{169, 170}

Quasi-experimental designs, on the other hand, construct an approximation of the counterfactual by either **comparing an intervention group to another group** or by **using repeated observations of the same population over time**.¹¹⁴ However, without random assignment, greater attention must be paid to whether the chosen comparison adequately reflects what would have happened in the absence of the intervention. Constructing a credible counterfactual using both design and analytical techniques is therefore particularly important in quasi-experimental evaluation.¹²⁰

3.4.2.2 Potential outcomes

The **causal inference framework** formalises counterfactual reasoning through the concept of potential outcomes, which represent the outcomes that would be observed under different intervention conditions.¹⁷¹ For every unit under study, whether an individual, organisation or population, one can conceptualise multiple potential outcomes corresponding to different intervention scenarios. However, only one of these potential outcomes can be observed in practice. The remaining potential outcomes are counterfactual.



As an example (adapted from Cunningham, 2021¹⁷²), suppose we have three patients with cancer and two treatments that we would like to compare in terms of post-treatment lifespan. The potential outcomes are shown in Table 20. In this example, each patient has two potential outcomes: one under treatment 1 and one under treatment 2. However, only the outcome corresponding to the treatment actually received can be observed. *The causal effect is defined as the contrast between these potential outcomes.* Because one potential outcome is always unobserved, individual causal effects cannot be measured directly and must instead be estimated.

Table 20 – An example of potential outcomes

Patient	O ₁ : life span after treatment 1	O ₂ : life span after treatment 2	Difference
A	7	(1)	6
B	(5)	6	-1
C	5	(1)	4

This table displays potential outcomes after two treatments and the difference between these potential outcomes. numbers in bold = actual observed outcomes;

numbers between brackets = unobserved counterfactual outcome if patients would have had the other treatment

3.4.2.3 Causal estimands

The potential outcomes framework also clarifies how causal effects are defined. In practice, researchers are rarely interested in the causal effect for a single individual. Instead, causal effects are typically averaged over a specified population, yielding explicit causal estimands.

A causal estimand specifies the effect of interest, including the population to which it applies, the intervention being compared, and the outcome on which the effect is measured.¹⁷³ In quasi-experimental studies, group assignment is not random and intervention uptake is typically selective, meaning that the ATE and ATT may differ substantially. Researchers should therefore be explicit about which causal effect is most relevant to their evaluation question.¹³⁷

These concepts can be illustrated by revisiting Table 20. The ATE for treatment 1, defined as the average effect of an intervention if it were applied to the entire target population regardless of receiving the treatment, is the average of the individual causal effects: $(6 - 1 + 4) / 3 = 3$ years. In contrast, the ATT for treatment 1, reflecting the average effect among the subset of individuals that actually received this treatment, is the average of the individual causal effects among patients A and C, yielding an effect of: $(6 + 4) / 2 = 5$ years.

It is up to the researchers to decide which of the two is most valuable prior to data analysis. For instance, evaluations of resource intensive programmes targeted at high need patients often focus on the ATT, reflecting interest in effects among those who receive the intervention. In contrast, evaluations of low cost system wide changes, such as standardized care pathways or information tools, may prioritize the ATE to inform population level implementation decisions.¹⁴⁹ Importantly, different quasi experimental designs implicitly target different causal estimands, as methods that construct comparators to resemble the intervention group typically estimate effects among the treated (ATT), whereas population level effects (ATE) require approaches that reweight or model the full eligible population.¹⁴⁸

The concepts of counterfactuals, potential outcomes and causal estimands form the foundation of modern causal inference. However, causal effects can only be identified under a number of assumptions, which are discussed in the next section.

3.4.2.4 Assumptions of causal inference

Because counterfactual outcomes are not observed, causal effects cannot be identified from observational data without additional assumptions about the underlying data-generating process. The causal inference framework makes these assumptions explicit. The **most important assumptions** are:¹⁷⁴

- **Consistency:** The outcome observed for an individual is the result of the specific intervention they actually received. This assumption requires that the intervention is sufficiently well-defined for all individuals classified as exposed to be considered as receiving the same intervention.
- **Exchangeability:** Also known as ‘no unobserved confounding’ (more information on confounding can be found in the next section). After accounting for an appropriate set of variables, treated and untreated individuals are comparable in terms of their potential outcomes.
- **Positivity:** Within any specific subgroup of the population (e.g. defined by characteristics such as age, sex, socioeconomic background), there is a non-zero probability that an individual could receive any of the interventions under study.

Importantly, the assumptions required for causal inference cannot be tested directly and are often difficult to justify in applied research. The exchangeability assumption is particularly demanding, requiring that those who did and those who did not receive the intervention are comparable after adjustment for observed variables. In observational settings, this condition is fragile, as unmeasured confounding can seldom be ruled out with confidence. Causal inference based on observational data, therefore, remains contested. Rather than resolving this uncertainty, the causal inference framework makes explicit which biases may persist and which assumptions, especially exchangeability, must be judged on substantive grounds.

3.4.3 Directed acyclic graphs

Directed acyclic graphs (DAGs) are graphical representations of assumed causal relationships between variables. In a DAG, variables are represented by nodes and causal relationships by arrows. The term *directed* refers to the fact that arrows have a direction, indicating the presumed direction of causality, while *acyclic* means that causal relationships cannot loop back on themselves. Importantly, a DAG does not represent statistical associations but rather the researcher's assumptions about the causal processes underlying the data. By making these assumptions explicit, DAGs help identify potential sources of bias and guide decisions about which variables should and should not be adjusted for during analysis.

The primary purpose of a DAG is to help researchers identify which variables should be adjusted for when estimating a causal effect. This decision is guided by the **backdoor criterion**, a principle stating that alternative pathways that could create a misleading association between the exposure and outcome should be eliminated, while the causal pathway of interest should be preserved. In practice, this typically means adjusting for confounders while avoiding adjustment for mediators and colliders.¹⁴³

A key feature of DAGs is that the role of a variable is determined by its position within the causal structure. Variables that are associated with both the exposure and outcome do not necessarily require the same analytical treatment. Table 21 summarises the most common relation types encountered in causal inference, while Textbox 6 illustrates them with an example.

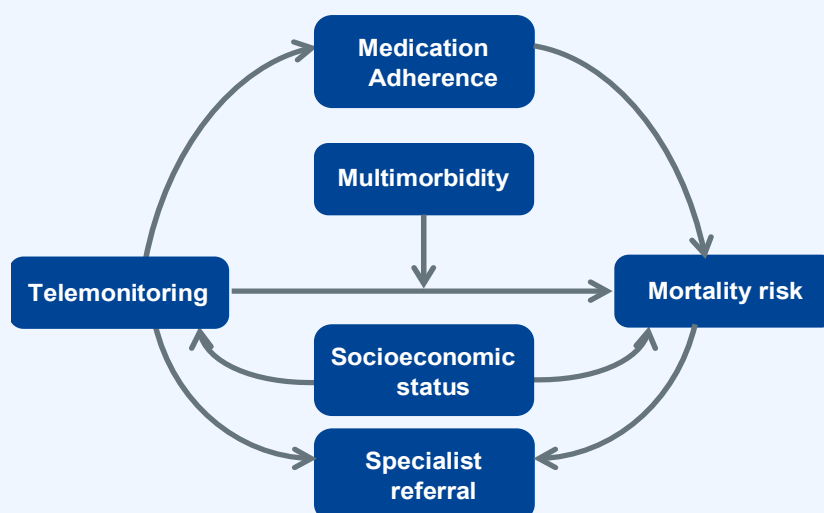


Table 21 – Most common relations and their adjustments

Relation	Graphical structure	Definition	Adjustment required?
Confounder		A variable that influences both the exposure and outcome, creating a non-causal pathway between them.	Yes. Adjustment helps eliminate alternative explanations for the observed association between the exposure and outcome.
Mediator		A variable that lies on the causal pathway between exposure and outcome and transmits part of the intervention effect.	No. Adjustment removes part of the causal effect being estimated.
Collider		A variable influenced by both the exposure and outcome, or by their causes.	No. Adjustment may introduce bias by opening a previously blocked non-causal pathway.
Moderator		A variable for which the magnitude or direction of the causal effect differs across subgroups.	Address through stratified analyses or interaction terms for estimating and reporting causal effects.

Textbox 6 – Illustrative example of a DAG and its components

Imagine a study assessing the effect of telemonitoring on mortality risk among patients with heart failure. The DAG below represents the assumed causal structure underlying this relationship.



This DAG contains the following components:

- Socioeconomic status (confounder): Socioeconomic status may influence both participation in telemonitoring programmes and mortality risk. For example, patients with higher socioeconomic status may have better access to technology and healthcare resources while also experiencing better health outcomes independent of the intervention.

- Medication adherence (mediator): Telemonitoring may improve patients' adherence to prescribed treatment through reminders, monitoring, and increased contact with healthcare professionals. Improved adherence can subsequently reduce mortality risk, placing medication adherence on the causal pathway between telemonitoring and mortality.
- Specialist referral (collider): Specialist referral may be influenced both by telemonitoring, which can identify patients requiring specialist input, and by worsening health status associated with increased mortality risk. Adjusting for specialist referral may therefore introduce bias by creating a spurious association between telemonitoring and mortality.
- Multimorbidity (moderator): The effect of telemonitoring on mortality may differ according to the number and complexity of chronic conditions a patient has. For instance, patients with multiple chronic conditions may benefit more from continuous monitoring than patients with a single condition.

While DAGs help make causal assumptions explicit and identify potential sources of bias, they do not by themselves estimate causal effects. A range of analytical methods has therefore been developed to translate these assumptions into effect estimates, each relying on specific assumptions and addressing confounding in different ways. A few of them will be discussed in the next section.

3.5 Analytical methods for causal inference

3.5.1 Propensity score methods

Propensity scores **summarise multiple baseline characteristics into a single quantity** representing the **probability that a participant receives the intervention**, conditional on observed pre-intervention covariates.¹⁴⁹ Thus, whereas the other techniques discussed in this section model the outcome, propensity score methods **model the assignment to the intervention**. In practice, propensity scores are most commonly estimated using logistic regression, yielding an estimated probability between zero and one.^{149, 175}

Propensity scores are not of substantive interest in themselves, but function as a **design tool to improve covariate balance** between exposure groups and thereby **support causal effect estimation**. Accordingly, the adequacy of a propensity score model is judged solely by its ability to balance observed confounders after application, rather than by how accurately it represents the true treatment assignment mechanism or how interpretable the model parameters are.¹⁷⁵

In this context, balance refers to intervention and control groups being similar with respect to their observed baseline covariates. Individuals with similar propensity scores may still differ on specific characteristics, so the aim is not exact one-to-one equivalence, but **comparability at group level**. A valid propensity score should achieve balance on all relevant confounders, including variables associated with both exposure and outcome. Because the full set of confounders can never be verified empirically, propensity score models must be justified by showing that the selected covariates plausibly capture the relevant confounders and that an acceptable balance is achieved after implementation.¹⁷⁵

There is no consensus in the literature on **which variables should be included** in a propensity score model. While it is theoretically motivated to include predictors of intervention assignment,¹⁴⁹ others argue that variables associated with the outcome should also be included, even if they are unrelated to assignment.¹⁷⁶ Because the full set of relevant (observed and unobserved) confounders can never be empirically verified, covariate selection must be justified substantively, by demonstrating (based on existing theory and empirical evidence) that the chosen variables plausibly approximate the true confounder set.¹⁷⁵ In practice, this often leads to the inclusion of all measured baseline characteristics, as many are likely to affect treatment assignment and/or the outcome.¹⁴⁹



Compared with conventional regression approaches, propensity scores reduce reliance on the correct specification of the exposure-outcome model and enable the simultaneous adjustment for a large number of confounders when constructing comparable groups.¹⁷⁵ However, this does not eliminate the need for modelling, but rather shifts it to the specification of the intervention assignment model, which must be correctly specified to obtain valid estimates. Furthermore, propensity score methods make it easier to assess whether baseline differences between groups are adequately controlled, as one can directly verify covariate balance between groups, whereas regression adjustment does not.¹⁴⁹ Finally, while both approaches rely on sufficient overlap in covariate distributions between groups, this requirement is particularly central for propensity score methods, as they explicitly depend on the ability to construct comparable groups within regions of common support.

Propensity scores have several applications, including matching, trimming, weighting, stratification, and covariate adjustment.^{147, 149, 171, 177} Stratification compares outcomes within groups of individuals with similar propensity scores, while covariate adjustment incorporates the score as a predictor in regression models.¹⁷¹ However, these methods tend to perform weaker.¹⁴⁹ Therefore, this chapter focuses on matching, trimming, and weighting, as these most directly target covariate balance at the design stage.

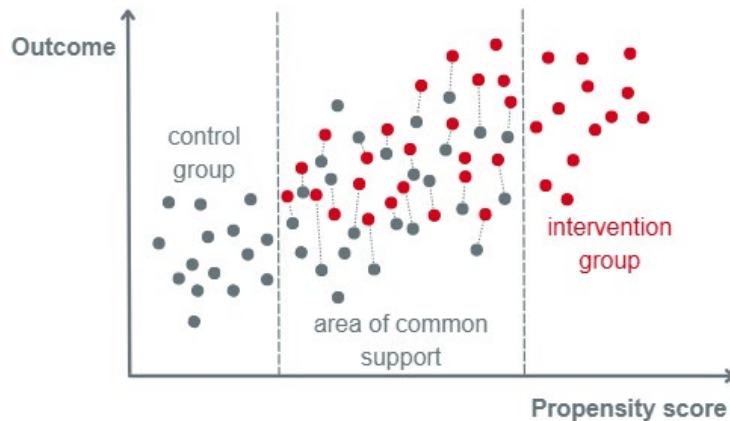
3.5.1.1 *Propensity score matching*

Propensity score matching addresses the practical limitations of matching on many characteristics (i.e. the aforementioned curse of dimensionality) by summarising multiple baseline variables into a single score. This makes matching more manageable and allows individuals in the intervention group to be paired with one or more control individuals who have similar propensity score values.¹¹⁵ In practice, propensity score matching involves estimating the score for each individual and then forming pairs or sets of individuals with similar values. This is typically done using nearest neighbour algorithms, optionally combined with restrictions such as calipers to avoid poor matches.^{115, 149} Importantly, this approach does not aim to make individuals identical, but to ensure comparability between intervention and comparison groups on a group level.

Checking the balance in baseline covariates between groups after propensity score matching is essential but often poorly done and reported. Relying on p-values is misleading because they reflect sample size rather than the magnitude of the imbalance. A more informative approach is to report standardised bias, which expresses group differences on a comparable scale and indicates how large any remaining imbalance is. Balance checking can be an iterative process, involving refinement of the propensity score model or matching parameters such as the caliper width. When imbalance persists for key covariates, adding interaction and higher-order terms should be considered to improve balance, or the use of automated propensity score model building methods.¹⁷⁵

It is important to note that, when propensity scores are near the extremes, comparable counterparts may not exist. This occurs when those who received the intervention have scores close to one, while others have scores close to zero, indicating a lack of common support. Matching is therefore typically restricted to the region where scores overlap, as depicted in Figure 22. As a result, causal estimates apply only to this matched subset rather than the full population. When overlap is very limited, comparisons must either be restricted to a small similar group or considered infeasible.¹⁴⁹ This also highlights a trade-off: more accurate propensity score models may reduce overlap between groups, leading to fewer matches and a smaller analysis sample, which can limit precision and generalisability.¹⁷⁸

Figure 22 – Propensity score matching and propensity score trimming



3.5.1.2 Propensity score trimming

As discussed in the context of propensity score matching, individuals located at the extremes of the propensity score distribution may lack comparable counterparts, indicating limited or no common support between intervention groups. In such cases, credible comparisons cannot be made for these observations. Rather than relying on matching procedures to implicitly exclude such individuals, propensity score trimming makes this restriction explicit by **defining a region of overlap in advance** and discarding observations outside this range.¹⁴⁷

Propensity score **trimming restricts the analysis to individuals for whom both intervention and comparison groups are sufficiently represented**, thereby improving comparability and reducing reliance on extrapolation into regions where no comparable observations exist. This typically involves excluding individuals with very high or very low propensity scores. By focusing on this region of overlap, trimming can improve internal validity, as estimates are based on more credible comparisons, but at the expense of external validity, since results no longer generalise to the full sample. The choice of trimming threshold reflects this trade-off, balancing the retention of a larger sample against improved comparability, and can be guided by formal procedures to identify an appropriate region of overlap.¹⁴⁷

Propensity score matching and trimming address complementary issues. Matching is suitable when sufficient overlap exists to identify comparable individuals across groups.¹⁴⁸ Trimming is particularly useful when this overlap is limited, as it restricts the analysis to observations for which credible comparisons can be made. In practice, trimming may be used to define a region of common support, after which matching is applied within that subset.¹⁴⁹

3.5.1.3 Propensity score weighting

Instead of creating matched subsets, **propensity score weighting uses the propensity score to reweight individuals based on the inverse of their probability of receiving the observed exposure**. In doing so, it reduces the association between confounders and exposure by giving greater weight to individuals who received an exposure level that was unlikely given their covariate profile.^{149, 179} The resulting weighted sample can be interpreted as a pseudo-population in which the distribution of measured confounders is similar across exposure groups, enabling a more credible comparison of outcomes.¹⁷⁹⁻¹⁸¹

In its most common form, inverse probability of treatment weighting (IPTW) weights are calculated as $1/PS$ for exposed individuals and $1/(1-PS)$ for unexposed individuals (with PS denoting the propensity score), after which the treatment effect is estimated in the weighted sample. In practice, this involves three steps:¹⁷¹

- estimating the propensity score,



- calculating the weights, and
- incorporating them into the outcome analysis.

A practical example of how these steps are carried out will be provided later in Textbox 8.

More generally, propensity scores can be used to construct weights that address other sources of bias. For example, inverse probability of censoring weighting (IPCW) can adjust for selective dropout, and inverse probability of sampling weighting (IPSW) can account for non-representative sampling. In each case, the aim is to create a pseudo-population in which the relevant selection mechanism is independent of observed characteristics.^{171, 181}

Compared with matching, propensity score weighting **retains most observations in the analysis**, which can increase effective sample size and improve statistical power.¹⁷⁹ It is also flexible and can be applied to a wider range of exposure types, including categorical and continuous treatments.¹⁷⁹ More generally, by summarising multiple confounders into a single score, weighting can be particularly useful in settings with many covariates or limited numbers of outcome events.¹⁷⁹ However, weighting methods are sensitive to extreme propensity scores, which can lead to unstable weights and reduced precision.^{182, 183} In addition, their validity depends on the correct specification of the propensity score model, as misspecification may result in residual imbalance and biased estimates.¹⁷⁹

3.5.2 G-computation

Rather than trying to balance exposure groups directly, g-computation adjusts for confounding by modelling the outcome as a function of both the intervention and relevant covariates, thereby accounting for the confounder-outcome relationship.¹⁷¹ The method builds on the potential outcomes framework mentioned in section 3.4.2.2, in which each individual has an outcome under exposure and an outcome under non-exposure, of which only one is observed.¹⁸⁴ The aim is to recover the ATE, defined as the mean difference between these two potential outcomes in the target population.^{171, 184}

In practice, g-computation proceeds in four steps. First, an outcome model is fitted on all observed data. Second, this model is used to predict counterfactual outcomes for each individual, both with and without the intervention. Third, individual treatment effects are calculated as the difference between these predictions and averaged to obtain the ATE. Finally, uncertainty is typically quantified using bootstrapping or Monte Carlo simulation.^{171, 180, 184} This process is illustrated in Textbox 7.

Textbox 7 – Illustrative example of g-computation

Imagine a study evaluating whether a structured medication review improves a medication appropriateness score. The DAG (directed acyclic graph, see section 3.4.3) created for this study showed age and number of chronic conditions to be important confounders. G-computation estimates the intervention effect as follows:

Step 1: Use available data to fit an outcome model and predict the counterfactual data:

$$\begin{aligned} \text{Predicted outcome}_i \\ = \beta_0 + \beta_1 \cdot \text{medication review}_i + \beta_2 \cdot \text{age}_i + \beta_3 \cdot \text{chronic conditions}_i + e_i \end{aligned}$$

Step 2: Predict outcome with and without the intervention and calculate individual treatment effect:

id	Age	Chronic conditions	Intervention: Medication review	Observed appropriateness score	Predicted appropriateness score without review	Predicted appropriateness score with review	Individual effect
1	68	1	0	12	9	12	3
2	75	3	1	22	21	30	9
3	70	2	1	18	18	25	7
4	82	4	0	35	25	34	9
5	66	1	1	10	10	10	0
6	78	3	0	28	20	28	8

Average: 6

Step 3: Average over the individual treatment effects to calculate the average treatment effect. Use Monte Carlo simulation and/or bootstrapping to account for uncertainty:

Interpretation: The medication review increases the appropriateness score by 6 (95% CI: [9; 3]) points on average.

G-computation can strengthen internal validity, provided that the key assumptions for causal inference are plausible, notably the **absence of unmeasured confounding**.¹⁸⁴

In addition to these standard assumptions, a fourth criterion is required for valid inference when applying g-computation: **correct specification of the outcome model**.¹⁸⁴ Because adjustment for confounding is achieved through the outcome prediction model, this model must adequately capture the true relationships between the outcome, exposure and confounders, including their interactions and functional form.¹⁷¹ If the model is misspecified, confounding may not be fully controlled, leading to biased estimates, even when the assumptions of exchangeability, positivity and consistency hold in principle.¹⁸⁴ Ensuring adequate model specification typically relies on **substantive expert knowledge** to correctly specify functional forms and interactions, although data-driven approaches such as machine learning may offer useful support in this step.^{185, 186} Another reason why the use of machine learning models is worth considering, is that the samples it requires to train the outcome prediction model may be less large than classical approaches.¹⁸⁶



3.5.3 *Doubly robust estimators*

For each of the methods mentioned above (regression adjustment, propensity-based methods, and g-computation), it is essential that the underlying models adequately reflect the true data-generating process⁹. These approaches rely on correct model specification to yield valid conclusions about intervention effectiveness. In practice, however, these processes are often complex, and perfect specification is unlikely. As a result, even small model misspecifications can introduce bias into the estimated effects.¹⁷¹

Doubly robust estimators offer a way to mitigate this issue by combining two modelling approaches, typically one for treatment assignment (such as propensity score methods) and one for the outcome (such as regression adjustment or g-computation). For instance, one of the most common doubly robust estimators combines a multivariable outcome regression model with inverse probability weighting. This provides two opportunities to capture the true data-generating process, so that the estimator remains unbiased as long as at least one of the two models is correctly specified.¹⁸⁷

A helpful way to understand this property is to view the doubly robust estimator as the sum of the intervention effect and an additional correction term. This term can be interpreted as the product of two bias components: one from the treatment model and one from the outcome model. If either model is correctly specified, its corresponding bias component is zero. Because the two components are multiplied, the correction term is eliminated, so the estimator remains unbiased even if the other model is misspecified. In practice, this also means that when both models are only slightly misspecified, the resulting bias may still be small, since it reflects the combination of two small errors. However, when both models are substantially misspecified, the bias can become large and the estimator may perform worse than either method used on its own.^{171, 181}

This increased robustness comes with a trade-off. While doubly robust estimators reduce bias when at least one of the two models is well specified, they are typically less precise than a correctly specified single-model approach, such as g-computation. In practice, this means that the estimated effects may fluctuate more from one sample to another, even if they are, on average, closer to the true effect. Thus, doubly robust methods prioritise protection against bias over precision, which can be advantageous in settings where model misspecification is a concern but may result in wider uncertainty around the estimated effects.^{188, 189}

⁹ Real-world process that produces both the treatment assignment and the outcomes, along with all observed data, according to underlying causal and stochastic mechanisms.

Textbox 8 – Practical example propensity score matching

Imagine a programme concerning a new staffing policy introducing flexible shift scheduling that improves nurse job satisfaction. To evaluate this programme, we compare nurses working in hospitals implementing this intervention to nurses working in hospitals without this policy. For both groups, data will be collected before and after the start of the intervention. In this example, the effect will be estimated with a DiD analysis using a doubly robust estimator combining inverse probability weighting. The doubly robust estimator calculates the intervention effect as follows:

Step 1: Calculate the propensity score for each individual in the dataset

$$\text{logit}(PS_i) = \gamma_0 + \gamma_1 \cdot \text{Age}_i + \gamma_2 \cdot \text{Years of experience}_i + \gamma_3 \cdot \text{Department}_i + e_i$$

Step 2: Calculate the IP weights for each individual as:

$$\omega_i = \frac{1}{PS_i} \text{ for intervention group}; \quad \omega_i = \frac{1}{1 - PS_i} \text{ for control group}$$

id	Intervention	Time	Job satisfaction	Age	Years of experience	Department	PS	IP weight
1	0	Before	6.2	34	8	Internal medicine	0.30	1.43
1	0	After	6.1	34	8	Internal medicine	0.30	1.43
2	1	Before	5.8	42	15	Surgical	0.60	1.67
2	1	After	6.8	42	15	Surgical	0.60	1.67
3	0	Before	6.5	37	10	Internal medicine	0.40	1.67
3	0	After	6.4	37	10	Internal medicine	0.40	1.67
4	1	Before	5.5	50	20	ICU	0.75	1.33
4	1	After	6.6	50	20	ICU	0.75	1.33

Step 3: Adjust for the confounders in the weighted (IP weights) outcome regression model

$$\text{Outcome}_i = \beta_0 + \beta_1 \cdot \text{Intervention}_i + \beta_2 \cdot \text{Time} + \beta_3 \cdot \text{Intervention}_i \cdot \text{Time} + \beta_4 \cdot \text{Age} + \beta_5 \cdot \text{Years of experience} + \beta_6 \cdot \text{Department}_i + e_i$$

Interpretation: The staffing policy increases nurse job satisfaction by approximately β_3 points.

3.5.4 Instrumental variables

A key challenge in estimating causal effects is that the observed relationship between an exposure or treatment and an outcome typically combines two sources of variation:

- **Confounded variation**, which reflects biases due to omitted variables
- **Unconfounded variation**, also called ‘as-if-random’, which carries the true causal signal.

Instrumental variables (IV) methods aim to isolate this latter component by leveraging an external factor, called an instrument, that influences the exposure but is independent of the confounders affecting the outcome. By using only the variation in the exposure that is induced by this instrument, IV mimics a randomised experiment by isolating variation in exposure that is as-if random.

This variation arises from individuals whose exposure changes in response to the instrument, and the resulting estimate therefore captures a causal effect for this subgroup that is not biased by confounding. Importantly, this effect pertains to individuals whose exposure is influenced by the instrument. This is known as a local average treatment effect (LATE). In other words, IV does not

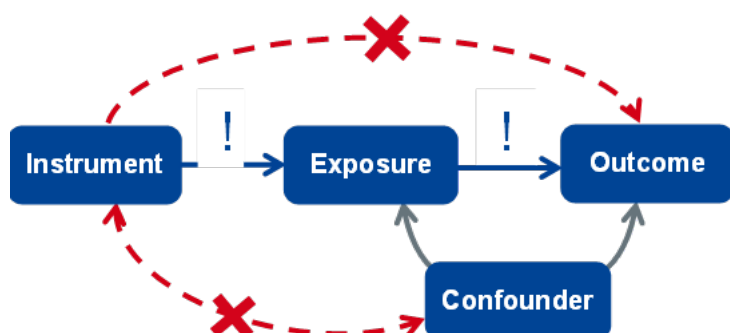


remove selection bias, it isolates variation in treatment that is not driven by selection and estimates the effect in that part of the population.^{117, 143} Textbox 9 this with an example.

A valid instrument must satisfy four conditions to ensure that it identifies a causal effect rather than introducing new biases. These criteria formalise the idea that the instrument should mimic random variation in the exposure while remaining otherwise unrelated to the outcome. The criteria (of which the first three are illustrated in Figure 23) are the following:^{118, 143, 190}

- **Relevance:** The instrument must be strongly associated with the exposure
- **Exclusion restriction:** The instrument should affect the outcome only through its effect on the exposure, and not through any alternative pathways
- **Independence:** The instrument must be independent of any unmeasured confounders that influence both the exposure and the outcome
- **Monotonicity:** An instrument must affect the treatment uptake in the same direction for all individuals, it cannot have the opposite effect on some.

Figure 23 – The criteria for an instrumental variable



In the analysis stage, IV is usually carried out in two steps. First, the exposure is predicted using the instrument, capturing only the part of the exposure that is driven by the instrument. Note that it is **possible to use multiple IVs** to estimate the effect of a single intervention.¹⁹¹

$$\text{Predicted exposure}_i = \gamma_0 + \gamma_1 \cdot \text{instrument}_i + \gamma_2 \cdot \text{covariates}_i + e_i$$

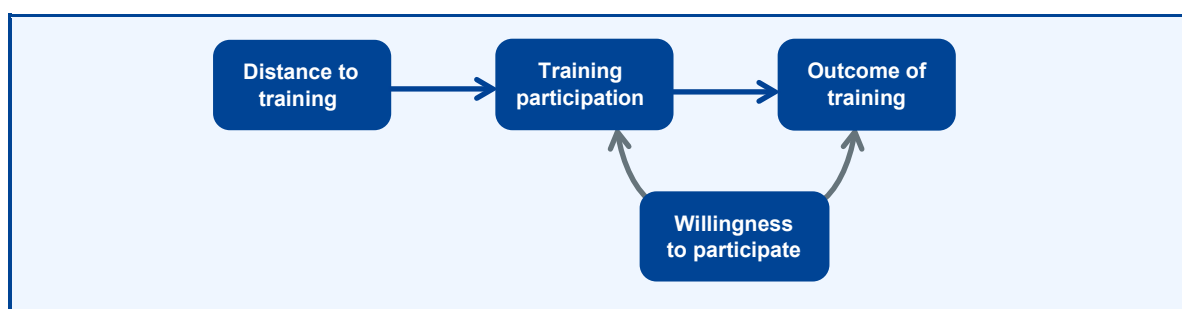
Then, the outcome is analysed using this predicted exposure, which allows estimation of the causal effect based on variation that is not influenced by confounding.

$$\text{Outcome}_i = \beta_0 + \beta_1 \cdot \text{Predicted exposure}_i + \beta_2 \cdot \text{covariates}_i + e_i$$

Textbox 9 – An illustrative example on instrumental variables

Imagine an evaluation of a training programme for health practitioners when participation is voluntary. Practitioners who choose to attend are likely to differ from non-participants in ways that also affect outcomes. For example, they may be more motivated, more open to changing their practice, or more convinced that the training will benefit them.

In this setting, the distance from a practitioner's workplace to the training location can serve as an instrument: it influences the likelihood of attending but is unlikely to affect outcomes except through participation. This helps isolate variation in training uptake that approximates random assignment. The resulting estimate can be interpreted as the effect of the training among practitioners whose participation is influenced by distance to the training location (i.e. the LATE). The figure below illustrates the DAG of this example.



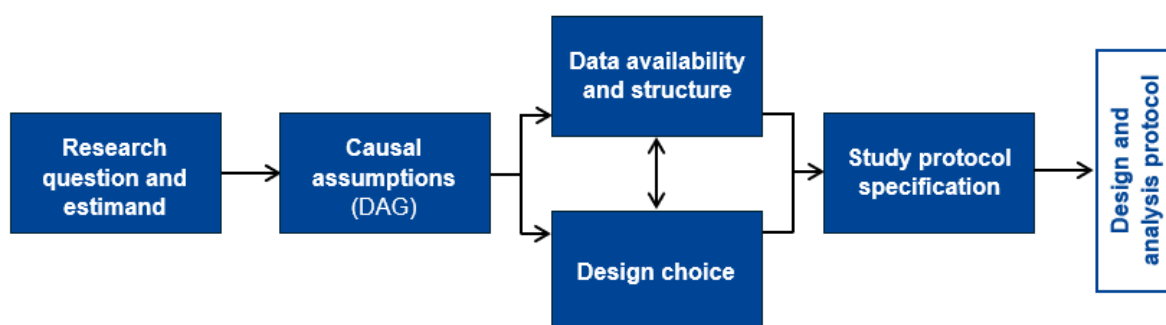
Instrumental variable methods offer an attractive solution to unobserved confounding in observational data.¹⁹⁰ By exploiting variation in exposure that is driven by an external instrument rather than by individual choice or programme implementation, IV focuses on changes in treatment uptake induced by that instrument, allowing for evaluation even in the presence of imperfect compliance or voluntary participation that would otherwise introduce selection bias.¹¹⁷ This makes IV particularly valuable in settings where randomisation is not feasible.

However, these advantages come with **important limitations**. Credible instruments are often difficult to find, and the validity of the approach relies on strong assumptions that cannot be directly tested.¹⁹¹ In particular, the main concern about unmeasured confounding between the exposure and the outcome is effectively replaced by assumptions about the instrument, such as its independence from unmeasured factors that affect the outcome and the absence of alternative pathways.¹⁹⁰ Moreover, as IV estimates a LATE, the findings may not be generalisable to the entire population but instead apply to the subgroup whose treatment is influenced by the instrument. It is therefore important to consider whether this subgroup is relevant to the research question of interest. Finally, IV estimates are typically less precise and therefore often require large sample sizes.¹⁹⁰

3.6 A practical guide to developing a quasi-experimental design and analysis protocol

The process of designing a quasi-experiment can be summarised in five components (see Figure 24), culminating in the design and analysis protocol as the final end product. Although the process is presented as broadly linear, this is an oversimplification because feedback between stages is often required. In particular, the stages of *data availability and structure* and *design choice* are considered together: the available data may constrain which designs are feasible, while the chosen design may determine which data need to be requested or collected. Each of the components is discussed in the sections below.

Figure 24 – Outline of the five components of quasi-experimental study design





3.6.1 Component 1: Research question and estimand

The first step in designing a quasi-experimental study is to establish a first conceptual definition of the research question. The research team must develop a clear understanding of the problem the study is intended to address and the practical and policy context. This requires some exploratory work, in close interaction with the programme organisers to establish the estimand of the project, which is defined by the following elements:

- **Population:** Who is affected by the intervention? (e.g. care providers, patients, informal caregivers, support/administrative staff, management)
- **Intervention:** What exactly does the intervention entail? It is particularly important to explicitly consider this for complex interventions consisting of a multitude of sub-interventions.
- **Comparator:** This is most commonly current practice.
- **Outcome:** It is important to gain an understanding of what the programme organisers aim to achieve with the programme. Potential adverse outcomes should also be considered. A primary outcome will later need to be selected for power considerations.
- **Time horizon:** How long will the project run? How likely is it that the supposed outcomes of the intervention could be observed within that timeframe?
- **Effect:** Are we interested in the average treatment effect in general, the average treatment effect on the treated or a local average treatment effect for a particular subgroup?

At this stage, these definitions can still be preliminary. They can be idealised and not yet constrained by feasibility or data, as the purpose is to clarify what should be estimated before selecting a design.

3.6.2 Component 2: Make causal assumptions explicit

Once the research question and estimand have been defined, the next step is to make the underlying causal assumptions explicit. In the context of evaluating a pilot programme, this step is critical to understand how the intervention may generate observed outcomes and where bias may arise.

The starting point is **to identify all variables that are causally related to the exposure, the outcome, or selection into the study**. This requires gathering information from multiple sources, such as expert input (including those familiar with the organisation and implementation of the pilot) and existing literature on similar interventions and outcomes. This information can be structured and interpreted using conceptual frameworks, such as the theory of change underlying the programme (see Chapter 1) or broader systems thinking perspectives (see Chapter 10), to capture interactions across organisational and care delivery levels.

Importantly, variables should not be restricted to those that are readily available in the data. Variables that are unmeasured, difficult to measure, or only partially observed should still be considered, as omitting them implicitly assumes that they play no causal role. The identified variables and their assumed relationships should then be formalised in a **directed acyclic graph (DAG)** (see section 3.4.3), to provide a transparent representation of the assumed causal structure underlying the intervention and its outcomes. A detailed, step-by-step guide to constructing and interpreting DAGs is provided in the work of Feeney et al.,¹⁹² which is recommended for further guidance in this stage of the study development.

Once specified, **the DAG serves as the foundation for subsequent design and analytical decisions**. It helps identify potential sources of bias, including confounding and selection mechanisms, and determines which variables should be adjusted for to obtain an unbiased estimate of the target effect. Equally importantly, it highlights variables that should not be adjusted for, such as mediators or colliders, as conditioning on these can introduce bias rather than remove it.

Developing the DAG is typically an **iterative process**. Initial versions should be refined through discussion with a broader group of experts until a reasonable level of consensus is reached.

Disagreements between experts are informative and should not be ignored. Instead, they can be formalised as alternative candidate DAGs, which can be used to guide sensitivity analyses.

Depending on the research question, **analyses may focus on only a subset of the relationships** represented in a broader DAG. Nevertheless, the broader causal structure remains important, as it provides the context for identifying potential sources of bias and assessing the plausibility of the assumptions underlying the analysis.

Finally, it is recommended to **pre-register the DAG** or at least document it before analysis begins. This ensures that the causal assumptions are specified independently of the data and results, and provides a transparent basis for design and analytical decisions.

3.6.3 Component 3: Data availability and structure

After defining the estimand and causal assumptions, the next step is to evaluate existing data and identify what additional information is needed, guided by the DAG and estimand. This requires mapping the variables in the DAG to available data sources (see also Chapter 11 – routinely collected data) and assessing whether they are observed, imperfectly measured, or missing.

Missing variables may signal residual confounding and should be addressed through design changes, sensitivity analyses, or refinement of the estimand. As measurement error and data gaps are common, these limitations should be acknowledged at the design stage and reflected in planned analyses, including sensitivity analyses to assess robustness. It is also important to consider how individuals entered the dataset, as selection mechanisms may introduce systematic differences that affect validity.

The data structure in which individuals, groups, and outcomes are observed over time has important implications for which evaluation designs can be implemented and how causal effects can be estimated. While data structures do not define causal designs themselves, they strongly influence which quasi-experimental approaches are most naturally applied. Different data structures and design combinations exist. Table 22 provides an overview of data structures with their implications towards design choice.

Table 22 – Overview of data structures and their potential applications

Data type	Description	Most common applications	Other possible but less common applications
Panel data	The same individuals (or clusters) are measured repeatedly over time, allowing direct observation of within-person change and control for time-invariant individual characteristics.	• Single-group pre-post designs • Difference-in-differences analysis	• Interrupted time series • Multiple baseline designs (NOTE: this requires sufficient repeated measurements)
Repeated cross-sectional data	Different samples are drawn from the same population at each time point, allowing assessment of population-level change but not change within individuals.	• Simple before-and-after comparisons • Difference-in-differences analysis	• Interrupted time series (NOTE: this requires sufficient repeated population aggregates)
Aggregated time series data	Repeated measurements are collected over time for an aggregate unit (e.g. hospital, school, region), capturing temporal variation without individual-level information.	• (Controlled) Interrupted time series designs • Multiple baseline designs	• Before-and-after comparisons • Difference-in-differences analyses in case of multiple time series (NOTE: Although these are feasible, methods that explicitly model temporal patterns are preferred)



3.6.4 Component 4: Evaluation design choice

Selecting an appropriate evaluation design requires balancing methodological rigor with the practical constraints imposed by the intervention and available data. Table 23 presents an indicative ranking of quasi-experimental designs according to their internal validity, assuming reasonable implementation and that key assumptions are met. Designs within the same category are broadly comparable, and the credibility of evidence depends on implementation quality, data quality, and the plausibility of assumptions.

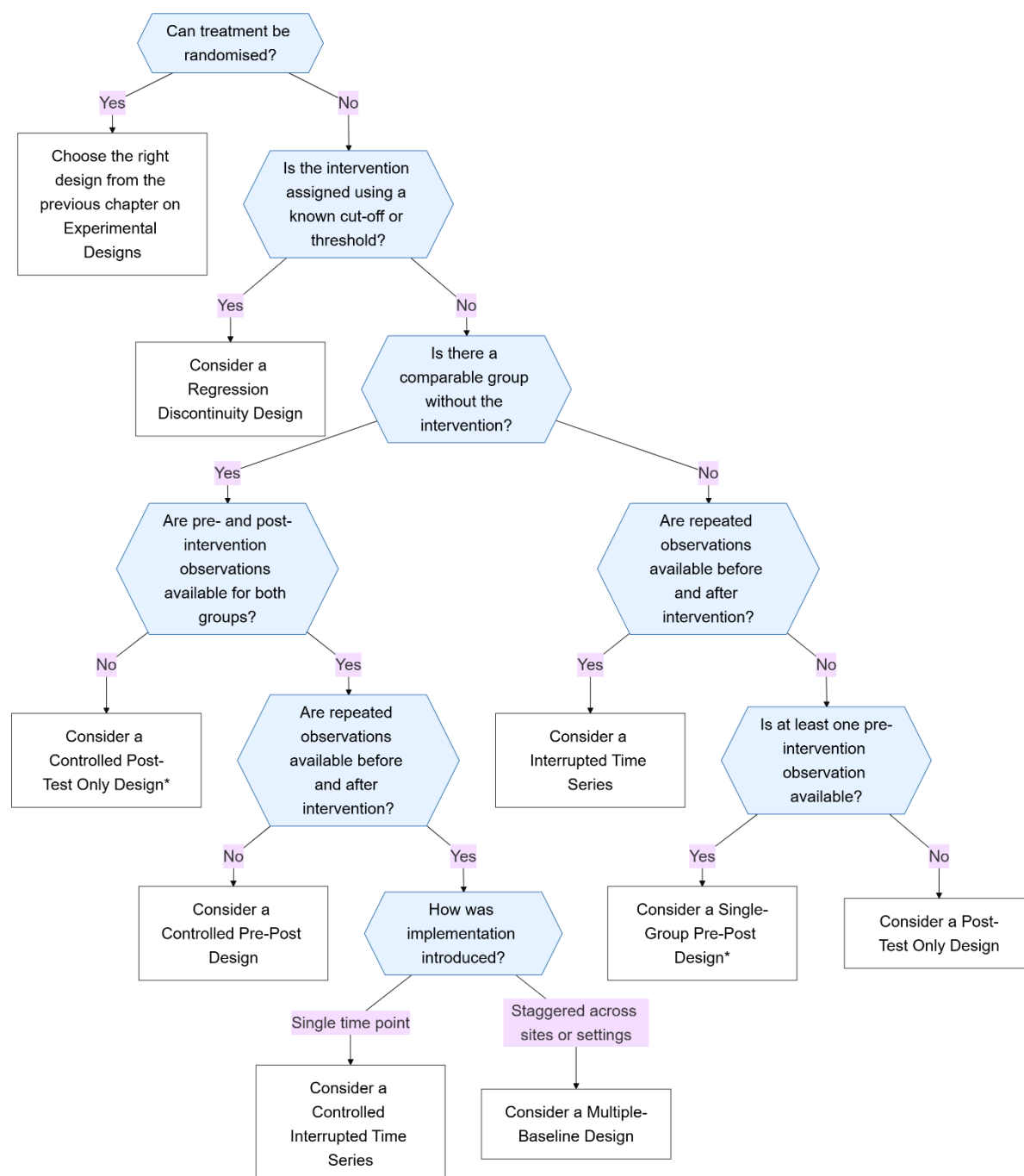
Table 23 – An indicative ranking of quasi-experimental designs

Internal validity	Designs	Comment
High	(Controlled) Interrupted Time series Multiple Baselines Design Regression Discontinuity Design	Strong protection against several sources of bias, provided key design assumptions are met.
Medium	Controlled Pre-post Design	More pragmatic and widely applicable, but more dependent on assumptions regarding group comparability.
Low	Single Group Pre-post Design Controlled Post-test Only Design	High susceptibility to bias. Researchers should carefully consider whether the design is capable of providing evidence that is sufficient to answer the research question and justify the resources required for the study.

The presented ranking should be interpreted as indicative rather than absolute. It reflects the typical internal validity of each design when considered in isolation. The validity of quasi-experimental evaluations also depends on the plausibility of the underlying causal assumptions and on the analytical methods used. Appropriate design adaptations and causal inference techniques may substantially strengthen a study (see sections 3.2.3, 3.3.3, 3.3.5, and 3.5), whereas violations of key assumptions may undermine even theoretically strong designs.

While some designs may generally provide stronger protection against bias than others, not all designs are feasible in every evaluation setting. The decision flowchart in Figure 25 is intended to help researchers identify designs that may be appropriate given the characteristics of the intervention and the available data. The flowchart should be viewed as a guide rather than a prescriptive algorithm. Identifying a design through the flowchart does not necessarily mean it is the most appropriate option. Researchers' own judgment is key in assessing whether the design's assumptions are plausible, whether it matches the available data structure, whether it can adequately answer the research question, and whether it captures the timing and nature of intervention effects.

Figure 25 – Decision flow chart to assist design choice



*: Recommendation: Consider whether the design is capable of providing the evidence needed to adequately answer the research question and justify conducting the study.

3.6.5 Component 5: Study protocol specification

This section builds on the target trial emulation framework introduced earlier and translates it into practical design considerations. For retrospective observational studies, the use of target trial emulation is worth considering such as the work of Seewald et al.¹²⁶ on TTE applied in health policy research.¹²⁶ However, the same TTE structure covered below can also guide the specification and operationalisation of prospective quasi-experiments, from eligibility criteria to power calculation.



UNITS AND ELIGIBILITY CRITERIA

This component specifies the policy-level units that could implement the policy (such as organisations) and the population or sub-units affected by the policy (such as individual care providers or patients), along with eligibility criteria for both, defined using information available at baseline. Relevant questions include:

- What are relevant inclusion and exclusion criteria?
- If prospective: Which subgroup of the target population is likely to be underrepresented?
- If retrospective: Are sufficient historical data available to establish eligibility at baseline?
- Is it required that all individuals in an organisation participate in the intervention? (e.g. all GPs in a participating GP practice? What about interns and temporary workers?)

EXPOSURE AND COMPARISON CONDITIONS

A thorough understanding of what policies exist around the topic of interest is critical, including mapping which (combination of) policies are considered the exposure and which policies could be potential confounders. Relevant questions include:

- What counts as treated? (e.g. Is there a minimum required exposure duration? Is there a minimum implementation fidelity? Should all aspects of the intervention be implemented as a whole or is partial adoption possible?)
- Is the recruitment of controls independent of that of the intervention group or not? (e.g. due to matching or the same stratification)
- Are there any contextual changes during the study period affecting intervention and exposure group unequally that may (have) impact(ed) the outcome? (e.g. changes to nomenclature or introduction of new policies) How can they be accounted for?

ASSIGNMENT MECHANISM

When random assignment is impossible, it is important to identify the factors that can cause differences in policy adoption and the outcome of the intervention. To strengthen internal validity, these should be accounted for using an appropriate analytical method, of which several are discussed throughout this chapter. Relevant questions include:

- In case of voluntary participation: Which factors can contribute to the decision to participate, and can they be measured (e.g. can motivation and belief in the positive impact of the intervention be quantified in surveys?) Or is there a variable that could potentially serve as an **instrument**?
- If necessary: How likely are parallel trends between the intervention and control groups?
- If necessary: Which analytical technique will be used to account for confounding and/or selection bias?

TEMPORAL STRUCTURE (TIME ZERO)

A clear time zero is defined as the point at which the policy could begin to affect outcomes. Incorrectly defining the baseline can introduce bias: defining it too late may capture anticipation effects, while defining it too early can dilute the observed treatment effect by including time before the treatment has any impact. Bias introduced by time zero misspecification is common in observational studies and, therefore, a key element in target trial emulation. For evaluation studies using observational data, the work of Kutcher et al.¹⁹³ is recommended for more information on the problem and potential solutions. Relevant questions include:

- What should be the exact rule for time zero? (e.g. first eligibility for intervention, actual initiation of intervention, policy implementation date, recruitment/enrolment date, ...)



- Are there likely anticipation effects?
- Is the intervention start on the same day for all participants or is implementation staggered?

OUTCOMES

Outcome variables should be operationalised, and the appropriate variable(s) should be selected for power calculation (usually the primary outcome). Relevant questions include:

- What is the primary outcome?
- Which type is each outcome variable? (e.g. binary, count, continuous, time-to-event)
- When and how often will the outcome be measured? (e.g. single post-baseline assessment; repeated measures over time; cumulative outcome over follow-up)

ESTIMAND

The initial estimand from component 1 (see section 3.6.1) should be revisited if necessary.

STATISTICAL ANALYSIS AND POWER ANALYSIS

The statistical considerations presented in Chapter 4 are applicable in quasi-experiments. In prospective studies, a sample size calculation should be made to establish recruitment targets. In retrospective studies, it can be useful in some contexts to calculate **power** from the available sample size. Relevant questions include:

- How will clustering or repeated measures be handled?
- What assumptions are required and how will they be assessed?
- Which sensitivity analyses would be useful?



4 STATISTICAL CONSIDERATIONS FOR EVALUATION STUDIES

KEY POINTS

- **Outcomes should be selected, operationalised and prioritised early**, as they influence data collection, sample size requirements, and the interpretation of study results. Different ways of measuring the same underlying concept may lead to different interpretations and different statistical properties.
- **The research hypothesis should be specified before data collection begins**, including whether the objective is to demonstrate a difference in any direction, superiority, non-inferiority, equivalence, or another predefined effect.
- **Statistical power quantifies the probability of detecting a true effect**. Sample size calculations provide an estimate of the number of participants required to address the research question with an acceptable probability of detecting a meaningful effect.
- **Power calculations depend on assumptions that require subject-matter expertise**. Estimates of effect sizes, variability and other key parameters should be informed by the best available evidence and (expert) knowledge of the study context.
- **Multiplicity increases the risk of false-positive findings**. Sources of multiplicity include multiple outcomes, multiple treatment comparisons, multiple subgroup analyses or adaptive interim analyses, and appropriate strategies should be considered during study planning.
- **Different analytical approaches answer different questions**. In general, intention-to-treat analysis is preferred over per-protocol analysis as it preserves the benefits of randomisation.
- **Intercurrent events should be anticipated and handled explicitly**. Important events that occur after implementation of the intervention can introduce bias in the study results.
- **Correlation between clustered observations is common in healthcare data and should be accounted for**, as they affect data collection, sample size requirements and analytical methods
- **Sensitivity analyses help assess robustness of the findings**. Conclusions that remain consistent across reasonable alternative analyses are generally more credible than conclusions that depend strongly on specific assumptions.

4.1 Defining the research questions

4.1.1 Outcomes

The choice of outcomes is a fundamental part of defining the research question, as outcomes determine how the success of an intervention will be assessed. Outcomes should be selected early during study planning because they influence data collection procedures, the planned statistical analyses, sample size requirements and ultimately the interpretation of the study results.

4.1.1.1 Primary and secondary outcomes

Studies evaluating complex interventions often assess multiple outcomes to capture their effects from different perspectives, including those of patients, healthcare providers, healthcare organisations and the healthcare system. **Primary outcomes** quantify the main measure of the effectiveness of the intervention and often drive design decisions such as sample size determination. **Secondary outcomes** provide additional information on other effects of the intervention. Secondary outcomes

may also include process outcomes, which assess whether the intervention changed the targeted processes of care. Because health policy interventions often affect multiple domains simultaneously, it is important to clearly prioritise outcomes and research questions to support both the statistical analysis and the interpretation of the findings. Programme Theory, as discussed in Chapter 1, can support such decisions.⁷⁰

When selecting key outcomes, the following aspects should be considered:

- **Relevance:** the outcome should reflect effects that are meaningful to stakeholders, including patients, healthcare providers, policymakers and insurance organisations.
- **Scientific acceptability:** the outcome should be consistent with current scientific knowledge and accepted practice within the field.
- **Feasibility of measurement:** the outcome should be measurable in a reliable and practical manner.
- **Follow-up duration:** the outcome should occur within a timeframe compatible with the study design.
- **Statistical considerations:** the outcome should allow appropriate and robust analytical methods
- **Ethically acceptability:** the endpoint and its measurement should not impose undue burden or risk on participants
- **Statistical power:** the endpoint should allow for a study with sufficient power given realistic sample size and budget constraints

In some studies, selecting a single primary outcome may be difficult because several outcomes are considered equally important. While it is possible to use multiple primary outcomes (also called 'co-primary endpoints'), this increases complexity. Multiple primary outcomes require additional decisions about how to define and interpret intervention success and also affect the required sample size (see section 4.2.3.1). A possible alternative approach is to combine multiple outcomes into a composite outcome. Another approach is to explicitly prioritise outcomes according to their importance and incorporate this prioritisation into the statistical analysis. Generalized Pairwise Comparisons (GPC), described in section 4.2.2.3, is an example of a statistical method that follows this principle. Both approaches require careful consideration and should be specified during study planning.^{70, 194}

4.1.1.2 Outcome operationalisation

Different types of outcomes require different statistical methods and may have implications for study design and sample size calculations. Table 24 presents some commonly encountered outcome types.

90 195

Table 24 – Types of outcomes

Outcome type	Description	Example
<i>Continuous</i>	Can take any value within a range	Cost of care; Visual Analogue Scale (VAS) for satisfaction with care
<i>Binary (dichotomous)</i>	Two possible categories	Hospital readmission within 30 days (yes/no); Received recommended preventive screening (yes/no)
<i>Count</i>	Number of events occurring during a defined period	Number of outpatient visits during follow-up; Number of emergency department visits in one year



Ordinal	Ordered categories	Likert scales of self-rated health; Cancer stage at diagnosis (stage I-IV)
Discrete score	Numerical score with a finite number of possible values	ADL score; Hospital quality of care score
Nominal	Categories without natural ordering	Primary mode of transport to access care; Type of healthcare provider first consulted (GP, specialist, emergency department, pharmacist, ...)

In general, variables that contain more information provide greater statistical power (for more on power, see section 4.2) and thus allow for smaller sample sizes. All else being equal, the following hierarchy can be considered: **continuous variables > discrete numeric variables (including counts) > ordinal variables > nominal and binary variables**. In clinical research, continuous outcomes are frequently dichotomised to facilitate interpretation or to align with clinical decision thresholds (e.g. using a binary indicator of length of stay being longer than five days or not instead of expressing length of stay in days). However, such dichotomisation comes at a cost, as it discards information, reduces statistical power, and may require substantially larger sample sizes to detect the same effect.¹⁹⁶

Outcomes represent concepts of interest (e.g. healthcare utilisation or health status) that must be operationalised into measurable variables. For example, healthcare utilisation can be measured as the number of hospital admissions, the admission rate, or time to first admission. Because different operationalisations may lead to different interpretations of intervention success, **outcome definitions should be explicitly discussed and agreed upon early with both the research team and relevant stakeholders**.

4.1.1.3 Surrogate outcomes

Sometimes the most relevant outcome is impractical to measure because it requires long follow-up or because its measurement is difficult, resource-intensive or burdensome for participants. In such cases, researchers may consider a surrogate outcome: an earlier or more easily measured outcome that is intended to substitute for the outcome of primary interest. For example, researchers may be interested in longer-term health outcomes, but only short-term indicators are available within the duration of the pilot programme. Process indicators or intermediate outcomes (see Chapter 1 on Programme Theory) may then be considered as surrogate measures of the expected policy impact.^{197, 198}

The use of surrogate outcomes requires caution because **a positive effect on a surrogate outcome does not guarantee a positive effect on the outcome of real interest**. For example, an intervention may improve guideline adherence or service uptake without improving patient outcomes. This is because the relationship between an intervention and a meaningful outcome is often more complex than can be captured by a single intermediate measure. Researchers should therefore critically evaluate whether the surrogate outcome is a credible indicator of meaningful intervention effects. Whenever feasible, direct measures of those effects should be used.

The suitability of a surrogate outcome should be evaluated using evidence from previous studies, by assessing both its association with the outcome of interest and the extent to which intervention effects on the surrogate reliably predict effects on that outcome.^{197, 198}

4.1.2 Hypotheses

4.1.2.1 Fundamentals of hypothesis testing

Once the key outcomes have been defined, the study must specify what it aims to demonstrate regarding those outcomes. This is done through one or more hypotheses, which underpin many subsequent statistical decisions.

In hypothesis testing, a null hypothesis (H_0) is assessed against an alternative hypothesis (H_a). The null hypothesis typically represents the absence of a meaningful intervention effect, while the alternative hypothesis represents the effect that the study aims to assess. In many studies, this corresponds to no difference between groups versus a non-null difference. However, other formulations may be more appropriate depending on the research question. The study evaluates the compatibility of the observed data with these competing hypotheses, commonly through confidence intervals and p-values. It is important to note that statistical significance should not be confused with a meaningful effect size.^{199, 200}

Because study conclusions are based on incomplete information, they are always subject to uncertainty. Two types of errors can occur:¹⁹⁹

- A **Type I error** (i.e. a false-positive result) occurs when a study concludes that an intervention has an effect when in reality it does not.
- A **Type II error** (i.e. a false-negative result) occurs when a study fails to detect an effect that truly exists.

Table 25 – Two types of error

		STUDY RESULT	
		H_0 rejected	H_0 not rejected
TRUTH	H_a true	OK <i>(true positive)</i> Probability = $1 - \beta$ (power)	Type II error <i>(false negative)</i> Probability = β
	H_0 true	Type I error <i>(false positive)</i> Probability = α	OK <i>(true negative)</i> Probability = $1 - \alpha$

4.1.2.2 Superiority, non-inferiority and equivalence studies

Different studies may aim to answer different research questions. The most common objective is to demonstrate that an intervention performs better than the comparator (superiority). However, other objectives may also be relevant.

Figure 26 illustrates several possible situations. The value 0 represents no difference between the interventions. The values $-C$ and $+C$ represent predefined margins whose interpretation depends on the research question.²⁰¹

- In a **superiority study**, the objective is to demonstrate that the intervention performs better than the comparator. In a **super-superiority study**, the objective is to demonstrate that the intervention exceeds a predefined threshold for a meaningful effect. In this case, $+C$ represents the minimum clinically or policy-relevant difference.
- In a **non-inferiority study**, the objective is to demonstrate that the intervention is not unacceptably worse than the comparator. Here, $-C$ represents the non-inferiority margin, i.e. the largest loss of effectiveness that would still be considered acceptable. Such studies may be

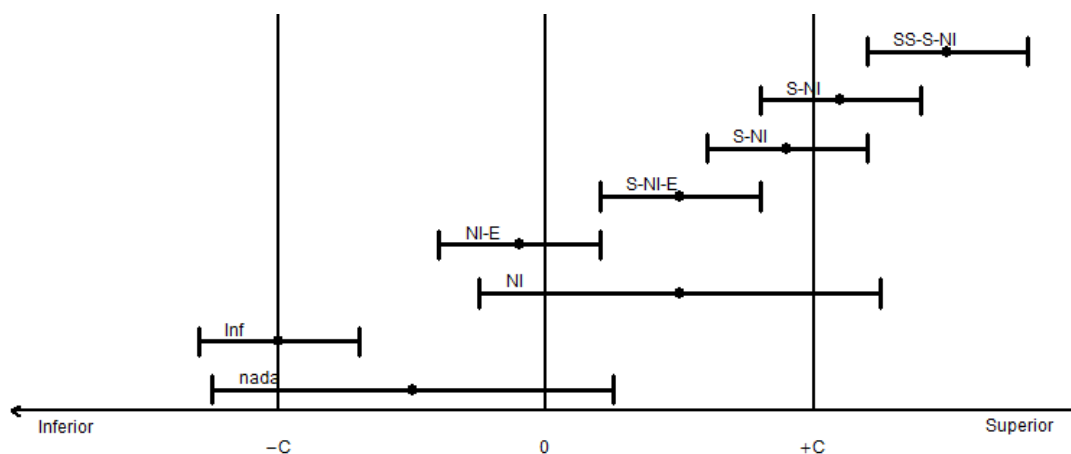


appropriate when an intervention offers other advantages, such as lower costs, easier implementation or improved accessibility.

- In an **equivalence study**, the objective is to demonstrate that any difference between interventions is sufficiently small to be considered unimportant. In this case, $-C$ and $+C$ define the lower and upper equivalence margins.

The research hypothesis also determines whether a one-sided or two-sided test is appropriate. A two-sided test considers both positive and negative differences between interventions and typically allocates 2.5% of a 5% significance level to each direction of the effect.²⁰¹ A one-sided test considers only one direction and allocates the full significance level to that direction, although a 2.5% one-sided level is often recommended for consistency with two-sided inference.⁸⁸ Within the KCE evaluation programme, two-sided tests are generally preferred because interventions may have both beneficial and unintended adverse effects. One-sided tests should only be used when justified by the research question (e.g. non-inferiority studies) and should be specified a priori.

Figure 26 – Superiority, non-inferiority and equivalence studies



Legend: S = Superior; NI = Non-inferior; E = Equivalent; $-C$ = Non-inferiority margin; SS = Super-superior; Inf = Inferior; $+C$ = Minimum important difference; nada = none of the above

Key considerations for the research team:

- Prioritise outcomes and decide which outcome(s) will determine the success of the intervention (see also Chapter 1).
- Operationalise outcomes carefully, taking into account feasibility, interpretability and the implications for statistical power.
- Agree on the research hypotheses for each key outcome before data collection starts.
- Consider whether surrogate outcomes are acceptable, and whether they are sufficiently credible proxies for the outcomes of real interest.

4.2 type II error: power analysis

4.2.1 Statistical power

In intervention studies, the number of study participants must be sufficient to distinguish a true change in the primary outcome measure from background noise. **Statistical power** reflects how likely a study is to achieve this. More formally, it is the probability that a study will detect an effect if such an effect exists. As introduced in section 4.1.2, power corresponds to $1 - \beta$, where β is the probability of a Type II error (false negative finding).¹⁹⁹

A study with insufficient power is referred to as **underpowered**. In such studies, an intervention may appear ineffective simply because the study was unable to detect an existing effect. Conversely, a study may be **overpowered** when substantially more participants are included than necessary. Although this increases the likelihood of detecting an effect, it may lead to unnecessary use of resources and to statistically significant findings that are too small to be of practical relevance.¹⁹⁹ Moreover, including more participants than required can lead to unnecessary costs, operational challenges, and ethical concerns, as participants may be exposed to research procedures and burdens without sufficient scientific justification.

Power analyses primarily address **random error** arising from chance variation in the data. Increasing the sample size generally improves the precision of estimated effects and reduces the influence of random error, thereby increasing the likelihood of detecting a true effect. However, larger sample sizes do not protect against **systematic error** (bias), which arises when study results are consistently distorted by aspects of the study design, implementation or data collection. Adequate power is therefore necessary but not sufficient to ensure valid study conclusions. Potential sources of bias must also be considered when designing and conducting a study, as discussed in Chapters 2 and 3.^{200, 202}

Sample size calculations can be based on different objectives. In studies focused on hypothesis testing, the required sample size is often determined through a power calculation. Power can be estimated given a set of key assumptions, including the expected distribution of the primary outcome and the magnitude of the anticipated effect size (this will be elaborated upon in section 4.2.2.2). The sample size is then chosen such that the study has a sufficiently high probability of detecting a clinically relevant effect if it exists.

However, not all studies aim primarily to test hypotheses. In some situations, the objective is to estimate a parameter, such as a prevalence, mean, or treatment effect, with a desired level of accuracy. In these cases, the sample size may be based on the desired precision of the estimate, which is commonly expressed through the width of a confidence interval. Larger sample sizes generally produce narrower confidence intervals and therefore more precise estimates. Precision-based sample size calculations are frequently used in descriptive studies, pilot studies, and studies that aim to provide reliable effect size estimates for future research.²⁰³

In prospective studies, either a power-based or precision-based calculation may be used to determine the required sample size, depending on the study objective. Studies based on existing data sources often have a fixed number of observations available. In such situations, the available sample size can be used to calculate either the statistical power to detect a given effect size or the precision with which key parameters can be estimated.

Sample size calculations are based on assumptions about quantities that are often unknown at the design stage, such as the expected effect of the intervention or the variability of the outcome. As discussed in the upcoming sections, several design choices and study characteristics may further increase the complexity of these calculations. In practice, transparency and plausibility of assumptions are often more important than methodological complexity when conducting sample size calculations.²⁰⁴ Therefore, **when key parameters are uncertain, a simple sample size calculation based on a limited number of transparent and well-justified assumptions may be preferable to**



a highly complex calculation that relies on many uncertain inputs. Additional complexity may not improve the reliability of the estimated sample size and can create a false sense of precision.

4.2.2 What do we need for power estimation?

4.2.2.1 Significant level (α) and power ($1-\beta$)

Power and sample size calculations require the specification of a significance level (α) and a desired statistical power ($1-\beta$). **Lower significance levels and higher levels of power require a larger sample size.**

In many studies, a significance level of 5% and a power of 80% or 90% are used. This corresponds to accepting a 5% probability of falsely concluding that an intervention is effective and a 20% or 10% probability of failing to detect a true effect. However, the consequences of false positive and false negative conclusions may differ between studies and should be considered during study planning. For example, stronger evidence (i.e. lower α) may be required before implementing a costly health policy intervention, whereas a more exploratory evaluation may tolerate a larger probability of type I errors.²⁰⁵

4.2.2.2 Effect size (δ)

The effect size represents the magnitude of the association or difference that the study aims to detect. Its exact interpretation depends on the study design and statistical analysis. For example, when comparing two groups, the effect size may be defined as the difference in average outcome values between the groups. In studies evaluating a dose-response relationship, the effect size may correspond to the slope describing the change in outcome associated with a one-unit increase in exposure or intervention intensity. More generally, the effect size is the parameter for which evidence is sought and on which the sample size calculation is based.

The larger the true effect, the easier it is to detect statistically, and therefore the smaller the sample size required to achieve a given level of power. This reflects the signal-to-noise ratio in the data: larger effects stand out more clearly against natural variability, whereas smaller effects are more easily obscured by random variation and typically require larger sample sizes to detect. A central task in sample size calculation is obtaining a plausible estimate of the effect size. Such estimates can be informed by existing data, expert knowledge, published literature, or evidence from studies conducted in similar settings or populations. For example, researchers may draw on effect estimates reported in previous studies, routinely collected healthcare data, or consultations with clinical experts and other stakeholders. Because these sources are often uncertain or imperfect, it is generally advisable to take a conservative approach when specifying the effect size. In practice, this means assuming a somewhat smaller effect, as larger assumed effects lead to smaller required sample sizes, whereas smaller assumed effects increase the sample size and reduce the risk of an underpowered study.⁸⁵

An alternative starting point is to consider the smallest effect that would be considered meaningful. This is sometimes referred to as the **minimum (clinically) important difference** (M(C)ID). Sample size calculations can then be based on the effect size that the study should be able to detect in order to inform decision-making. This is often a relatively conservative approach, as meaningful effects may be smaller than the effects reported in previous studies, resulting in larger required sample sizes. Published M(C)IDs are available for a range of outcomes, particularly for patient-reported outcome measures (PROMs). These are typically derived in dedicated methodological studies that aim to identify the smallest change in an outcome that is perceived as meaningful by patients, for example by relating changes in the outcome measure to patients' own assessment of improvement.^{206, 207}

Both perspectives are useful and ideally should be considered together. An effect may be plausible but too small to justify implementing an intervention, or conversely, meaningful but unrealistic to achieve. Discussing both the expected effect and the minimum meaningful effect can help determine whether a study is both feasible and worthwhile.

4.2.2.3 Variability (σ)

Higher variability reduces the precision of the estimate and therefore increases the sample size required to achieve a given level of statistical power.

Estimates of the standard deviation are sometimes available from previous studies, pilot data, or the published literature. However, relevant data or literature may not always exist. In such cases, expert opinion can be used to inform plausible values, or estimates may be sought from data obtained in similar settings or populations. In general, it is advisable to adopt a conservative approach. This involves assuming a slightly larger standard deviation, which increases the required sample size but reduces the risk of an underpowered study.⁸⁵

When uncertainty about the standard deviation remains high, researchers may consider incorporating a planned reassessment of this assumption after the first participants have been enrolled (also called a sample size assumptions reassessment interim analysis). In many settings, this can be done without examining differences between study groups, for example by estimating the pooled standard deviation across all participants, thus maintaining blinding.²⁰⁸

4.2.2.4 Allocation ratio

Studies do not always assign equal numbers of participants to each intervention arm. Unequal allocation may be chosen for ethical, practical, recruitment, or cost reasons (see Chapter 4). However, **allocating more participants to one arm than the other reduces the statistical efficiency of the comparison, meaning that a larger total sample size is required to achieve the same level of confidence in the results.** The extent of this increase depends on the allocation ratio, and the assumed variance in both arms. In many examples these are assumed to be equal, but this is not necessarily the case. The reason for this increase is that unequal allocation inflates the variance of the estimated difference between intervention arms, thereby reducing statistical efficiency.^{203, 209}

As a practical rule of thumb, to achieve the same statistical power as a study with equal allocation (1:1), a study with 1:2 allocation ratio requires 12% more participants in total, while a 1:3 allocation ratio requires 33% more participants.^h

4.2.2.5 Outcome-specific parameters required for sample size calculation

In addition to the aforementioned commonly needed parameters (α , $1-\beta$, δ , σ), sample size calculations require outcome-specific parameters. The exact information needed depends on the type of outcome and the corresponding statistical analysis. Table 26 provides an overview of the parameters that typically need to be specified for commonly encountered outcome types. As with the target effect size, these parameters are often obtained from previous studies, pilot data, routine data sources, or expert judgement.^{203, 209}

^h Under otherwise equal circumstances, the required total sample size is proportional to $(1+r)^2/(4r)$, where r denotes the allocation ratio. The percentages reported above are derived directly from this relationship.



Table 26 – Additional required parameters for sample size calculation by outcome type

Outcome type	Main parameters needed												
Continuous outcome data: Difference between two normally distributed means	δ = expected effect size under H_a σ = standard deviation in all experimental arms												
Binary outcome data: Difference between two proportions	π_c = proportion in control arm π_e = proportion in experimental arm (based on the expected effect size)												
Time to event data	- HR + (median time to event or landmark proportion in one arm), or - Median time to event or landmark proportions in both arms - Operational assumptions in order to calculate number of participants - Accrual rate and pattern - Follow-up time after accrual - Dropout rate - Preferred timing of analysis												
Count data	λ_C and λ_E: expected number of events per time period D_E, D_C : study duration in each arm												
Ordinal, discrete score-type data and nominal data	Expected proportions for each category, in each arm, e.g. <table><tr><th>Category</th><th>Control</th><th>Experimenta l</th></tr><tr><td>Poor</td><td>12%</td><td>8%</td></tr><tr><td>Fair</td><td>25%</td><td>20%</td></tr><tr><td>Good</td><td>30%</td><td>32%</td></tr></table>	Category	Control	Experimenta l	Poor	12%	8%	Fair	25%	20%	Good	30%	32%
Category	Control	Experimenta l											
Poor	12%	8%											
Fair	25%	20%											
Good	30%	32%											

Sources: Based on Cytel et al. (2018)²⁰⁹ and Chow et al. (2017)²⁰³

4.2.3 Factors affecting sample size calculation

4.2.3.1 Multiple primary outcomes

In studies of complex interventions, it is common to focus on **multiple key endpoints** rather than a single primary endpoint. Multiple primary outcomes have implications for sample size calculations for two reasons. First, different outcomes may require different sample sizes because they differ in their type, frequency, variability, or expected intervention effect. **The overall sample size is therefore typically determined by the outcome with the most demanding power requirement.** Second, the required sample size depends on how intervention success is defined. For example, if all primary outcomes must demonstrate a positive effect for the intervention to be considered successful ('co-primary endpoints'), a larger sample size may be required to ensure sufficient overall power. As a result, studies with multiple primary outcomes often require more participants than studies with a single primary outcome^{i,70}

4.2.3.2 Clustered designs

In reality, not all data are independent of each other as is often assumed in standard sample size calculations. This may occur when participants are grouped within the same hospital, general practice, or family (clustered data), or when the same individuals are measured repeatedly over time

ⁱ When multiple co-primary endpoints must all show a significant effect, the overall power equals the probability that all endpoints achieve significance simultaneously. Assuming independence, this can be approximated as the product of the individual powers. For example, two endpoints each powered at 80% yield an overall power of 64% (0.8×0.8), whereas two endpoints each powered at 90% yield an overall power of 81% (0.9×0.9). Because co-primary endpoints are often positively correlated, this approximation is generally conservative.

(longitudinal data). Correlation between data affects the amount of information available and therefore has implications for both sample size calculations and data analysis (more on this in section 4.4).²¹⁰

Observations within the same cluster tend to be more similar to one another than observations from different clusters. As a result, the effective sample size is smaller than the actual number of participants. For between-group comparisons, **the stronger the within-cluster correlation, the larger the sample size required to achieve a given level of statistical power.** These considerations are discussed further in section 4.4 and Appendix 2.3. In contrast, for within-group comparisons based on longitudinal data, observations from the same individual are often positively correlated. Thus, **for within-group comparisons, stronger correlations reduce the variability of changes over time, which increases statistical power and may reduce the sample size required to detect a given effect.**²¹⁰

Ignoring these correlations can lead to an overly optimistic sample size calculation and an underpowered study. Correlation also has important implications for the statistical analysis of the data. The main considerations for the analysis of clustered data are discussed in section 4.4, while Appendix 2.2 and Appendix 2.3 provide additional technical details on the sample size calculations in case of clustered data.^{107, 108}

4.2.3.3 *Drop-out and missing data*

Standard sample size calculations often assume that outcome data will be available for all participants. In reality, some participants may withdraw, be lost to follow-up, or have incomplete outcome measurements. Because these participants may contribute less information to the primary analysis, the effective sample size is often smaller than the number of participants initially recruited. Ignoring this issue can result in a study with insufficient power.^{203, 211}

Studies therefore commonly increase the planned sample size to account for anticipated drop-out or missing data. When the primary analysis is based only on participants with observed outcome data, the required inflation is often determined directly by the expected loss of participants. For example, if 100 evaluable participants are required and a 20% drop-out rate is anticipated, approximately 125 participants would need to be recruited.

When the primary analysis aims to estimate the treatment effect for all enrolled participants despite missing outcome data (i.e. an intention-to-treat analysis as will be explained in section 4.5), the impact of drop-out on sample size is more complex. The required adjustment depends not only on the proportion of missing data but also on how missing outcomes and other post-enrolment events are handled in the analysis (i.e. on the strategy chosen to handle intercurrent events, as will be explained in section 4.5). Consequently, there is no single inflation factor that can be recommended for all studies. Further details are provided in Appendix 2.4.

4.2.3.4 *Change from baseline*

Many intervention studies collect measurements before and after the intervention. Incorporating baseline measurements into the analysis can reduce variability, thereby improving the efficiency of the study and reduce the required sample size.²¹²

The magnitude of this gain depends on how strongly baseline and follow-up measurements are related. Because this relationship is often uncertain during study planning, sample size calculations that make use of baseline measurements require additional assumptions. Further details on methods such as change-from-baseline analyses and baseline-adjusted analyses are provided in Appendix 2.5.²¹²



Key considerations for the research team:

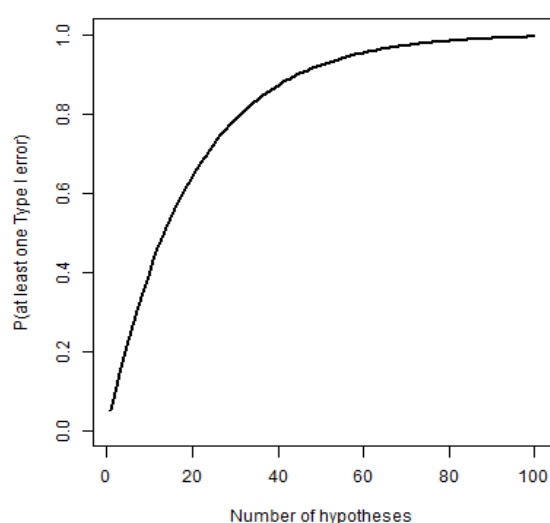
- Perform a sample size calculation to estimate how many participants are required to answer the research question with sufficient power.
- Recognise the main factors that may increase the required sample size:
 - Using a more stringent significance threshold
 - Requiring a higher probability of detecting a true effect
 - Trying to detect smaller effects
 - Greater variability (noise) in the data
 - Unequal allocation of participants between groups
 - For between-group comparisons: Stronger clustering of participants
 - For within-group comparisons: Weaker clustering among observations of the same participant
 - Higher expected levels of drop-out or missing data
- Collect the best available evidence to inform assumptions regarding effect size, variability and other required parameters (e.g. from the literature, routine data, pilot studies or expert input).
- Discuss both the expected effect and the smallest effect that would be considered meaningful for decision-making.
- Keep assumptions transparent and proportionate. A simple calculation based on a limited number of well-justified assumptions is often preferable to a highly complex calculation that relies on many uncertain inputs.
- Treat a sample size calculation as an estimate rather than a hard cut-off, as it depends on assumptions that may ultimately prove incorrect.

4.3 Managing type I error: multiplicity

The type I error may be compromised when a study includes multiple statistical comparisons or analyses, a phenomenon referred to as multiplicity. As a consequence, the overall probability of making at least one type I error may exceed the intended significance level (e.g. 5%) unless appropriate adjustments are made. Figure 27 illustrates how this probability increases rapidly as the number of statistical tests grows^j. Consequently, any planned analyses or interim evaluations should be justified by a clear scientific objective.²¹³

^j More specifically, the probability of at least one type I error increases according to the formula $1-(1-\alpha)^m$, with m being the number of tests.

Figure 27 – Probability of at least one type I error increasing with the number of conducted tests



Number of tests	Probability of at least one Type I error
1	0.05 = 5%
2	0.098 ≈ 10%
3	0.143 ≈ 14%
4	0.185 ≈ 19%
5	0.226 ≈ 22%
10	0.401 ≈ 40%
15	0.537 ≈ 54%
20	0.642 ≈ 64%

4.3.1 Multiple testing

Multiple testing occurs whenever a study includes more than one formal hypothesis test. Common examples include studies with multiple outcomes, multiple treatment comparisons, multiple subgroup analyses or adaptive interim analyses. Because each additional test creates another opportunity to demonstrate an effect, the probability of obtaining at least one false-positive finding increases unless appropriate adjustments are made.²¹³

Numerous methods have been developed to control the overall risk of a Type I error when multiple hypotheses are evaluated. From a study design perspective, the most important distinction is whether hypotheses are treated as equally important or whether some are prioritised over others.

If all hypotheses are considered equally important, the available significance level can be distributed across them. The well-known **Bonferroni correction** (illustrated in panel A of Figure 28) is one example, although less conservative alternatives are often used in practice (e.g. Tukey, Hochberg, Holm, ...).²¹⁴

A common approach in studies where hypotheses are not considered equally important is **hierarchical testing** (panel B of Figure 28), in which hypotheses are tested in a predefined order. Testing proceeds to the next hypothesis only if the previous test is statistically significant. Another approach is the **split alpha** strategy (panel C of Figure 28), in which the overall significance level is distributed across hypotheses according to their relative importance. Greater priority can be given to key outcomes by assigning them a larger share of the available Type I error budget.²¹⁴

Hierarchical testing and split alpha can also be combined in what is called a **fallback procedure**. In this approach, hypotheses are prioritised, while part of the Type I error budget is initially reserved for lower-priority hypotheses. If a higher-priority hypothesis is statistically significant, its allocated significance level can be passed on, allowing subsequent hypotheses to be tested at a more lenient threshold. For example, in panel D of Figure 3, the primary outcome (H1) is tested at $\alpha = 0.03$ and the secondary outcome (H2) at $\alpha = 0.02$. If H1 is significant, its α -allocation is transferred to H2, which can then be tested at the full significance level of $\alpha = 0.05$. If H1 is not significant, H2 remains tested at $\alpha = 0.02$ only.²¹⁴

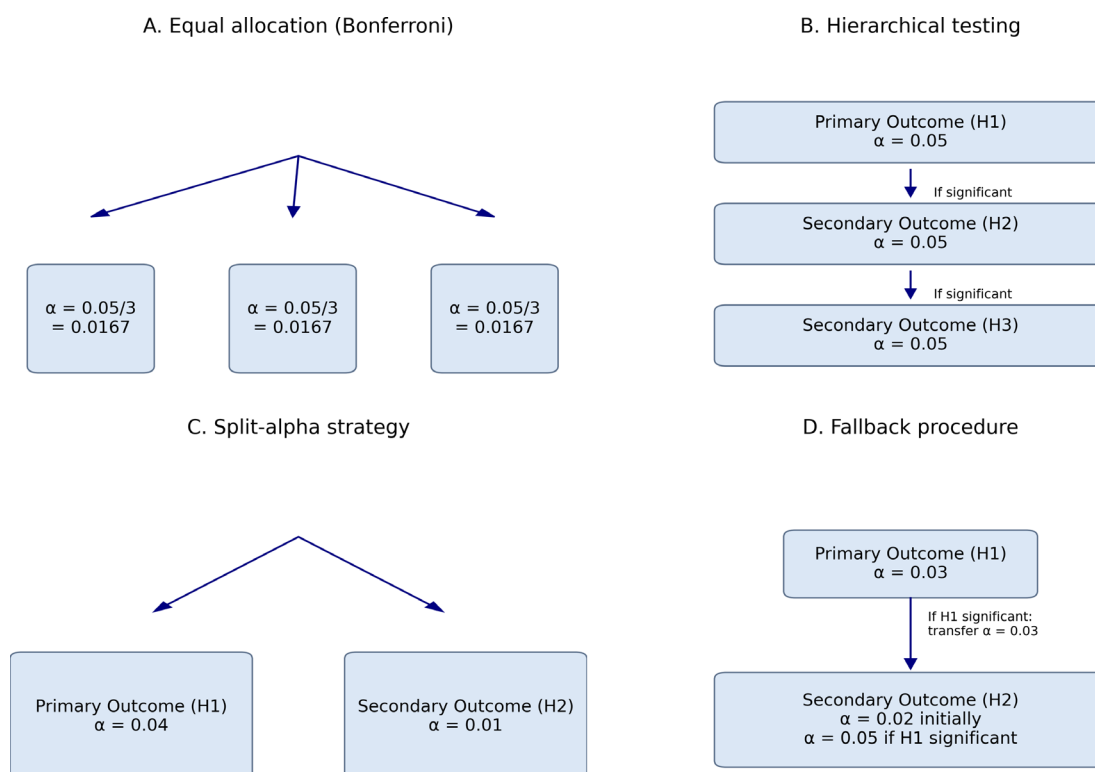


Regardless of the specific method, multiplicity adjustment often requires explicit decisions about which outcomes or hypotheses are most important. These prioritisation decisions should be made by the research team when creating the protocol, before the start of the study.

Finally, it is important to note that there are also contexts in which formal adjustment for multiple testing is not required. Whether adjustment is needed depends on whether a statistically significant result could contribute to concluding the intervention is effective. Examples where formal adjustment may not be required include:

- **Analyses performed for descriptive purposes**, where the objective is to describe observations rather than to support a formal claim. Descriptive results should, by definition, not be subjected to formal hypothesis testing.
- **Sensitivity analyses**, which are intended to assess the robustness of the main findings rather than to establish the effectiveness of the intervention. Multiple outcomes combined into a composite outcome, resulting in a single formal hypothesis test. (Note: more information on sensitivity analyses can be found in section 4.6.)
- **Co-primary outcomes** for which all outcomes must be statistically significant for the study to be considered successful. In this case, the overall Type I error rate is already controlled, although statistical power is reduced because all outcomes must be significant (see Section 2.2.4.2).²¹⁵
- **When success is defined by a coherent pattern of effects across these outcomes rather than by statistical significance of any single outcome.** Consequently, the interpretation of the study is unlikely to change because one additional outcome happens to reach statistical significance. Instead, the pattern of findings across outcomes is used to assess whether the intervention operated as expected. This is typically the case in studies of complex interventions, where a theory of change often specifies a limited set of key outcomes through which the intervention is expected to operate. Conventional multiplicity adjustments may therefore be less relevant.⁷⁰

Figure 28 – Illustration of four different adjustment methods for multiple testing ($\alpha = 0.05$)



4.3.2 Interim analysis

Interim analysis refers to an evaluation conducted while a study is ongoing, before completion of data collection. Common applications of interim analyses include:

- **Interim monitoring**, which supports study oversight by assessing aspects such as participant recruitment, data quality, adherence to study procedures, or participant safety. More information on how this is ideally conducted can be found in Textbox 10.

Example: Monitoring participant recruitment may reveal that enrolment is substantially slower than anticipated, prompting modification to the recruitment strategy (such as extending the recruitment period or enrolling additional study sites) or reconsideration of the study timeline.

- **Reassessment of assumptions underlying the study design**, which allows researchers to verify whether assumptions used during study planning remain reasonable.²⁰⁸

Example: Estimates of the variability of the primary outcome may be used to reassess the sample size required to achieve the desired statistical power.

- **Adaptive study designs**, in which pre-specified modifications to the study are made based on accumulating data. Examples include stopping a study early because the intervention appears clearly beneficial, harmful or unlikely to achieve its objectives (futility), reassessing the sample size, discontinuing ineffective intervention arms, or modifying eligibility criteria. Such approaches may improve study efficiency and help avoid unnecessary use of resources, but they require careful planning and specialised statistical methods to preserve the validity of study conclusions. Adaptive designs are most useful when important uncertainties remain at the design stage and when information accumulated during the study may meaningfully inform subsequent study decisions. Further details are provided in Appendix 2.6.²¹⁶

Example: A study protocol may specify that an intervention arm will be discontinued if interim results suggest that it is highly unlikely to achieve the desired effect.

From the perspective of multiplicity, the key distinction is whether interim analyses involve evaluating study outcomes. Interim monitoring activities and reassessment of design assumptions generally do not raise concerns regarding multiplicity because they do not involve repeated evaluation of treatment effects. In contrast, adaptive designs often rely on interim assessment of study outcomes and therefore create additional opportunities to demonstrate an effect. Unless appropriate adjustments are made, such repeated evaluations may increase the probability of a Type I error.²¹⁷

Textbox 10 – Data monitoring committees

Monitoring of study conduct is generally considered part of good study oversight and should be considered in most intervention studies. Interim findings may occasionally lead to recommendations to modify or terminate a study, for example because of clear evidence of benefit, futility, poor data quality, slow participant enrolment, or because the research question has already been answered elsewhere.

When interim results may influence important study decisions, access to such information should preferably be restricted to an independent multidisciplinary group with sufficient knowledge of the intervention and study context. This helps to reduce the risk that knowledge of interim results influences study conduct, data collection, or interpretation of the findings. The need for formal oversight structures and procedures for blinding interim results should be assessed on a case-by-case basis, taking into account the study design, potential risks to participants, and the broader study context.²¹⁸

This approach originates from clinical drug trials, where interim results are typically reviewed by an Independent Data Monitoring Committee (IDMC), but may also be relevant in other intervention studies.^{219, 220} An IDMC typically is composed of voting members (a number of expert physicians



and a biostatistician) and a nonvoting independent statistician, involved in creating the interim statistical report and who presents these results to the voting members.²²¹ For the KCE evaluation programme, the need for a data monitoring committee, might be discussed at the protocol meeting. As these formal structures do not exist, they can be set up on an ad-hoc basis, following the same principles.

Key considerations for the research team:

- Identify potential sources of multiplicity, such as multiple outcomes, multiple treatment comparisons, subgroup analyses or interim analyses.
- Consider whether all planned analyses are necessary and scientifically justified.
- Decide whether some hypotheses or outcomes are more important than others, as this may influence the choice of multiplicity adjustment strategy.
- Consider whether interim analyses will be performed and whether they create additional opportunities to conclude that the intervention is effective.
- Assess whether independent oversight of interim results is warranted, for example through an independent multidisciplinary monitoring group.

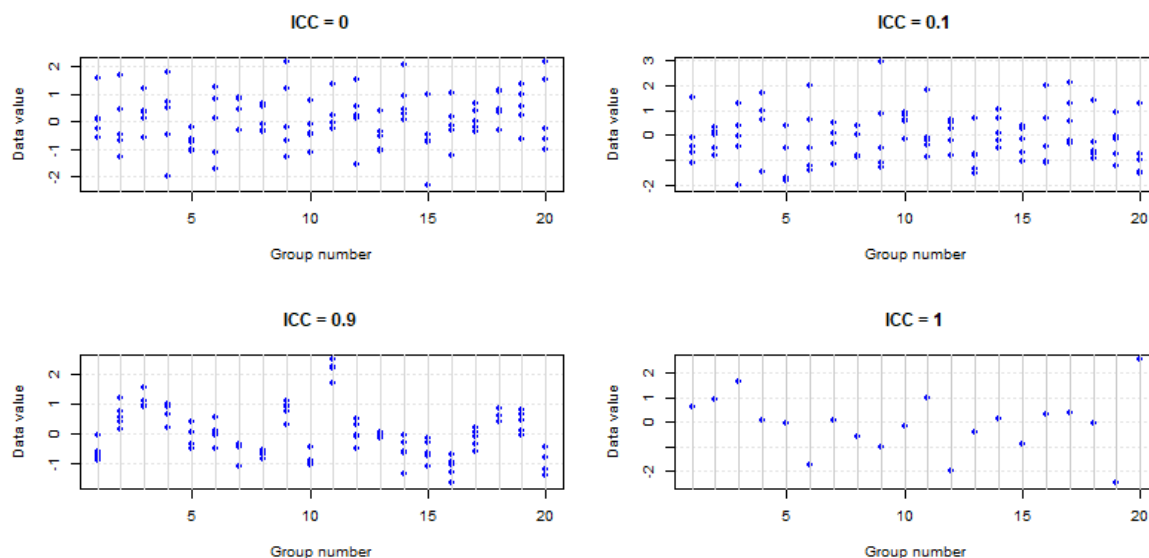
4.4 Clustered and longitudinal data

Many statistical methods assume that observations are independent of one another. In practice, however, observations often occur in groups and may therefore be related. Such data are referred to as clustered data.

A common source of correlation is **clustering**, where individuals share a common context. Examples include patients treated in the same medical practice. Because these individuals are exposed to similar care processes, healthcare professionals, and organisational environments, their outcomes are often more alike than those of individuals from different settings.²¹⁰

However, the mere existence of clusters does not necessarily imply substantial correlation. The degree of similarity between observations within the same cluster may vary considerably depending on the context and the outcome being studied. The strength of the correlation between observations within the same cluster is often quantified using the intraclass correlation coefficient (ICC). An ICC close to zero indicates little or no within-cluster similarity, whereas larger ICC values indicate stronger correlation between observations from the same cluster (illustrated in Figure 29). More information on how to estimate the ICC is presented in Textbox 11, while further details on the ICC are provided in Appendix 2.2.

Figure 29 – Intraclass correlation coefficient (ICC)



Textbox 11 – Estimating the intraclass correlation coefficient

The intraclass correlation coefficient (ICC) is typically unknown at the planning stage. It must therefore be estimated based on:

- previous studies,
- published literature,
- expert opinion,
- or data from similar settings.

Typical ranges (for continuous outcomes) depend on the setting. Literature is available in varying contexts. For example, the following ICC values were reported according to Crespi 2025:¹⁰⁸

- Countries or cities: very small (<0.01-0.001)
- Schools or worksites: 0.01-0.05
- Primary care studies: up to 0.1
- Families: >0.2

Because of this uncertainty, it is good practice to use conservative values (larger) and to perform sensitivity analyses over a plausible range of ICC values. If feasible interim re-estimation of the ICC and sample size during the trial can be considered, ideally blinded to treatment allocation, in order to prevent operational bias.¹⁰⁴

Correlation also arises in **longitudinal data**, where data are collected from the same individual repeatedly over time. For example, a participant may complete a questionnaire before an intervention and again during follow-up. Responses from the same individual are typically more similar than responses from different individuals. In this sense, longitudinal data can be viewed as a special form of clustered data, where repeated measurements are grouped within individuals.²¹⁰

For longitudinal data, it is important, during the planning phase of a study, to carefully consider the measurement schedule. Designing the study with common, predefined timepoints with respect to every individual's baseline, for all participants facilitates the use of simpler statistical models, allowing



pre-specification in the protocol. It also allows easier effect estimation at earlier timepoints. In practice, measurement schedules are often defined using time windows to allow for some flexibility. For example, a follow-up visit may be scheduled monthly, with an allowable window of ± 7 days.

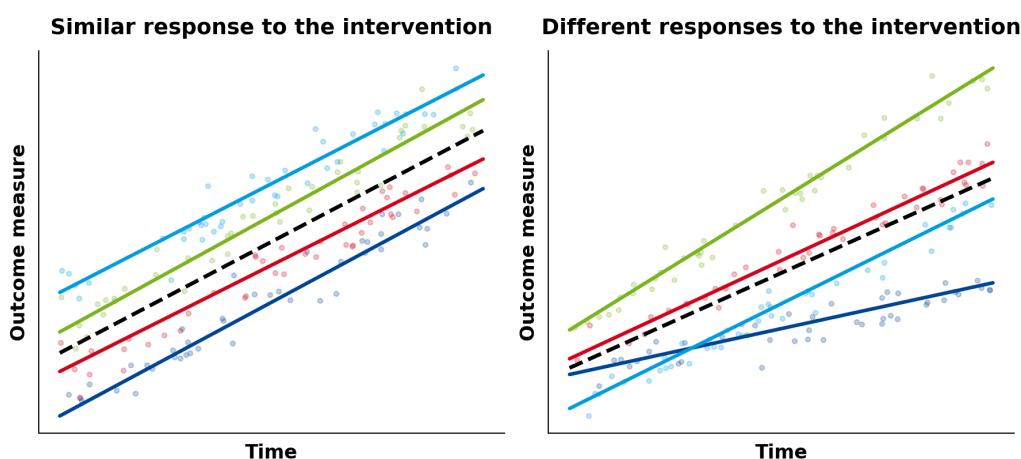
In many intervention evaluations, several forms of clustering occur simultaneously. Patients treated by the same physician may have more similar satisfaction with care because they are exposed to similar care processes. Additional correlation may arise when physicians work within the same hospital and share a common organisational environment. Repeated measurements from the same patient are also correlated. For example, previous negative experiences with healthcare may lead to consistently lower satisfaction scores over time, regardless of the intervention.

Correlated observations are not completely independent from one another. As a result, an additional observation from the same cluster generally contributes less new information than an additional observation from a different cluster. The amount of information in a dataset therefore depends not only on the number of observations, but also on how strongly those observations are related to one another. This is important because most standard statistical inference techniques assume that observations are independent. When correlation is present, this assumption is violated. **If the correlation is ignored, the effective amount of independent information in the data may be overestimated, uncertainty may be underestimated, and confidence intervals may become too narrow.**²¹⁰

For this reason, it is important to identify potential sources of clustering and, in longitudinal studies, to **ensure that measurements from the same individual can be linked across waves**. These design decisions determine which analyses are possible and how reliably intervention effects can be estimated.

The statistical methods used to account for correlation in the data (e.g. including mixed-effects models and generalized estimating equations models) are beyond the scope of this process note. Nevertheless, the structure of the data is not merely a statistical consideration. Subject-matter expertise is often needed to identify potential sources of correlation and to anticipate how groups or individuals may differ in their response to the intervention (see Figure 30).

Figure 30 – Illustration of cluster response to intervention



In the left panel, groups differ primarily in their baseline level but show a similar evolution over time. In the right panel, groups differ both in their baseline level and in their response to the intervention. The dashed line indicates the overall mean, while the full line indicates different clusters.

Key considerations for the research team:

- Identify potential sources of correlation in the data, such as clustering within providers, organisations, families or repeated measurements within individuals.
- In case of longitudinal data, carefully consider the measurement schedule.
- Consider whether individuals within the identified clusters are likely to be more similar to one another than to individuals outside the cluster with respect to characteristics or experiences that may influence the outcome.
- Ensure that the necessary identifiers are collected/requested to allow clustered and longitudinal analyses.
- Consider how clusters may differ, both at baseline and in their response to the intervention.
- Recognise that clustering and repeated measurements affect both sample size requirements and analytical methods.
- Consult statistical expertise early when clustered or longitudinal designs are anticipated.

4.5 Defining the treatment effect of interest

4.5.1 *Intention-to-treat and per-protocol analyses*

Intervention studies do not always proceed exactly as planned. Participants may discontinue the intervention or otherwise deviate from the intended protocol. As a result, different analytical approaches may answer different questions about the effect of an intervention. Two commonly used approaches (both illustrated in with an example in Textbox 12 in the next section) are:^{222, 223}

- **Per-protocol (PP):** This approach restricts the analysis to participants who adhered sufficiently to the intervention as originally planned. The PP approach therefore aims to estimate the effect of the intervention under ideal or near-ideal implementation conditions.
- **Intention-to-treat (ITT):** Participants are analysed according to the intervention group to which they were originally assigned, regardless of whether they fully adhered to the intervention. The ITT approach therefore estimates the effect of *offering* or *implementing* the intervention in practice, taking into account that some participants may not fully comply.

A third approach is the **as-treated (AT)** analysis, in which participants are analysed according to the intervention actually received rather than the intervention originally assigned. AT analyses may be informative when actual exposure differs substantially from intended exposure, for example when uptake of an intervention is variable, implementation fidelity differs across organisations, or participants cross over between study groups. However, because participants or organisations receiving different levels of exposure may differ systematically from one another, AT analyses are generally more susceptible to bias and are therefore usually considered supplementary or sensitivity analyses rather than primary analyses.^{224, 225}

In most intervention studies, the ITT approach is preferred as the primary analysis because it preserves the comparability of the study groups established at the start of the study, guaranteed by randomisation in case of an RCT. Participants who adhere to an intervention often differ systematically from those who do not. For example, more motivated participants may be more likely to remain in a programme and may also achieve better outcomes, irrespective of the intervention itself. Excluding participants after such deviations can therefore introduce bias and lead to an overly optimistic estimate of the intervention effect.²²²

Nevertheless, PP analyses may still be informative in some situations, particularly when there is interest in understanding the effect of an intervention under conditions of good adherence. For this



reason, some studies report both ITT and PP analyses, while clearly identifying which analysis addresses the primary research question.²²²

The choice between ITT and PP should be driven by the research question and specified before the start of the study. To make this decision, it is necessary to consider how events occurring after implementation of the intervention will be handled. These events are discussed in the next section.

4.5.2 Intercurrent events

Intercurrent events^k are events occurring after randomisation and implementation of an intervention that may affect either exposure to the intervention or the interpretation of the outcome of interest. The importance of an intercurrent event depends on the research question being addressed. Events that are irrelevant for one treatment effect may be highly relevant for another.²²⁶

Examples of intercurrent events in health services and policy evaluations include:

- A participant, healthcare provider or organisation withdraws from the intervention before the end of the study.
- A participant receives an additional intervention that may affect the outcome of interest.
- A reimbursement or policy change alters exposure to the intervention for part of the study population during the study period.

Textbox 12 – An illustrative example on per-protocol analysis, intention-to-treat analysis and intercurrent events

Consider a pilot programme introducing a new financing model for home nursing. Some nurses may withdraw because participation negatively affects their income. Such withdrawal is an example of an **intercurrent event (ICE)**.

Nurses who leave the programme are unlikely to be a random subset of participants. For example, self-employed nurses may be more financially affected by the new financing model than salaried nurses and may therefore be more likely to withdraw. As a result, the nurses who remain in the programme may **systematically differ from the original study population**.

Excluding nurses who discontinue participation in a **per-protocol analysis** may therefore lead to biased estimates of the intervention effect. Similar concerns apply to an **as-treated analysis**, where nurses who discontinue participation may be reclassified according to their actual exposure status, potentially increasing differences between the comparison groups. An **intention-to-treat analysis** would retain these nurses in the comparison, thereby preserving the original study groups and reducing the risk of such bias.

Nevertheless, per-protocol and as-treated analyses may provide useful complementary information alongside an intention-to-treat analysis. For example, if an intervention appears ineffective in an ITT analysis, these analyses may help distinguish between a lack of effectiveness of the intervention itself and challenges related to uptake, adherence, or implementation. Their results should be interpreted with caution because both approaches may be susceptible to selection bias.

^k This terminology was introduced in the context of the estimand framework.²²⁶

Intercurrent events should not be confused with missing data. An intercurrent event may occur even when outcome data continue to be collected. For example, a nurse may withdraw from a pilot financing programme but continue to provide follow-up data. Conversely, missing data may occur without a clear intercurrent event, for example when a participant cannot be contacted for follow-up. Whenever feasible, outcome data should therefore continue to be collected even after an intercurrent event has occurred.

The way intercurrent events are handled determines the treatment effect being estimated. Several strategies are possible.²²³

- Under a **composite strategy**, the intercurrent event is incorporated into the outcome definition. For example, withdrawal from a programme may itself be considered an unfavourable outcome.
- Under a **while-on-treatment strategy**, only outcomes observed before the intercurrent event are considered. This strategy may be relevant when the effect of the intervention is only of interest while participants remain exposed to it.
- Under a **treatment policy strategy**, outcomes are analysed regardless of whether an intercurrent event occurs. This approach is conceptually closest to an intention-to-treat analysis and is often appropriate when the objective is to estimate the effect of implementing an intervention in routine practice.
- Under a **hypothetical strategy**, the objective is to estimate what would have happened if the intercurrent event had not occurred. In this case, post-ICE data are ignored, and the unobserved hypothetical outcome must be estimated or imputed using model-based or causal inference methods.
- The **principal stratum strategy** focuses on the subgroup of patients who would not experience the intercurrent event under either treatment. Since this subgroup cannot be directly observed, causal inference methods are again required.

The most appropriate strategy depends on the research question rather than on statistical considerations alone. For this reason, important intercurrent events and the planned approach for handling them should be identified during study design and specified in the study protocol before the analysis is conducted. These choices in turn guide data collection, data labelling and planning of the statistical analysis.

4.6 Evaluating the robustness of conclusions

All statistical analyses rely on assumptions and analytical choices. Examples include how missing data are handled, which statistical model is used, how outliers are treated, or which covariates are included in the analysis. Because these choices may influence the results, it is often useful to perform sensitivity analyses.

Sensitivity analyses assess whether the main study conclusions remain unchanged under alternative, but still reasonable, assumptions or analytical approaches. Consistent findings across sensitivity analyses increase confidence that the conclusions are not driven by a particular analytical choice.

Common examples of sensitivity analyses include:

- Comparing different assumptions used in the analysis
- Using alternative definitions or operationalisations of an outcome
- Using an alternative definition of the effect (i.e. intention-to-treat vs. per-protocol)
- Comparing alternative statistical models (e.g. parametric versus non-parametric methods)
- Including or excluding specific covariates in the statistical model



- Using different approaches to handle missing data
- Evaluating the influence of outliers

The choice of sensitivity and supplementary analyses should be specified whenever possible before data analysis begins. It is most relevant to prioritise sensitivity analyses that challenge conclusions likely to influence decision-making.

Key considerations for the research team:

- Identify the most important assumptions underlying the primary analysis, such as assumptions regarding missing data, model choice, or outcome definitions.
- Plan a limited number of relevant sensitivity analyses that assess whether the main conclusions remain robust under alternative reasonable assumptions.
- Pre-specify sensitivity analyses whenever possible, rather than designing them after seeing the results.

4.7 Guidance with respect to protocol, statistical analysis plan and reporting

It is advisable to consider relevant guidance from the start of the study planning. The CONSORT-SPIRIT website (<https://www.consort-spirit.org/>) provides recommendations for reporting both RCT protocols (SPIRIT) and results (CONSORT). In addition, the website <https://www.consort-spirit.org/extensions> offers a range of extensions tailored to specific study designs, including crossover trials, factorial trials, non-inferiority/equivalence trials, pragmatic trials, cluster randomised trials, cluster crossover trials and stepped-wedge trials.

Although the estimand framework (see above) is recognised as an important concept in trial design and analysis by CONSORT and SPIRIT authors, it has not yet been formally incorporated in the central SPIRIT and CONSORT checklists. Yet, they have been included in the SPIRIT and CONSORT extensions for factorial RCTs.²²⁷

Incomplete and inconsistent reporting of study interventions limits their implementation and reproducibility. The TIDieR checklist (<https://tidierguide.org/>) was developed to address this issue by providing guidance to ensure interventions are described in sufficient detail for authors, reviewers, and readers to understand and replicate them.

The EQUATOR Network (<https://www.equator-network.org/>) is an international initiative that promotes transparent and high-quality reporting of health research. It provides a comprehensive collection of tools and guidelines to support the design, conduct and dissemination of research studies.

Finally, the KCE web page (<https://kce.fgov.be/en/kce-trials/resources/useful-links-resources-for-investigators>) includes a lot of references useful in the planning stage of a study.

5 PROCESS EVALUATION

KEY POINTS

- Process evaluations complement and deepen evidence from effectiveness evaluations. They **help to understand how, why, for whom, and under what circumstances** the intervention works or does not work.
- Process evaluations help to open the 'black box' between activities related to the intervention and outcomes. It seeks to provide information on implementation, mechanisms, and context in complex interventions, supporting accurate interpretation of outcomes and their transferability.
- Process evaluations are conceptualised around three core dimensions and their interrelations: **implementation** assesses whether the intervention was implemented as planned and how this was achieved; **mechanisms of impact** seek to clarify the underlying causal pathways; and **context** involves identifying the factors that may have influenced what happened with the intervention.
- **Theory, model and framework (TMF)**
 - The use of a TMF is essential in a process evaluation, as it provides an analytical structure. It helps to guide data collection and analysis, and strengthens methodological rigour.
 - The selection of a TMF should be aligned with the research question(s) and the specific aspect(s) of the intervention under investigation, such as implementation, mechanisms of impact and/or context.
 - The TMFs may be applied in full or in part, but the reasons behind each decision must always be thoroughly discussed and documented.

5.1 Introduction

In general, most topics that will be evaluated in the context of the KCE evaluation programme will concern complex interventions for which traditional approaches to evidence production (largely grounded in linear cause-effect models) are insufficient.¹⁵ **Process evaluations** expand the scope of the evaluation as it goes beyond the question 'Does this intervention work?' Instead, it examines how an intervention is implemented, how it generates its effects, and how these processes are shaped by the context in which it is delivered and therefore answering why an intervention works or fails to work. It complements and deepens effectiveness evaluation by addressing questions that move beyond whether an intervention works. The process evaluations contribute to the understanding of 'how, why, for whom, and under what circumstances' outcomes are achieved. Accordingly, process evaluations shift the interpretation of findings from a conclusion such as 'the intervention has no effect' to a more nuanced understanding: 'under these conditions, and when implemented in this way, specific mechanisms were (or were not) activated'.^{9, 15, 25, 228, 229}

5.2 What is process evaluation?

5.2.1 Process evaluation: beyond effectiveness evaluation

Process evaluations enable a more nuanced **understanding** of intervention outcomes. Effectiveness evaluation indicates whether an intervention works, but does not necessarily explain **why**.²³⁰ In particular, when an effectiveness evaluation shows no effect, process evaluations can help to distinguish between **theory failure** (also called intervention failure) and **implementation failure**.²³⁰ As such, it will avoid to incorrectly interpret a null result as evidence of ineffectiveness when it actually might work under some conditions, for some target groups, etc.



Theory failure implies that the theoretical basis of the intervention is flawed: even if the intervention is delivered as intended, it cannot produce the desired effect. In contrast, implementation failure indicates that the theory is sound, but the intervention has been poorly or inadequately delivered (e.g. inadequate number of training sessions, low adherence, insufficient resources).^{231, 232} A hypothetical example¹ can be drawn from a school-based programme aimed at reducing obesity. Theory failure would occur if the programme assumes that providing nutritional information alone is sufficient to change behaviour. In this case, pupils may understand the information but fail to modify their habits, indicating that the underlying theory is inadequate. On the other hand, implementation failure would occur if a programme were designed to include workshops, physical activities, and follow-up, but schools delivered only the workshops, with no physical activities and limited follow-up. In this case, the intervention would not have been implemented as intended.

In the case of implementation failure, process data may further identify issues such as poor delivery or contextual mismatch.^{230, 233, 234} For example, poor delivery occurs when an intervention with potential effectiveness is inadequately implemented (e.g. low intensity, reduced exposure, poor participation, or incomplete delivery).^{230, 234} In the school obesity example, poor delivery would occur if teachers received insufficient training, activities were delivered irregularly, or only a small proportion of pupils participated in the programme. Contextual mismatch arises when an intervention may be effective in principle but is not suited to the context in which it is introduced (e.g. cultural norms, available resources, healthcare systems, or environmental factors).²³² For example, the obesity programme may prove ineffective in schools located in some areas where pupils have limited access to healthy food or safe spaces for physical activity outside school. In this case, the intervention's limited effects would not necessarily reflect a flawed theory or poor delivery, but rather a mismatch between the intervention design and the broader social and environmental context.

Taken together, these findings highlight an important consideration: outcome evaluation alone may lead to the conclusion that “the intervention does not work”. However, without process evaluations, it remains unclear whether:

- the underlying theory is flawed,
- the implementation is inadequate,
- the intervention is poorly adapted to its context, or
- the intervention is genuinely ineffective.¹⁵

5.2.2 Other roles of process evaluations

Beyond providing a detailed understanding of intervention effects, process evaluations also enable the assessment of the **transferability** of findings (e.g. which mechanisms are context-sensitive and what adaptations may be required in other settings) and their **scalability** (e.g. which implementation conditions are necessary and which organisational capacities are critical).^{228, 235}

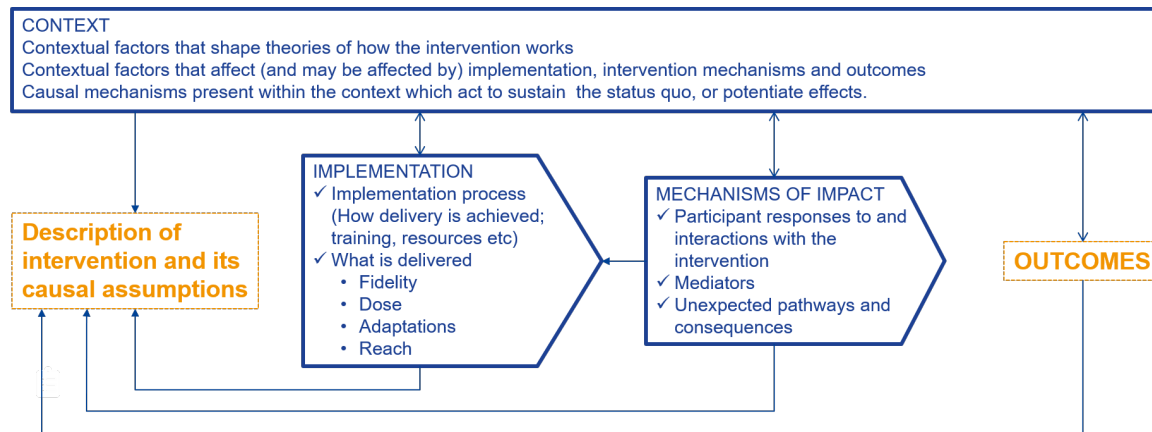
5.3 Three main dimensions: implementation, mechanisms of impact, context

Several frameworks have been developed to guide the design and conduct of process evaluations. Among these, the framework proposed by the Medical Research Council (MRC) provides a widely used and structured approach.^{10, 25} The framework emphasises the causal assumptions underlying the intervention and the use of evaluation to understand how the intervention works in practice. These causal assumptions may come from different sources, including the social sciences, common sense, or prior experience.²⁵ Process evaluations are conceptualised around three core dimensions and their interrelations: implementation, mechanisms of impact, and context. Together they offer a

¹ This hypothetical example on a school-based programme to reduce obesity was generated by AI.

comprehensive basis for examining how complex health interventions are delivered and how they produce effects in real-world settings (Figure 31).

Figure 31 – Key functions of process evaluations and their interrelations



Source: Moore *et al.*²³⁶

5.3.1 Implementation

This point describes the process by which a planned intervention becomes a real one.^{25, 237} To do so, process evaluations aim to provide a detailed understanding of implementation by examining ‘how’ it was delivered and ‘what’ was actually delivered, covering four aspects:

- fidelity: was the intervention delivered as intended?
- dose of the intervention: ‘how much’ (quantity) was delivered?
- adaptations: what changes were made in the intervention?
- reach: was the intended audience reached? What is the participation rate?

Some authors have also suggested that examining user experiences can complement these assessments by providing additional insight into intervention uptake.²³⁸

Alongside this, the ‘how’ refers to the process by which the intervention was delivered. It includes the characteristics of the providers such as their professional profile, training and experience. It also covers the mode of delivery, for example whether it is face-to-face, digital or group-based. In addition, it considers the contextual and organisational conditions in which the delivery takes place. Finally, it looks at the quality of implementation, including the level of structure and standardisation, as well as how acceptable the delivery approach is.

This is essential for determining whether observed outcomes can be attributed to the intervention itself or to variations in its implementation.

5.3.2 Mechanisms of impact

Exploring the mechanisms through which interventions bring about change is crucial to understanding not only how the effects of the intervention occurred, but also how these effects might be replicated by similar interventions in the future. Analyses of mechanisms of change can test, refine or construct hypotheses with regard to both anticipated and unanticipated outcomes, using quantitative and/or qualitative methods. Process evaluation may test hypothesised causal pathways using quantitative data complemented with qualitative methods to better understand complex pathways or to identify unexpected mechanisms. Mechanisms of change are also referred to as mediators, whereas contextual factors are referred to as moderators, which influence the strength of impact but are no integral part of the intervention-outcome relationship.^{25, 237, 239}



5.3.3 Context

Process evaluations also analyse the contextual factors that shape theories of how the intervention works and the factors that affect/may be affected by implementation, intervention mechanisms and outcomes. As we can see in Figure 31, it goes beyond the context itself and also includes the relationships between all the components of the model (the context is transversal to the implementation and the mechanisms of impact).^{25, 237}

Contextual factors should be understood as *'anything external to the intervention that may act as a barrier or facilitator to its implementation, or its effects'*.²⁵ These include for example epidemiological, ethical, geographical, legal, political, socio-cultural or socioeconomic context. From a system perspective, this involves examining how the intervention interacts with, and is reshaped by, the context in which it operates (see Chapter 10). Such analysis helps to explain why an intervention may be effective in some settings but not in others.²⁵

The MRC framework emphasises that these three dimensions are not independent but are dynamically interrelated.²⁵ Context may shape implementation, and implementation in turn influences mechanisms of impact. These mechanisms may also vary according to contextual conditions, underscoring the need for a systemic perspective.

In this regard, theories of change and logical models (see Chapter 1) help to improve transparency and understanding of how interventions work, while also structuring the evaluation in line with the context.

5.3.4 When to do a process evaluation?

A process evaluation can be undertaken at different points in the lifecycle of an intervention, including during its development, evaluation, and implementation. Each stage serves a distinct purpose, and the focus of the process evaluations vary accordingly. While currently KCE will not develop interventions in the context of the evaluation programme, it can be asked to evaluate the feasibility of an intervention that is under development. A process evaluation can, in that stage, support the assessment of the practicality of the intervention and help to optimise both its design and its evaluation methods. Nevertheless, a process evaluation within the context of the KCE evaluation programme will be most relevant to complement an effectiveness evaluation. When assessing effectiveness, it helps to understand what was actually implemented in terms of quantity and quality. It also involves examining how contextual factors may influence outcomes and the generalisability of the findings. This, in turn, strengthens confidence in effectiveness conclusions.^{25, 237} A process evaluation can also be considered as the only component of a study conducted in the context of the KCE evaluation programme. This should preferably be restricted to understand the implementation aspects of an intervention that is evidence based.

5.4 Theories, models and frameworks (TMFs)

The development of theories, models and frameworks (TMFs) to support process evaluation is a domain that is constantly evolving. In this section, we present only a selection of TMFs to provide general direction for the KCE team responsible for conducting a process evaluation. However, the team should explore the literature to identify relevant alternative TMFs, depending on the topic under study and the research questions addressed in the process evaluation.

5.4.1 Why use TMFs?

The use of TMFs in process evaluations is essential to provide a structured analytical approach through which interventions can be understood. TMFs function as conceptual tools that make the underlying assumptions of an intervention explicit. They also help to clarify its expected causal pathways, and to identify the mechanisms through which outcomes are anticipated to occur.

TMFs may also help to guide the different stages of the evaluation process, by informing the data collection, shaping analytical strategies, and support the interpretation of findings. This approach strengthens methodological rigour and transparency by grounding analytical choices in a clearly defined conceptual framework.^{25, 240, 241}

The selection of an appropriate TMF should be closely aligned with the research question and the primary focus of the process evaluation. This will require careful preparation (e.g., review of literature, talking to commissioner(s) as well as document analysis to understand the questions that should be addressed) and discussions within the KCE research team.²⁵

The increasing number of available TMFs reflects the growing recognition that different evaluation questions require different conceptual lenses. Rather than competing with one another, TMFs are often complementary, each focusing on specific aspects of implementation, context, mechanisms of impact or outcomes. Consequently, combining multiple TMFs within a single process evaluation is increasingly common, particularly when evaluating complex interventions.²⁴²

Different TMFs emphasise the three distinct dimensions. For example, if the evaluation aims to better understand intervention fidelity and delivery, **implementation**-focused TMFs may offer a particularly relevant lens, while still allowing consideration of other dimensions such as context and mechanisms. A study primarily aiming to understand the causal pathways through which an intervention generates its effects will prioritise TMFs that focus on **mechanisms of impact**. When the evaluation aims to understand the impact of external influences, the KCE team will choose a TMF that is focused on a **contextual analysis**. Several TMFs may be used within a process evaluation, reflecting the diversity of research questions, and to reflect the complexity of the intervention. Nonetheless, the researchers should explain why the framework(s) was chosen and how it was used.^{9, 25, 240}

5.4.2 Main characteristics of a selection of TMFs

In the next sections, we present a selection of TMFs commonly used or relevant in the process evaluations of complex health interventions. These models are grouped according to their primary analytical focus (mechanisms of impact, context and/or implementation). This pragmatic classification is adapted from the taxonomy of Nilsen (2015)^{m, 240}

As this is a rapidly evolving field, the overview of TMFs is not intended to be exhaustive. Therefore, representative TMFs are presented in the main text, with additional frequently cited models included in Appendix 3. Nevertheless, KCE teams are encouraged to undertake a targeted search tailored to their specific evaluation study.

Table 27 – Main characteristics of a selection of TMFs

TMFs	Main characteristic	See details
<i>RE-AIM</i>	Evaluation of the impact and reach of interventions through five dimensions from which indicators are derived: reach, effectiveness, adoption, implementation, and maintenance	Table 28
<i>CICI</i>	Multidimensional analysis of context and its influence on the implementation and outcomes of complex interventions	Table 29

^m TMFs are grouped according to their primary analytical focus. This classification should not be interpreted as suggesting that the other dimensions are ignored; rather, it reflects the lens through which the framework primarily approaches process evaluation, while still considering implementation, mechanisms, and context as interconnected components.



TMFs	Main characteristic	See details
CFIR	Comprehensive identification of multilevel determinants influencing implementation (intervention, individuals, organisation, outer context, process)	Table 30
ENCOMPASS evaluation framework	Evaluation of interventions within complex adaptive systems, with attention to interactions and evolving contexts	Table A. 1*
NASSS	Understanding factors influencing (non-)adoption, scale-up, spread, and sustainability of health and care innovations, particularly technology-enabled interventions	Table A. 2*
BANANA	Structured and in-depth contextual analysis before and during implementation	Table A. 3*
TDF	Analysis of behavioural determinants influencing practices and behaviour change	Table A. 4*
i-PARIHS	Analysis of implementation success through the interaction between innovation, recipients, context, and facilitation, with facilitation acting as the active ingredient of implementation	Table A. 5*
EPIS	Process-oriented approach to the different phases of implementation, from exploration to sustainment	Table A. 6*
Proctor's Implementation Outcomes Framework	Conceptualisation and measurement of implementation outcomes distinct from service and clinical outcomes (e.g., acceptability, adoption, appropriateness, feasibility, fidelity, penetration, sustainability, implementation cost)	Table A. 7*
FRAME	Systematic documentation of adaptations and modifications made to interventions	Table A. 8*
FRAME-IS	Systematic documentation of adaptations specifically related to implementation strategies	Table A. 9*

*Table available in Appendix 3.

The research team may choose to use all or only selected components of the TMF (e.g. constructs or domains, depending on the terminology adopted within the framework). Likewise, within each component, either all subcategories or only a subset may be used. This flexibility is intended to ensure the framework can be tailored appropriately to the research questions under investigation. Yet, the choices should be clearly documented and justified.

5.4.3 Focus on the mechanisms of impact and outcome: the 'outcome- and system-oriented' TMFs

Outcome- and systems-oriented TMFs focus on assessing the effects of interventions across multiple levels, including intended and unintended outcomes and system-level changes. Within this group of TMFs, the **RE-AIM framework** is frequently used (Table 28).²⁴³ In addition, the ENCOMPASS framework may be considered and is presented in Appendix 3.1.⁵⁵

Table 28 – RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance

RE-AIM	
Origin of the TMF	Developed by Russell Glasgow and colleagues (1999; expanded over time)
Objectives	▪ Evaluate the public health impact and external validity of interventions, including scalability and transferability ▪ Assess outcomes at both individual- and organisational-levels ▪ Support translation of evidence into practice ▪ Guide intervention planning, implementation, evaluation, and reporting
Main components	5 dimensions: - Reach: participation rate and representativeness of individuals - Effectiveness: impact on outcomes, including unintended effects - Adoption: uptake by settings and staff - Implementation: fidelity, adaptations, cost - Maintenance: sustainability over time at individual and organisational levels All five dimensions should ideally be considered to provide a comprehensive assessment of intervention impact and potential for dissemination and sustainability; however, specific dimensions may be prioritised according to study objectives
When to use?	Can be used to design, implement and evaluate interventions To structure evaluation of intervention impact in real-world settings (including in public health and health services research)
Comments	Applicable prospectively (planning) and retrospectively (evaluation) Can be used to conceive the intervention, to adjust implementation or to evaluate its impact Used for comparative and pragmatic evaluations

Sources: Glasgow et al.²⁴³, Jung et al.²⁴⁴ and the RE-AIM website²⁴⁵

Based on the RE-AIM framework but broader in scope, the PRISMⁿ (Practical, Robust Implementation and Sustainability Model) examines how an intervention interacts with its recipients and environment through multiple contextual factors. It also considers how these contextual factors influence reach, effectiveness, adoption, implementation, and maintenance. PRISM can be used during the planning, implementation, and evaluation of health interventions.^{246, 247}

PRISM encompasses four main domains:

- **intervention:** the characteristics of the programme (including its complexity, costs, and adaptability);
- **recipient characteristics:** the perspectives and characteristics of the ‘implementers’ (e.g. clinicians) and the users (e.g. patients or community members);
- **implementation and sustainability infrastructure:** the processes, training, and resources available to support the initial implementation and the long-term sustainment of the intervention;
- **external environment:** policies, financial incentives, and regulatory frameworks that may facilitate or hinder the implementation and sustainability.

PRISM also examines how these four domains interact with the RE-AIM dimensions.^{246, 247}

ⁿ PRISM is presented within this subgroup as it builds on the RE-AIM framework by explicitly integrating contextual and system-level factors that influence implementation and outcomes.



5.4.4 Focus on the context: 'context and system complexity' TMFs

These TMFs address the complexity of systems and the contextual interactions, including how they affect implementation and outcomes. The **Context and Implementation of Complex Interventions (CICI) model** is presented in Table 29.²⁴⁸ The Non-adoption, Abandonment, Scale-up, Spread and Sustainability (NASSS) framework and the Basel Approach for coNtextual ANALysis (BANANA) are presented in Appendix 3.2.^{249, 250}

Table 29 – CICI: Context and Implementation of Complex Interventions model

CICI	
Origin of the TMF	Developed by Lisa Pfadenhauer and colleagues (2017)
Objectives	▪ Provide a structured and comprehensive conceptualisation of context and implementation in complex interventions ▪ Distinguish while linking context, implementation and setting ▪ Support systematic assessment of how contextual factors influence implementation processes and intervention outcomes ▪ Improve the evaluation, synthesis, applicability and transferability of complex interventions → By understanding why an intervention functions differently across contexts, it becomes possible to enhance the transferability, applicability and transferability of interventions
Main components	Three dimensions interacting with each other and with the intervention dimension: 1. Context (7 domains): - Epidemiological - Ethical - Geographical - Legal - Political - Socio-cultural - Socio-economic 2. Implementation (5 domains): - Implementation theory - Process - Strategies - Agents - Outcomes 3. Setting: specific physical location in which the intervention is delivered and implemented, where intervention, implementation and context interact. Context, implementation and intervention-related phenomena may operate at micro-, meso- and macro-levels and interact across these levels → Multi-dimensional structure (components can be used selectively depending on study focus)
When to use?	To analyse the influence of context on intervention implementation and outcomes To design, evaluate and synthesise complex interventions To assess transferability and applicability across settings To structure context reporting in implementation and process evaluations
Comments	Interacting constructs rather than hierarchical levels Avoid rigid distinctions between context, setting, and implementation Strong emphasis on contextual complexity and multi-level interactions Checklist available to facilitate the pragmatic application of the CICI framework

Source: Pfadenhauer et al.²⁴⁸

5.4.5 Focus on implementation and context: ‘determinant’ TMFs

These TMFs address the following questions: Why does implementation succeed or fail? and Why is an intervention (not) effectively implemented (including the role of the contextual factors)?

The **Consolidated Framework for Implementation Research (CFIR)** helps to identify barriers and facilitators across multiple levels (e.g. micro, meso and macro), that may influence whether the intervention is implemented as intended.^{241, 251} Its main characteristics are summarised in Table 30. The Theoretical Domains Framework (TDF) and the Integrated-Promoting Action on Research Implementation in Health Services (i-PARIHS) are presented in Appendix 3.3.^{252, 253}

Table 30 – CFIR: Consolidated Framework for Implementation Research

CFIR	
Origin of the TMF	Developed by Laura J. Damschroder and colleagues (2009; updated 2022)
Objectives	▪ To identify, predict and explain determinants (barriers and facilitators) influencing implementation outcomes and effectiveness → What contextual factors influenced the implementation? ▪ Adapt an intervention to a given context ▪ Improve the effectiveness and sustainability of projects ▪ Compare implementations between sites or contexts
Main components	<u>CFIR 2.0 comprises five domains containing multiple constructs and subconstructs that capture determinants of implementation across multiple levels</u> 1. Innovation domain (characteristics of the innovation that may influence implementation): - Innovation source - Evidence base - Relative advantage - Adaptability - Trialability - Complexity - Design and packaging 2. Outer setting domain (external contextual influences): - Critical incidents - Local attitudes - Local conditions - Partnerships and connections - Policies and laws - Financing - External pressure (societal, market, performance measure pressure) 3. Inner setting domain (organisational context influencing implementation): - Structural characteristics (physical infrastructure, information technology infrastructure, work infrastructure) - Relational connections - Communication - Culture (human equality-, recipient-, deliverer-, learning-centeredness) - Implementation climate - Readiness for implementation - Other innovation implementation and innovation delivery determinants 4. Individuals' domain (actors involved in implementation and innovation delivery): - High-level leaders - Mid-level leaders - Opinion leaders - Implementation facilitators - Implementation leads - Implementation team members - Other implementation support



	- Innovation deliverers - Innovation recipients - (+ subdomain on characteristics subdomain: need capability opportunity, motivation) 5. Implementation process domain (implementation activities over time): - Teaming - Assessing needs (innovation deliverers, innovation recipients) - Assessing context - Planning - Tailoring strategies - Engaging (innovation deliverers, innovation recipients) - Doing - Reflecting and evaluating (implementation, innovation) - Adapting The 2022 update refined construct definitions and terminology, strengthened equity considerations, clarified distinctions between innovation, recipients and context, and improved the representation of multilevel interactions → Not all constructs are required in a single study
When to use?	CFIR can be used to inform choice of implementation strategies that may best address contextual determinants, generate hypotheses to prospectively guide predictions of implementation outcomes, or retrospectively explain implementation outcomes by assessing differences in determinants across implementation settings
Comments	The CFIR is designed to be modular and adaptable to the context of the study. <u>Selection of the constructs:</u> based on their relevance to the research question, their plausibility as determinants of implementation, and the feasibility of investigating them within the available resources → Depending on the aims and scope of the evaluation, an entire CFIR domain, a targeted selection of constructs, or the entire framework may be used. The selection of constructs should be justified and their role in relation to the development of the interview guide, the coding framework, the analytical process, and the interpretation of findings explicitly described. The CFIR is primarily designed to capture perspectives from stakeholders who influence, experience or are affected by implementation processes and outcomes. The CFIR must be fully operationalised prior to use in a project: 1) Define the subject of each domain for the project (see guidance for each domain below) 2) Replace broad construct language with project-specific language if needed 3) Add constructs to capture salient themes not included in the updated CFIR

Sources: Damschroder²⁴¹ and the CFIR website²⁵¹

5.4.6 Focus on implementation: 'process' TMFs

These TMFs address how implementation unfolds, is operationalised, and can be evaluated throughout the implementation process. Some frameworks, such as the Exploration, Preparation, Implementation, Sustainment (EPIS), adopt a structured phase-based approach to implementation, emphasising clearly defined stages.²⁵⁴ Others, such as Proctor's Implementation Outcomes Framework, focus on evaluating key implementation outcomes including acceptability, adoption, appropriateness, feasibility, fidelity, penetration, and sustainability.²⁵⁵ Appendix 3.4 presents their main characteristics.

5.4.6.1 *Focus on adaptation and fidelity in implementation: 'adaptation and fidelity' TMFs*

A subcategory focuses on adaptation and fidelity in implementation. The main idea is that adaptations are inevitable and should therefore be documented, addressing this question: 'How was the intervention modified, by whom, why, and with what potential consequences for transferability?', covering modification or maintenance of the intervention over time. The Framework for Reporting Adaptations and Modifications-Enhanced (FRAME) focuses on adaptation of the intervention, while the Framework for Reporting Adaptations and Modifications to Evidence-based Implementation Strategies (FRAME-IS) focuses on the adaptation of the implementation strategies.²⁵⁶⁻²⁵⁸ The main characteristics of both frameworks are presented in Appendix 3.5.

5.4.7 *Focus on mechanisms of impact: 'mechanism-oriented' TMFs*

TMFs focusing on mechanisms of impact aim to explain how specific components of an intervention produce change. Within this perspective, realist evaluation is commonly used to examine Context-Mechanism-Outcome (CMO) configurations, providing a structured way to understand how outcomes are generated in particular settings.^{229, 259} The RAMESES II reporting standards may be used for reporting on realist evaluations.²⁶⁰ Chapter 6 elaborates on Realist evaluation.

5.5 Data collection methods in process evaluations

Process evaluations are not method-specific and should align methods with the research questions, ensuring an appropriate balance between breadth and depth. As illustrated in the MRC framework, process evaluations typically focus on three interconnected domains: context, implementation, and mechanisms of impact (see Figure 31), and data on intervention outcomes may also be collected to facilitate interpretation of the process evaluation findings. Data collection should therefore be designed to capture these different levels, using appropriate methods at each stage.²⁵ The choice of data collection methods should also be driven by the constructs defined by the selected TMF.²⁴²

Quantitative methods (e.g. surveys, routine monitoring data) are commonly used to assess implementation and outcomes, including reach, fidelity, and dose. **Qualitative methods** (e.g. stakeholder, implementer, and participant interviews, documentary analysis, and observation) are used to explore context, implementation processes, and mechanisms of change. Mixed-methods designs enable integration across these components, allowing findings from one source to inform and contextualise another.^{8, 25}

Study designs may be cross-sectional or longitudinal, including repeated measures to capture change, adaptation, and dynamics over time. Process evaluations should also be embedded within randomised controlled trials or quasi-experimental studies, allowing process data to be collected alongside effectiveness evaluation. Flexibility should be maintained throughout the evaluation, as preliminary findings may highlight unanticipated issues that warrant further investigation.⁸

5.6 Conduct of a process evaluation

The following steps are recommended when a process evaluation is conducted.

Process evaluations may require careful **planning** and a transparent definition of the **evaluators' role**, ensuring an appropriate balance between collaboration with stakeholders and methodological independence. From the outset, it may be important to clarify whether the evaluators will take an active role (e.g. providing feedback during implementation) or a more passive role, and to report this position clearly. It is recommended that the research team adopts a more active role during the intervention development and feasibility (pilot) phases, and a less active role during the subsequent evaluation and implementation phases.^{8, 237} As KCE will primarily conduct process evaluations in these later phases, an independent evaluator role should be adopted.



Adequate resources and staffing should be foreseen. This may include expertise in quantitative, qualitative, and mixed-methods approaches. Integration across these approaches should be planned, as well as the relationships within the evaluation team (e.g. coordination between process and outcome evaluators, strategies to support the integration of findings).^{8, 237} Depending on the complexity of the intervention and the evaluation objectives, involving expertise in implementation science may strengthen the design, conduct, and interpretation of a process evaluation.²⁶¹

The **evaluation design** should begin with a detailed description of the intervention and its underlying theory. This includes making explicit the assumptions about how the intervention is expected to be implemented and to work within its context, as well as specifying key components such as resources, activities, mechanisms, and expected outcomes. A logic model can be used to visually represent these elements (see Chapter 1).^{8, 237}

Research questions should be derived from the theoretical assumptions but remain flexible to allow for emerging questions during the evaluation. **Methodologically**, process evaluations typically combine quantitative and qualitative approaches, taking into account their respective strengths and limitations, and ensuring a balanced sampling strategy (neither too broad nor too narrow). The methods should be chosen based on the research questions. The findings should not be generated in parallel but inform one another throughout the evaluation process.^{8, 237}

Analysis should include, in most cases, both quantitative and qualitative data, as well as the integration of both.^{8, 237}

Finally, **reporting** should be comprehensive and transparent. There is no single model applicable to all contexts. It is important that the KCE team clearly reports the choices that were made during the process evaluation as well as the interconnection between the different evaluation components (e.g. effectiveness, process evaluation, economic evaluation). Reporting should also include the dissemination of the results to several audiences, including to policymakers and stakeholders.^{8, 237}

6 REALIST EVALUATION

KEY POINTS

- A realist evaluation assesses **what** works, **for whom**, **in what circumstances** and **why** rather than only evaluating if a programme works. This theory-driven approach aims to understand the causal processes that generate (or fail to generate) change in specific contexts.
- A realist approach can be applied **at all stages of the evaluation** of complex interventions, as described in the MRC guidelines. Yet, the KCE evaluation programme is not equally involved in each stage.
 - While realist evaluation is particularly suited for the **development phase** (i.e. building the programme theory) the development stage of complex interventions is, currently, **not within the scope of the KCE evaluation programme**.
 - In terms of **feasibility**, a small-scale realist evaluation study may be performed when there is **uncertainty** about the **programme theory** underlying an intervention or about the feasibility of its **implementation** and evaluation. Such studies can help strengthen the programme theory and inform the design of a subsequent larger-scale evaluation.
 - In the **evaluation phase**, realist evaluation can be used to understand **why** and **how** an intervention **works** (or does not work) in a certain context.
 - In the **implementation phase**, realist evaluation goes one step further than confirming whether an intervention worked. It **clarifies** which **mechanisms** drove that effect, for whom, and under which conditions – insights that directly inform how implementation strategies should be adapted for scale-up or transfer to new settings.
- The **realist evaluation research cycle** includes the following steps:
 - Developing and/or understanding the **programme theory** (see Chapter 1), i.e. identifying context, mechanisms, and outcomes, ultimately translated into Context-Mechanisms-Outcomes configurations (CMOCs) and Context-Mechanism-Outcomes configuration hypotheses (including rival theories).
 - Collecting and analysing **data to test the CMOCs** and look for additional Context-Mechanism-Outcome patterns in the data.
 - **Synthesising the results** and formulating conclusions on what works, for whom, in what circumstances.
 - Providing a **feedback loop** from the findings and conclusions into the existing theory.
- Nevertheless, in the KCE evaluation programme the evaluation may also end with policy relevant output, informing policymakers about what works, for whom under which conditions.

6.1 Introduction

In this chapter, we describe the realist evaluation approach in the context of the KCE evaluation programme. Throughout the text, the realist evaluation concepts will be made more tangible by means of examples from a former KCE project about the remote monitoring of patients with COVID-19.²⁶³ In December 2020, remote monitoring of COVID-19 patients was launched to save hospital beds and relieve strain from hospital workforce and primary care providers, while still addressing patients' care needs. Twelve pilot projects were set up and KCE was commissioned to carry out a study to evaluate this new way of providing care remotely within each of these pilot projects. While this single example



is used for consistency throughout the chapter, the realist approach has been applied across different research domains.²⁶⁴

6.2 What is realist evaluation?

Realist evaluation is a theory-driven approach, originally developed by Pawson and Tilley.²⁶⁵ Instead of evaluating whether a programme works, it assesses *what works, for whom, in what circumstances, and why*.^{260, 265, 266} Rather than comparing outcomes between those who did and did not receive an intervention (as in a randomised controlled trial where the standard cause-effect question asks ‘Does this work?’), realist evaluation compares Context-Mechanism-Outcome Configurations (CMOCs) within programmes, asking why the same intervention produces different outcomes in different settings and for different populations (i.e. across implementation sites).²⁶⁷⁻²⁶⁹ As such, in a realist evaluation, the basic premise in how reality works is that interventions do not produce *outcomes* mechanically or uniformly, but by triggering underlying causal processes i.e. *mechanisms* in people who reason and respond differently depending on their circumstances (*context*).

6.2.1 The core conceptual framework: Context-Mechanisms-Outcome Configurations

The realist evaluation approach identifies three fundamental components of an intervention, namely context (C), mechanism (M) and outcome (O) (the so-called Context-Mechanism-Outcome Configuration or CMOC, see section 6.3.2) and examines how they interact.

The basic causal logic is: **Context + Mechanism = Outcome** (for example: dry, confined space (C) + spark (applied to gunpowder) (M) = explosion (O)).^{265, 268, 270} A practical way to read this, for example, is: ‘*Interventions achieve a successful outcome when theory-driven mechanisms are activated (‘fired’) under appropriate contextual conditions*’.^{260, 265} In other words, it is assumed that interventions ‘work’ in different ways for different people in different situations.^{260, 265, 266}

Some researchers have extended the basic CMOC to make additional analytic elements explicit, such as CIMO (Context – Intervention – Mechanism – Outcome) and ICAMO (Intervention – Context – Actor – Mechanism – Outcome).²⁷¹ The ICAMO configuration is particularly valuable for health policy, where a national policy instrument passes through multiple layers of actors, i.e. national, regional, local, individual, before reaching the end beneficiary.

Adding explanatory factors is possible and can be insightful depending on the study’s scope and aims. A minimum requirement is, however, that the **causal link** between **context**, **mechanism**, and **outcome** (CMO) is explained.

6.2.1.1 Context (C)

Context refers to the pre-existing conditions, i.e. cultural, social, institutional, financial conditions, existing policies or structural, that influence whether and how mechanisms are triggered (‘fired’).^{265, 272} Context shapes the causal process as different mechanisms may be activated (and produce different outcomes) depending on the setting.

Textbox 13 – Example of context (C) from the project on telemonitoring in COVID-19 patients

When telemonitoring teams are exempt from other hospital duties (Context), they will have higher patient inclusion rates (Outcome), then when they must do the telemonitoring in combination with their regular hospital duties.

6.2.1.2 Mechanism (M)

Mechanisms are the deeper, often unobservable, causal processes through which an intervention generates change.^{265, 266, 273, 274} Pawson and Tilley (1997)²⁶⁵ originally proposed three ways of conceptualising mechanisms: as programme components (the active ingredients delivered), as participant reactions to those components, and a generalisable explanatory accounts of causal processes (so-called 'demi regularities').^{260, 275} They state that mechanisms will only activate in the right conditions, providing the guiding formula: C (context) + M (**mechanism**) = O (outcome). While this typology clarifies what the concept of mechanism encompasses, it has been criticised for creating confusion in implementation science, particularly between what is delivered and why it works (or not).^{273, 274} Dalkin et al. (2015)²⁷⁴ has addressed this concept by explicitly disaggregating mechanism into two related but analytically separable components: resources and reasoning. It helps to understand the difference between the resources offered by the intervention and the ways in which this changes the reasoning of participants, which in turn helps to distinguish between a context (i.e. pre-existing conditions) and a mechanism. This approach shows the mechanism is not the resource itself (i.e. a device, a protocol, staffing) but the change in reasoning that the resource triggers in a particular context: how participants make sense of, respond to, and act upon what is offered. For that, Dalkin et al. stress to understand the formula: (**M (Resource)** + C (Context) -> **M (Reasoning)** = O (Outcome). This explains that resources should be introduced into a pre-existing context, which together induces an individual's reasoning, leading to an outcome.

Textbox 14 – Example of mechanism (M) from the project on telemonitoring in COVID-19 patients

The pulse oximetry and daily monitoring are 'the resources'; the 'mechanism' is the sense of safety and clinical ownership these resources trigger in patients and clinicians respectively. Whether that 'reasoning' is activated depends on the 'context' (e.g. health literacy, workload, organisational support).

6.2.1.3 Outcome (O)

Outcomes are the intended/expected or unintended/unexpected, short- or long-term consequences/results of the intervention on the micro-, meso-, or macro-level.^{266, 276} In realist evaluation, outcomes are disaggregated by context and mechanism to test whether outcome patterns match CMOC assumptions revealing for whom the intervention works, under what conditions, and through which pathways. In the telemonitoring example, outcomes may be e.g. patient inclusion rates and timely escalation when deterioration is detected.

Textbox 15 – Example of outcome (O) from the project on telemonitoring in COVID-19 patients

In the telemonitoring projects it was observed that some sites had higher patient inclusion rates and deterioration was detected earlier (Outcomes). Consequently, the question is raised how and why these sites achieved these outcomes while other sites did not, despite the same policy. For that we need to understand the (un)observable mechanisms at work. It was seen through interviews, focus groups, and document analysis on the organisational structure that, although the same policy was in place, there were contextual differences between sites, such as whether:

- telemonitoring teams were freed from regular hospital duties,
- had decision-making authority,
- had clear escalation protocols available.



The identified mechanism that ‘fired’ was the telemonitoring teams’ experience of clinical ownership over the monitoring process.

6.2.2 *Applying realist evaluation principles across the MRC framework*

The UK Medical Research Council (MRC) framework for evaluating complex interventions recommends development, feasibility testing and piloting, evaluation and implementation as the key elements of the development and evaluation process (Figure 1).¹⁷ We agree with Fletcher et al. (2016) that a realist approach can be integrated across the different phases within the latest MRC guidance on evaluating complex interventions, especially since the latest version places increased emphasis on causal mechanisms and theory development compared to the initial MRC model.²⁷⁷

A realist approach can be applied at all stages of the evaluation of complex interventions, but the KCE evaluation programme is not equally involved in each stage.

6.2.2.1 *Intervention development phase*

While realist evaluation is particularly suited for the development phase (i.e. building the programme theory) the development stage of complex interventions is, currently, not within the scope of the KCE evaluation programme. However, KCE could be involved in intervention development outside the evaluation programme. For example, a preparatory study (e.g. consisting of a realist review) could be part of the classic annual KCE programme which could then be used by policymakers (e.g. RIZIV – INAMI, FOD – SPF) to develop a policy intervention. In a next step, KCE could be asked to evaluate this intervention within the context of the KCE evaluation programme.

In the intervention development phase **realist reviews** – also called realist syntheses²⁷⁸ – are particularly useful. When systematic reviews provide input for intervention development, the focus is typically on synthesising quantitative studies answering ‘what works’ at the expense of understanding how, in what context and for whom.²⁷⁷ Realist reviews are a form of systematic literature review. They have been suggested as an alternative or adjunct to address the lack of attention for CMOs in evidence syntheses. They synthesise evidence from existing research into elements of the programme theory. This might include, for example, studies identifying outcome evaluations, research trials, literature reviews, or comparative case studies.^{260, 279} Reviews can also use qualitative comparative analysis (QCA), a methodology that enables the analysis of multiple cases in complex situations and aims to explain why change happens in some cases but not in others. Both can be applied to secondary as well as primary data.²⁸⁰

6.2.2.2 *Feasibility testing and piloting phase*

Feasibility and pilot studies can also adopt a realist approach to assess what is feasible for whom under what circumstances. It can also be used to refine hypotheses about potential mechanisms and how these vary by context. This way, the programme theory and the intervention design can be optimised prior to the evaluation phase (e.g. realist RCT).²⁷⁷

This is consistent with the KCE evaluation programme. Commissioners may request a small-scale evaluation study when there is uncertainty about the programme theory underlying an intervention or about the feasibility of its implementation and evaluation. Such studies can help strengthen the programme theory and inform the design of a subsequent larger-scale evaluation.

^o Note that realist reviews or realist syntheses are different from section 6.3.7 describing ‘synthesis’ as a step within the realist evaluation research cycle.

6.2.2.3 *Evaluation phase*

How to use realist evaluation in the evaluation phase of the MRC framework is explained in 6.3, which describes this step by step.

Note that within the realist evaluation approach, a realist RCT is a particular design. A realist RCT layers a CMOC based explanatory analysis on top of a standard RCT (comparing outcomes between those who did and did not receive an intervention) to explain why and for whom to intervention worked. It combines the advantages of bias minimisation in estimating intervention effects via randomisation with the ability to theorise the mechanisms underlying these effects and how they differ by setting.⁹² In addition, realist RCT methods can be combined with process evaluation to understand mechanisms and context.²⁵

6.2.2.4 *Implementation phase*

Where interventions are scaled up after successful evaluation, realist approaches can examine long-term variations by context. This kind of implementation studies can occur over a wide range of settings, populations and time periods, to understand how context shapes outcomes. When an intervention programme is already running, the central question shifts from 'How might this work?' (development phase) to 'Is it working as theorised, for whom, where, and why not everywhere?' (implementation phase).

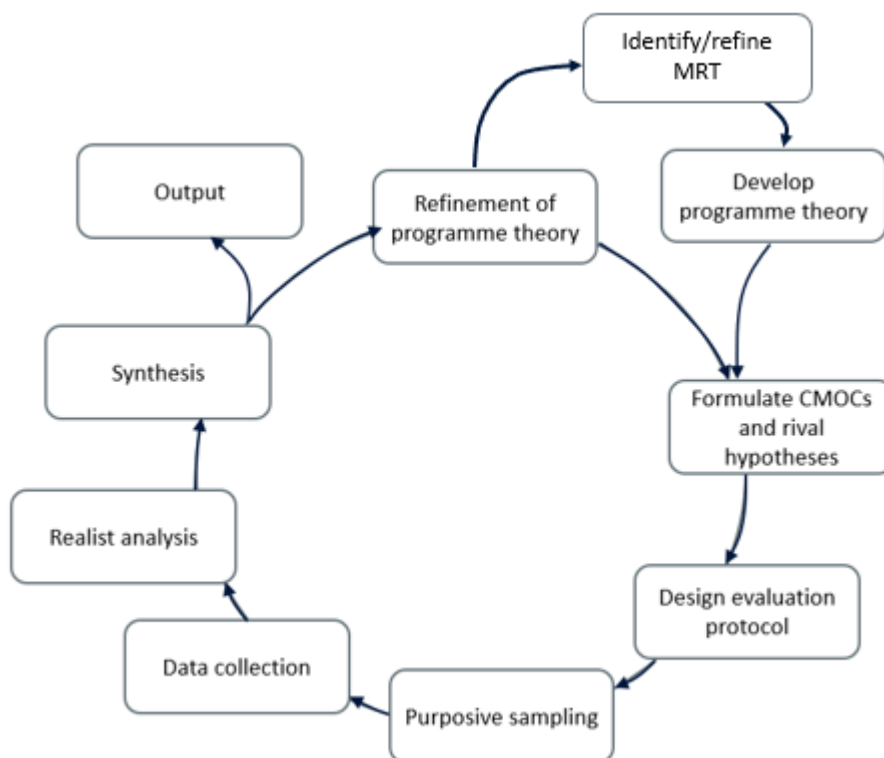
In the context of the KCE evaluation programme, realist evaluation can therefore play an important role during the implementation phase. A strong evidence base for an intervention does not necessarily imply that the promising results observed under study conditions will automatically be realised when an intervention is scaled up and broadly implemented in routine practice. By identifying which mechanisms are triggered, for whom, and under which contextual conditions, a realist approach can provide actionable insights to adapt and improve implementation strategies.

6.3 How to use realist evaluation: the realist evaluation research cycle

The original realist research cycle as formulated by Pawson and Tilley (2004)²⁸¹ presents the steps of hypotheses generation, data collection and analysis, and theory testing as baseline model. However, several authors added elements or specifications.^{262, 268, 278, 281, 282} Figure 32 is the result of an integration of several representations of the realist research cycle into one.



Figure 32 – The realist evaluation research cycle



Source: Adapted from Pawson and Tilley, 1997)²⁶⁵

In the following paragraph, we will elaborate on each step in **the realist evaluation research cycle**. In short, realist evaluation starts from developing and/or understanding the programme theory (Chapter 1), i.e. identifying context, mechanisms, and outcomes, ultimately translated into CMOCs and CMOC hypotheses (including rival hypotheses, see section 6.3.2).²⁶⁵

With the programme theory established, the realist evaluation proceeds through data collection and data analysis to test the CMOCs and look for additional CMO patterns in the data. Finally, the data are synthesised and conclusions are formulated on what works, for whom, in what circumstances, i.e. the policy relevant theory.²⁷⁸ Ultimately, the cycle feeds back to existing theory which represents the cumulative, iterative nature of realist evaluation.^{265, 283} Each evaluation produces a refined programme theory which, when synthesised across evaluation of similar programmes, contributes to the refinement of Middle Range Theories (MRTs) and a broader generalisable evidence base.^{278, 283} This is what distinguishes realist evaluation from a one-off evaluation. It is designed to build generalisable knowledge, not just programme-specific findings. Instead of cycling back to more MRT refinement, the evaluation may also end with a policy relevant output, informing policy makers about what works, for whom under which conditions (see Figure 32).

Although the testing and refining of the programme theory is an explicit purpose of a realist evaluation, as much as it is to determine whether and how the intervention worked in a particular context, KCE targets only the latter. Refining the programme theory as the start of a new research cycle is rather unlikely, unless several related studies are commissioned covering different phases of the realist evaluation research cycle. In that case, the refined CMOCs from one evaluation become the starting hypotheses for the next. In all other cases, for KCE purposes the realist evaluation research cycle is not part of an iterative process but ends with providing policy relevant output (see Figure 32).

6.3.1 Develop programme theory

Realist evaluation is theory-driven and the evaluation cycle therefore starts with theory development.^{265, 284} The purpose of the cycle is to start with the identification of a MRT from which the programme theory can be deduced. **Middle Range Theories (MRTs)** are specific enough to generate hypotheses and general enough to apply across different situations. They are a result of previous research endeavours and can be used as input for programme theory specification at the start of the cycle. An MRT is not invented for a specific programme but brought to it as theoretical lens. The evaluator's task is to assess whether and how these general patterns manifest in the specific programme under evaluation. A realist programme theory specifies the mechanisms that are expected to generate the outcomes in which context.²⁷⁰

A well-established example of a MRT with particular relevance for implementation research is **Normalization Process Theory (NPT)**, which describes how new practices become embedded in routine work (see Textbox 16). An example of a more substantive MRT on self-efficacy and professional motivation is provided in Textbox 17.

Textbox 16 – Normalization process theory (NPT)

Normalization process theory (NPT) can be applied as a Middle Range Theory (MRT) for implementation research. NPT is a sociological MRT that explains how new practices, technologies, or complex interventions become or fail to become embedded in routine work.²⁸⁵⁻²⁸⁹ It focuses on the work that people need to do to ensure that interventions become *normalised*.²⁸⁹ NPT can also be used as a framework to support the optimisation of a trial intervention by supporting intervention design, describing the context of a trial, and supporting the interpretation of a trial's results.²⁸⁹ It is particularly well-suited as a theoretical lens for realist evaluation of complex health interventions, because its four constructs map directly onto candidate mechanisms in a CMOC.^{289, 290}

These four constructs are:²⁸⁹

- **Coherence:** the sense-making work people do to understand what a new practice is, why it matters, and how it differs from existing practice.
- **Cognitive participation:** the relational work of building and sustaining commitment among those who must enact the practice.
- **Collective action:** the operational work of enacting the practice in everyday settings, including the division of labour and workability.
- **Reflexive monitoring:** the appraisal work of assessing whether the practice is working and adjusting accordingly.

Each construct can serve as a candidate mechanism in a realist programme theory. For example: in contexts where practitioners already share a clear understanding of why a new monitoring programme is valuable (coherence as context/mechanism), collective action is more readily sustained (outcome: successful embedding), particularly when roles and responsibilities are clearly assigned (context). Conversely, where coherence is absent, for instance because telemonitoring is perceived as an addition to existing workload rather than an improvement to care, collective action is harder to sustain and the practice fails to normalise.

Taking the telemonitoring in COVID-19 example, NPT-derived candidate mechanisms include clinicians developing shared understanding of the purpose of remote monitoring (coherence), a clinical leader actively enrolling colleagues into the pathway (cognitive participation), a workable and clearly distributed monitoring workflow (collective action), and teams reflecting on escalation outcomes to refine their approach over time (reflexive monitoring). The four constructs broadly follow a sequential logic over time – from sense-making and commitment-building, through enactment, to appraisal – though they frequently overlap in practice.²⁸⁹



NPT has been validated as a useful theoretical lens across diverse healthcare implementation studies and has been shown to be complementary with realist evaluation, with its constructs lending themselves directly to CMO hypothesis formulation and realist interviewing design (see Textbox 20).²⁸⁹⁻²⁹¹ A structured coding manual linking NPT constructs to CMO elements is available to support analysis.²⁹¹

Theory development is not a main aim in the KCE evaluation programme, but MRTs can be part of the programme theory.

Textbox 17 – Example of a MRT in the telemonitoring COVID-19 project

In the telemonitoring in COVID-19 project, the programme theories were deduced from MRTs about self-efficacy and professional motivation. More specifically, when clinicians responsible for monitoring are provided dedicated time, clear protocols, and decision-making authority, they develop clinical ownership over the monitoring process. This ownership drives proactive patient outreach, confident escalation when needed, and sustained programme engagement over time.

On the contrary, when monitoring is added on top of regular hospital duties without dedicated resources, the mechanism of professional ownership fails to fire, resulting in reactive rather than proactive monitoring.

6.3.2 Translating realist programme theory into CMOC hypotheses

Realist programme theories are general and should be translated into specific, testable propositions for a specific intervention with specific measurable outcomes. These are called **Context-Mechanism-Outcome-Configurations (CMOC) hypotheses**.

For a complex intervention several MRTs can be relevant from which a programme theory can be generated.^{265, 266, 274, 282} Once the programme theory is clearly defined, several CMOCs can be deduced. The KCE research team's task is to identify the hypotheses (CMOCs) that have the highest policy relevance and are most feasible to investigate.

The standard formula for formulating a realist hypothesis or CMOC is:

We hypothesise that in [X context], [Y mechanism] will cause [Z outcome].^{265, 284}

Not all CMOCs are equally suited to be tested within a single evaluation study. The KCE team will prioritise the CMOCs in collaboration with the commissioner of the evaluation study. Which CMOCs will be part of the evaluation will depend on the policy questions as well as the feasibility given the available resources.

The CMOC hypotheses are more specific than the programme theory in three ways: they are specific to the intervention, they identify the contextual conditions, and they point to measurable outcomes.

For each CMOC hypothesis, the KCE research team should also explore alternative causal explanations that could produce the same observed outcome pattern through a different mechanism or pathway. These are called **rival hypotheses or rival theory^P** and testing them systematically is what characterises rigorous realist evaluation.²⁹² A rival theory asks: if we observe this outcome pattern, is our proposed mechanism the most plausible explanation, or could something else account for it equally well?

Rival theories take different forms.^{259, 260, 265}

^P The terminology is often used interchangeably.

- The outcome may have been produced by a **different contextual condition** rather than the hypothesised mechanism (see Textbox 18);
- The same outcome may have been produced by a **different mechanism** entirely (see Figure 33);
- The observed outcome pattern may reflect a **measurement artefact** rather than a genuine causal effect (see Textbox 18).

These rival theories are built into the data collection design from the outset, instead of being introduced retrospectively when findings are unexpected. In practice, this might mean, for instance, that realist interviewing guides (see Textbox 20) explicitly probe the rival explanations, and that contextual data are collected specifically to enable discrimination between the primary hypothesis and its rivals, whether or not through interviews.

Confirming a CMOC goes beyond finding supporting evidence; it requires proving this evidence while actively eliminating other plausible explanations (i.e. rival theories) as well. This dual requirement is what gives realist evaluation its explanatory rigor and distinguishes it from evaluation approaches that confirm their hypotheses simply by finding supporting quotes or correlations. Textbox 18 provides an example of a CMOC and rival theories from the telemonitoring in COVID-19 project.

Textbox 18 – Example of CMOCs and rival theories from the telemonitoring COVID-19 project

General formulation CMOC: *We hypothesise that in [X context], [Y mechanism] will cause [Z outcome].*

- **CMOC:** We hypothesise that in contexts where the telemonitoring teams are exempted from their regular hospital duties and have decision-making authority (Context), their clinical ownership of the monitoring process (Mechanism), will lead to higher patient inclusion rates (Outcome).
- **Rival theory (different contextual condition):** Higher patient inclusion rates (Outcome) may relate to differences in case-mix rather than team effectiveness.
- **Rival theory (measurement artefact):** Higher recorded inclusion rates may reflect more complete administrative registration at some sites rather than a true difference in patient inclusion.

6.3.3 Design the evaluation protocol

When the programme theory is established and the CMOC hypotheses and rival theories are defined, the realist evaluation proceeds with the design of the evaluation protocol specifying the sampling (see 6.3.4), data collection (see 6.3.5), realist analysis (see 6.3.6) and synthesis (see 6.3.7), as is described in the following subsections. The evaluation protocol should additionally meet criteria of iterativity (remaining open to revision as programme theory is refined), theoretical purposiveness (ensuring that sampling and methods are driven by the CMO hypotheses rather than convenience), and methodological adequacy (selecting methods capable of generating data on both mechanisms and context). The RAMESES II project has developed **quality criteria** (e.g. transparency of programme theory development, rigour in data collection and analysis, adequate engagement with stakeholders, etc.) and **reporting standards** (e.g. report details of participants, report how programme theory was (further) developed, etc.) for realist evaluation which can be used by the KCE research team when drafting an evaluation report.²⁶⁰

Realist evaluation is method-neutral, which means it does not impose the use of particular methods.²⁶⁵ Therefore, the wording 'realist approach' is used in this process note. The choice of the method is guided by the type of data needed to answer the evaluation question and test the initial programme theory.²⁷⁸ Both quantitative and qualitative data can be useful, often with quantitative data being focused on context and outcomes and qualitative data on generative mechanisms.²⁶⁵



6.3.4 Purposive sampling

In realist evaluation, it is important to include the variation that naturally occurs across implementation sites, populations, and time periods in the study.^{265, 282} The logic behind this is that the same policy might produce different outcomes in different contexts because mechanisms can be triggered differently depending on the contextual conditions across sites.^{265, 277} Variation is therefore considered as an important analytical resource, not a confound to be controlled for.

The sampling method used in realist evaluation must allow to test specific CMOC hypotheses. As such, a site needs to be selected, for instance, because it enables the testing of a specific hypothesis. It is the KCE research team's task to explain why outcomes differ by identifying which mechanisms fired (or did not fire) in which contexts. In this way, the size of a sample is less important than how cases are utilised to interpret and explain research findings.²⁹³

In realist evaluation, sampling is therefore **purposive**, i.e. cases should enable testing of the initial programme theory in all its dimensions.²⁷⁸ So it should represent a *range of different settings* (for example selecting high-performing versus low-performing implementations), varied contextual conditions (e.g. rural vs urban, well-staffed vs. under-staffed, different population demographics, difference in financing), different health service delivery models or degrees of local adaptation, different stages of implementation maturity, etc. (see Textbox 19). The KCE research team could for example select a purposeful sample based on secondary data (e.g. administrative or survey data) about contextual conditions or settings.

Theoretical sampling is also commonly used in realist evaluation. It's a kind of purposive sampling based on theoretical arguments.²⁶ For more information on theoretical and other sampling strategies, we refer to Nick Emmel's book (2013).²⁹³

Textbox 19 – Sampling in the telemonitoring COVID-19 project

To test the CMOC and rival theories, specific sites can be selected – for example sites with dedicated telemonitoring FTE versus sites where telemonitoring is added to regular hospital duties.

6.3.5 Data collection

Realist evaluation data collection is often organised, using **mixed-methods**, around different data topics to enable the testing of CMOC hypotheses. Data from different topics are collected simultaneously, not sequentially.^{260, 265, 278, 282}

For KCE purposes, these data topics are most relevant during the evaluation phase of the MRC model, less in the development and pilot/feasibility phase. Implementation will not be common in the KCE evaluation programme, but in general the topics also apply to that phase.

The most frequent and informing topics in realist evaluation are about:

- Data that provide information on **implementation fidelity**.^{265, 294} This concerns data to answer questions such as 'What service/intervention is actually being delivered, and how does it differ from the intended design?' It captures both the formal and informal components of service delivery, including real-time adaptations. Sources to gather data on implementation fidelity include fidelity assessments, structured observation, document analysis of local protocols and guidance, and activity logs and service records.
- Data that provide information on **actor reasoning**.^{265, 266} This concerns data to understand how people reason and respond to what the policy or intervention offers them and to detect the mechanism at work. It is the reasoning of actors interpreting and responding to the components of the intervention programme in a given context. Informants are selected for their capacity to provide insights into specific operational contexts and causal mechanisms. This can be done during realist interviews (see Textbox 20), focus groups, etc.

- Data that provide information on **the context**.^{260, 265, 282} Outcomes may vary depending on economic, geographic, historical, social and political circumstances and the cultural values of participants. Variations within the intervention programme's target group can also influence which mechanisms operate. These data are collected to understand how, and for whom, an intervention programme works. Sources to gather data on the context may include routine administrative data (e.g. staffing levels, population demographics), organisational data (e.g. governance structures), policy data (e.g. funding streams, co-existing policies), and historical or cultural data (e.g. prior attitudes, trust in services).
- Data that provide information on **the outcomes** (disaggregated).^{260, 265, 282} In a realist approach, within the same programme, sites or groups are compared and set out against each other. For this reason, no control group is required and data are collected disaggregated (i.e. not aggregated into a single mean effect) by context and mechanism to test whether outcome patterns match the CMOC hypotheses.

Textbox 20 – Realist interviewing

A data collection method that is specific to realist evaluation and differs from interviewing techniques in other KCE projects is realist interviewing. It is a specific kind of in-depth and a theory-driven approach to interviewing.²⁹⁵ As explained by Pawson and Tilley and further developed by Manzano, the use of traditional interviewing techniques is not sufficient to explain what works, for whom, and under what circumstances. Therefore, the purpose of realist interviewing is to test, confirm, refute, or refine a specific CMOC hypothesis. The technique was proposed by Pawson (1996) *'to create a situation in which the theoretical postulates/conceptual structures under investigation are open for inspection in a way that allows the respondent to make an informed and critical account of them'*.^{296, 297} Thus, the interviewer arrives with a theory and uses the interview to interrogate it.

Consequently, a realist interview guide is not a topic list as in traditional interviewing techniques. It is a structured theory-testing conversation, built around the CMOC hypotheses, with probes designed to surface contextual conditions, reasoning processes, and unexpected pathways.^{270, 278, 297}

- Why something did (not) work? (not only whether it worked)
- Presenting tentative theories to interviewees and inviting confirmation, refutation, or refinement
- Probing for contextual condition that enabled or blocked engagement
- Exploring unintended consequences and unexpected causal pathways.

When implementation theories such as Normalization Process Theory (NPT) have been used to develop the programme theory (see section 6.3.1)^{289, 290}, their constructs can serve as a concrete and reusable source of testable hypotheses to present to interviewees, making the theory-testing conversation more structured and specific. The interviewer may probe whether practitioners developed a shared understanding of the intervention (coherence), whether they personally committed to it (cognitive participation), what practical work it required (collective action), and how problems were noticed and addressed (reflexive monitoring); each representing a specific, testable mechanism hypothesis. These NPT constructs may also structure the analysis of the data. A coding manual explicitly linking NPT constructs to CMOCs was developed by May et al. (2022).²⁹¹

Key participants to realist interviews are typically frontline practitioners, commissioners, beneficiaries, etc.

Apart from conducting these 'realist interviews' in the theory refining (i.e. evaluation) phase for data collection, interviews can also be conducted to provide input for the initial programme theory (i.e. initial exploratory interviews to map out basic ideas) and for theory consolidation (i.e. final interviews



conducted during synthesis to understand the structural barriers and specific contextual boundaries).²⁹⁵

For more information on how to carry out a realist interview, we refer to Manzano (2016)²⁹⁵ and Mukumbang et al. (2020)²⁹⁷.

6.3.6 Realist data analysis – Retroduction

The main analytical approach in realist evaluation is **retroduction**, i.e. reasoning backwards from observed outcome patterns to the underlying mechanisms that must have produced them.^{265, 278} Retroduction uses both inductive and deductive reasoning. In addition, it is combined with researcher's theoretical knowledge and insight, to identify the hidden causal forces lying behind identified patterns. The central question is: *given this outcome pattern, what mechanisms must have been operating or absent to produce it, and what contextual conditions explain why they fired here but not there?*

A practical analytical tool is a CMOC matrix, i.e. a table with columns for Context, Mechanism, Outcome, and Evidence (confirming / refuting / refining), with one row per hypothesis.^{266, 270} This keeps the analysis anchored in the programme theory and prevents drifting off into generic thematic analysis.

Table 31 provides an example of a CMOC matrix applied to the telemonitoring COVID-19 project. It is a translation of the hypotheses presented in Textbox 15 in the format of a CMOC matrix.

Table 31 – An example of a CMOC matrix

Hypothesis	C	M	O	Evidence status	Evidence source
H(x)	Dedicated FTE, GP collaboration, clear inclusion protocols	Clinical ownership	Proactive outreach Higher inclusion	Confirmed where present; suppressed elsewhere	Team interviews
H(x) rival theory (a different mechanism)	Poor calibration, manual entry errors	Alarm fatigue	Reduced team responsiveness	Supported	Alarm reports, team interviews

6.3.7 Synthesis

Following the analysis, the refined programme theories are synthesised across cases or sites to adapt MRTs. Gilmore et al. (2019) describe this as a distinct analytical phase within realist evaluation, not to be confused with realist synthesis (also called realist review) as a separate review methodology.^{278, 298}

The synthesis following the analysis identifies demi-regularities in which the programme theories and CMOCs are brought together and common patterns are identified and grouped. During this synthesis process the KCE research team will have to pay special attention to outcome patterns that occur across most but not all cases (i.e. demi-regularities) which are the analytical core of the synthesis. During the synthesis, the KCE research team may consult stakeholders who have a system-level view of the intervention to explain discrepancies between cases. It is also an option to compare the generated theory with relevant formal theories from the literature as a plausibility check. Situating findings within the broader theoretical context can help to further refine the explanatory mechanism(s).

Ultimately, the KCE research team must identify how and under what conditions change occurs. This should result in a transferable, policy-relevant theory that can inform implementation in new contexts.

6.3.8 Output of the realist evaluation

Realist evaluation can be applied within the KCE evaluation programme in different ways. The expected output will differ depending on the study type and the phase of the MRC framework in which the realist approach is embedded. When the realist approach is:

- **applied alongside an impact evaluation**, e.g. embedded within a study that also assesses clinical or cost-effectiveness, it helps to explain why the observed outcomes occurred. It does not replace the effectiveness evaluation but complements it by opening the 'black box' by identifying which mechanisms fired, under what contextual conditions, and for whom. The output in this scenario is primarily explanatory and is directed at understanding the conditions under which the intervention produced (or failed to produce) its effects. This output can subsequently inform how the intervention should be rolled out or adapted when moving from the evaluation to the implementation phase.
- **the primarily methodology in an implementation study**, e.g. evaluating the implementation of an intervention programme, the output is directly actionable for implementation steering. In this scenario, the realist evaluation will help to refine the programme theory. This will give insight in which contextual conditions are necessary for the intervention to work, which elements of the implementation are necessary, and where adaptations are required. The main question to answer is: how do we make it work here, for these patients, in these organisational conditions?

Realist evaluation provides an evidence base for transferability, i.e. guidance on which causal configurations are likely to hold in new contexts and which are context-specific. This is relevant to inform the implementation phase but also to inform scale-up to new settings or populations.

Therefore, in the final section of the realist approach cycle the KCE research team could answer:

- What does this refined programme theory tell policymakers/commissioners about how to implement this intervention in a new context?
- Which contextual conditions are necessary for key mechanisms to fire?
- Which aspects of implementation are non-negotiable, and which can be adapted?

Generally, the final conclusion does not provide a straightforward answer to whether an intervention 'works'. Rather, it identifies the conditions under which the intervention is likely to be effective. In response to the question of whether the intervention should be scaled up, the answer might be that the intervention may reduce strain under the following conditions. This should be aligned with the needs of the commissioners. For future research, a research framework can be built around the CMOCs to enable prospective testing of the refined hypotheses, with standardised data collection across sites. Of note, when seeking to primarily determine average effect size of a more simple intervention (i.e. an intervention administered in a limited range of conditions) an RCT or systematic review may be more appropriate.²⁶⁰ Consider that realist evaluation is resource-intensive in terms of time, expertise, and iterative stakeholder engagement, and may therefore be disproportionate for evaluations with a narrow scope or short timeframe. The decision to apply a realist approach should therefore be deliberate and justified in relation to the evaluation question, the complexity of the intervention, and the type of evidence needed by the commissioners.

Textbox 21 – Example of a plausible output from the telemonitoring COVID-19 project

Reporting by CMOC

- Professional ownership fires when teams have dedicated FTE, clear protocols, decision-making authority, and GP collaboration. It is systematically suppressed when monitoring is added to regular hospital duties without resource allocation. The convention's assumption of



GP involvement was structurally undermined by lack of data integration, GP workload, and unclear role definitions. Primary care partnership requires formal operational agreements.

CMOC refinement

- If the mechanism and contextual explanation hold and the rival is ruled out, the initial hypothesis is refined: the professional ownership mechanism requires not just physician leadership but dedicated coordination time and an established patient pathway, conditions absent in Site A and present in Site B.

Output to the commissioner with regard to whether telemonitoring should be scaled up, may be:

- Telemonitoring may reduce the strain on the hospital workforce and primary care providers under the following conditions: dedicated staffing, active two-way communications, GP and home nurse involvement.

7 EVALUATION OF ECONOMIC IMPACT

KEY POINTS

- It is important to integrate economic considerations into the assessment of healthcare interventions because all interventions use resources and therefore have economic consequences. This process note covers **economic evaluation, budget impact analysis (BIA) and costing**.
- Economic evaluations can consider both health and non-health outcomes. Within this evaluation programme, the **focus will often be on non-health outcomes** that may be relevant to the decision problem, such as access, capacity, productivity, quality of care or organisational performance.
- The **choice of economic evaluation** method should reflect the nature of the outcomes and the evidence available. For guidance on conducting a cost-utility analysis (CUA) or a cost-effectiveness analysis (CEA) we refer to KCE report 400. These require an effect evaluation on health outcomes. In this process note, we give methodological guidance for the following types of economic evaluation:
 - cost-minimisation analysis (CMA) requiring demonstrated equivalence of relevant outcomes;
 - cost-consequence analysis (CCA) presenting multiple costs and consequences separately; and
 - cost-benefit analysis (CBA) expressing both costs and consequences in monetary terms and to be used when these can be valued credibly.
- **A BIA estimates the financial consequences of introducing an intervention for a specified budget holder.** It compares expenditure under the current and future treatment mix, taking into account the target population, expected uptake and diffusion, changes in resource use and costs, and the timing of these changes.
- **Costing identifies, measures and values the resources required to implement and/or deliver an intervention.** It can be part of an economic evaluation or BIA or undertaken as a stand-alone study; the costing approach should be determined by the purpose of the analysis and its perspective.
- **Resources** should be identified and measured in appropriate natural units and valued using relevant unit costs. Implementation costs and routine delivery costs should be distinguished where relevant.
- The **costing method** should provide sufficient precision for the decision problem. Top-down or bottom-up and gross-costing or micro-costing approaches can be used alone or in combination, with more detailed approaches applied where resource use or cost drivers are particularly important.

7.1 Introduction

Economic considerations are relevant to the assessment of healthcare interventions because they require resources and hence represent an economic cost. Economic evidence should inform decision-makers about the value of an intervention relative to the resources required and about the financial implications of its implementation. Taking these resource implications into account is particularly important when healthcare resources are limited and decisions need to balance expected health benefits against the resources required to achieve them.



Healthcare interventions may consume a wide range of resources, including healthcare professionals' and informal caregivers' time, medicines, medical equipment, infrastructure and hospital capacity. They may also affect costs borne directly by patients, such as out-of-pocket payments and time spent receiving care.

A distinction should be made between economic evaluation and the assessment of economic impact. An **economic evaluation** compares the costs and consequences of two or more alternatives and requires an estimate of the incremental effect of the intervention relative to an appropriate comparator.²⁹⁹ This effect may be estimated using a randomised controlled trial (RCT) or, when randomisation is not feasible, using an appropriate quasi-experimental or observational design (see Chapter 3).

Effectiveness, implementation and costs are closely related but represent distinct dimensions of an intervention. Effectiveness concerns whether an intervention produces the intended outcomes, whereas implementation concerns how the intervention is delivered in practice, including its uptake, fidelity, reach, feasibility and the resources required for delivery (see Chapter 8). Costs represent the value assigned to the resources consumed in implementing the intervention and delivering care. Resource use can encompass healthcare professionals' and patients' time, medicines, medical equipment, infrastructure and hospital capacity. It is important to note that an intervention may have different resource requirements or effects in routine practice than under controlled study conditions. For example, implementation may require more staff time, training or infrastructure than initially anticipated. Understanding implementation can consequently provide important information for the economic assessment by identifying resource requirements, cost drivers and differences in implementation between settings.

This process note does not provide a comprehensive methodological framework for conducting full economic evaluations in which costs are compared with clinical outcomes, such as cost-effectiveness or cost-utility analysis. For methodological guidance on these analyses, we refer to KCE Report 400.²⁹⁹ Instead, the process note focuses on **the economic assessment of interventions where the outcomes of interest include non-health outcomes, as well as on budget impact analysis (BIA) and the identification, measurement and valuation of resource use and costs (costing)**. This reflects the nature of the evaluation programme, in which a focus on clinical effectiveness alone (such as is typically done in the evaluation of medicines and health technologies) will in most cases not be available or feasible, making other approaches to assessing economic impact more relevant.

These approaches address different but complementary questions. Full economic evaluation assesses whether the additional resources required for an intervention are justified by its additional health benefits, typically expressed in terms of costs per QALY (quality-adjusted life year) or other health outcomes. Economic analyses including non-health outcomes may instead assess outcomes such as capacity, productivity, access or organisational effects. BIA estimates the financial consequences of adopting an intervention for a specified budget holder over a defined time horizon, while costing identifies, measures and values the resources consumed by an intervention, service or patient group. Costing may form part of any of these analyses but is not, in itself, an economic evaluation.

These distinctions are particularly relevant in Belgium, where healthcare financing is distributed across different sectors and levels of government and where patients may bear part of healthcare costs through co-payments and other out-of-pocket payments. An intervention may, for example, be cost-effective from a societal or healthcare-system perspective while generating substantial additional expenditure for a particular budget holder in the short term. Conversely, it may reduce expenditure for one budget holder while increasing costs elsewhere in the healthcare system. The appropriate analytical approach should therefore be determined by the decision question, the outcome of interest and the perspective of the analysis.

7.2 Economic evaluation when effectiveness can be assessed

An economic evaluation compares the costs and consequences of at least two alternative interventions. It requires:

- a relevant intervention and comparator; and
- an estimate of the relative effect of the intervention compared with the relevant comparator.

The relative effect can be estimated using a randomised controlled trial (RCT) or, where randomisation is not feasible or appropriate, using other study designs combined with appropriate methods for causal inference (see Chapters 2 and 3). Depending on the research question, intervention and available data, methods such as difference-in-differences, interrupted time-series analysis, regression discontinuity, instrumental variables, matching or propensity-score methods may be appropriate. The choice of method should be guided by the decision problem, how the intervention is introduced and delivered in practice (e.g. timing, who receives it, whether there is a comparison group), the available data and the assumptions required to identify a causal treatment effect.

Evidence on comparative effects can then be combined with information on resource use and costs to estimate the incremental costs and consequences of the intervention compared with the relevant alternative.

This section briefly introduces the two main types of economic evaluation that require a health outcome: cost-utility analysis (CUA) and cost-effectiveness analysis (CEA). A CUA compares the costs and health outcomes of interventions using quality-adjusted life years (QALYs), enabling comparisons across different diseases and clinical areas. A CEA uses a more specific outcome measure, such as life-years gained (LYG), and may be presented alongside CUA when important survival benefits are not fully captured by QALYs. More detailed methodological guidance on these approaches is provided in KCE Report 400, which defines the Belgian reference case for health economic evaluations in the context of reimbursement decisions. The treatment of these methods is deliberately brief in this process note because, within the evaluation programme, a measurable health outcome will in most cases not be available or feasible to assess. Greater attention is therefore given to other approaches to economic assessment that are more relevant to the types of evaluations expected within the programme.

7.2.1 Economic evaluation with non-health outcomes

The outcomes considered in an economic evaluation may be health outcomes or non-health outcomes, depending on the decision problem. While health outcomes may be relevant for some evaluations, they will in most cases not be available or feasible to measure within the evaluation programme. This section therefore focuses on economic evaluations that include non-health outcomes, such as changes in capacity, access, productivity, resource use or organisational performance, while health outcomes may also be considered where relevant.

7.2.1.1 Cost-minimisation analysis

Cost-minimisation analysis (CMA) compares the costs of two or more alternatives when their relevant consequences can be considered equivalent. The analysis therefore focuses on identifying the least costly alternative. CMA should only be used when there is sufficient evidence that the alternatives have equivalent effects on the outcomes relevant to the decision. Importantly, CMA should not be used simply because differences in outcomes have not been measured: absence of evidence of a difference is not evidence of equivalent effects. Where equivalence has not been established, cost-consequence analysis (CCA) (see section 7.2.1.2) or, where a single relevant outcome can be defined and measured, CEA may be appropriate. Within the evaluation programme, CMA may therefore be appropriate when the relevant outcomes of the alternatives can be shown to be equivalent. A CMA should, however, not be used as a substitute for assessing differences in outcomes when these have not been adequately established.



For example, suppose an evaluation compares two organisational models for delivering home nursing care. If there is strong evidence that both models provide equivalent access to care, quality of care and other outcomes relevant to the decision, a CMA could be used to determine which model requires fewer resources to implement and operate. If, however, one model is associated with differences in access, quality, staff capacity or other relevant outcomes, these differences should be incorporated into the economic analysis (e.g. a CCA) rather than assuming equivalence.

7.2.1.2 Cost-consequence analysis

Cost-consequence analysis (CCA) presents the costs and relevant consequences of an intervention separately, without combining them into a single summary measure such as an incremental cost-effectiveness ratio. It is particularly appropriate when an intervention has multiple relevant outcomes that cannot or should not be combined into a single measure, including non-health outcomes such as access, capacity, productivity, quality of care, organisational performance or patient experience. CCA allows these different consequences to be presented alongside the associated costs, leaving their relative importance to be considered as part of the decision-making process.

CCA is particularly relevant within the evaluation programme because interventions may affect several dimensions of performance simultaneously, and these effects may not be adequately aggregated into a single outcome measure. For example, an intervention may increase costs but improve access and quality of care while reducing staff workload. Presenting these costs and consequences separately makes such trade-offs explicit rather than requiring them to be reduced to a single metric.

CCA can be used whether the consequences are health or non-health outcomes. The outcomes selected should reflect the decision problem and should be specified in advance where possible. When consequences move in different directions, the analysis should make these trade-offs transparent and provide sufficient information for decision-makers to assess their relative importance. When all relevant consequences move in the same direction, the interpretation may be more straightforward, but the same principle of presenting costs and consequences separately applies.

7.2.1.3 Cost-benefit analysis

In a cost-benefit analysis (CBA), both the costs and the consequences of an intervention are expressed in monetary terms. This allows the benefits of an intervention to be compared directly with its costs and, in principle, enables an assessment of whether the intervention generates a positive net monetary benefit. CBA may be particularly relevant within the evaluation programme when an intervention has consequences that extend beyond health outcomes or when the evaluation requires a broad perspective. For example, an intervention may affect healthcare expenditure, patients' out-of-pocket costs, productivity, informal care and other consequences. Expressing these consequences in monetary terms allows them to be considered within a common analytical framework.

However, assigning monetary values to health and other outcomes can be methodologically challenging. The valuation of clinical, health and non-health outcomes in monetary terms has long been debated, and approaches such as willingness-to-pay can produce estimates that depend substantially on the methods used and on the populations considered. CBA should therefore be used with caution, particularly where the monetary valuation of important outcomes is subject to substantial uncertainty.

Within the evaluation programme, CBA may be used as a stand-alone approach when the decision problem can be adequately addressed by expressing the relevant costs and consequences in monetary terms. It may also be combined with other economic analyses where this provides additional information, for example when some relevant consequences can be monetised while others are more appropriately presented in their natural units. The choice of CBA, alone or in combination with another approach, should therefore be guided by the decision problem, the relevant outcomes and the feasibility and validity of their monetary valuation.

Textbox 22 – An example comparing CMA, CCA and CBA

For example, suppose an evaluation compares two organisational models for delivering home nursing care.

Cost-minimisation analysis (CMA)

If there is strong evidence that both models provide equivalent access to care, quality of care and other outcomes relevant to the decision, a CMA could be used to determine which model requires fewer resources to implement and operate. If, however, one model is associated with differences in access, quality, staff capacity or other relevant outcomes, these differences should be incorporated into the economic analysis (e.g. a CCA) rather than assuming equivalence.

Cost-consequence analysis (CCA)

If the models differ in their effects on access to care, quality of care, staff capacity or other outcomes relevant to the decision, a CCA could be used to present these consequences alongside the costs of each model. For example, one model may require more resources but improve access and reduce staff workload, while the other may be less costly but provide lower access. Presenting the costs and consequences separately allows decision-makers to assess these trade-offs without requiring the different outcomes to be combined into a single measure.

Cost-benefit analysis (CBA)

If the relevant consequences of the two models can be credibly expressed in monetary terms, a CBA could be used to compare their total costs with their monetised benefits. For example, one model may require higher expenditure but generate monetary benefits through improved productivity, reduced staff workload, fewer hospitalisations, lower patients' out-of-pocket costs or reduced informal care needs. These costs and benefits can be valued in monetary terms and combined to estimate the net monetary benefit of each model. If some important consequences cannot be credibly monetised, these could be presented separately alongside the CBA results.

7.2.2 *Key principles of economic evaluation*

The **target population** should be clearly defined and should reflect the population, organisations or other units for which the intervention is intended to be used or implemented in routine practice. The relevant unit of analysis and characteristics will depend on the type of intervention and the evaluation question. These may include demographic characteristics, disease or risk profile, treatment setting, previous treatment, organisational characteristics, staffing levels, level of service provision, geographical characteristics, or other factors that may influence the outcomes, resource use, costs or baseline conditions relevant to the evaluation. Belgian data should be provided where available.

The target population considered in the evaluation may be broader than the population directly represented in the available evidence, provided that extrapolation to the broader population is sufficiently justified. The assumptions underlying such extrapolation should be clearly stated, and any associated uncertainty should be explored where relevant.

Relevant subgroups should be identified based on the decision problem and the outcomes considered in the evaluation. Depending on the intervention, relevant subgroup characteristics may relate to differences in health status, access to care, service needs, organisational context, baseline risk, resource use, costs, implementation conditions or other factors that may affect the outcomes of interest. Subgroups should, where possible, be specified in advance, particularly when there are plausible reasons to expect differences between them.

Subgroup-specific analyses should be conducted where differences between subgroups are relevant to the decision and are sufficiently supported by the available evidence. Where evidence is insufficient to establish meaningful differences between subgroups, the analysis should avoid introducing unsupported assumptions. Relevant differences in baseline conditions, resource use, costs or other



parameters may nevertheless be incorporated where supported by the evidence. Post-hoc subgroup analyses that were not prespecified should be clearly identified as exploratory and interpreted with appropriate caution.

For example, an evaluation of a new hospital-based intervention should clearly define the population or hospital services for which the intervention is intended in routine practice, taking into account characteristics relevant to the evaluation question. These may include age, disease severity and admission type, but may also include organisational characteristics such as hospital size, staffing levels or the availability of specific services. If the evaluation concerns access to care rather than clinical effectiveness, other characteristics may be more relevant, such as geographical location, socioeconomic circumstances or distance to healthcare services. Subgroup-specific analyses should then be undertaken where differences between these groups are relevant to the decision and sufficiently supported by the available evidence.

Cost calculation is an essential component of an economic evaluation. It involves identifying, measuring and valuing the resources consumed by the intervention and its comparator(s), with a focus on the incremental resource use and costs. Future costs and outcomes are **discounted**. The methods used to identify, measure and value resource use and costs are described in section 7.4.

The **perspective** of the analysis determines whose costs and consequences are considered relevant (see section 7.4.2). Depending on the decision problem, this may be the perspective of the healthcare payer, healthcare provider or society. The perspective should be specified in advance, as it determines the scope of the costs and consequences included in the analysis.

The **time horizon** should be specified in advance and should be sufficiently long to capture all relevant differences in costs and consequences between the intervention and the comparator. Its duration should therefore reflect the period over which the intervention is expected to affect relevant costs and outcomes. When the appropriate time horizon extends beyond the period for which empirical evidence is available or can be collected, longer-term costs and consequences may need to be estimated through modelling or extrapolation. The assumptions underlying such extrapolation should be clearly described and explored in sensitivity or scenario analyses. Where relevant, a shorter time horizon corresponding to the period for which empirical evidence is available may be considered as a scenario, particularly when there is substantial uncertainty about longer-term effects. The same time horizon should be applied to costs and consequences.

Economic evaluations are usually based on evidence and assumptions. The potential impact of **uncertainty** should be explicitly assessed and reported, distinguishing between uncertainty about the values used in the analysis (parameter uncertainty), uncertainty arising from the methods or assumptions used to conduct the analysis (structural and methodological uncertainty) and contextual uncertainty.

Parameter uncertainty concerns uncertainty about the values of individual inputs, such as intervention effects, costs, probabilities, resource use, utilities or other outcome measures. Where relevant, parameter uncertainty should be assessed using probabilistic sensitivity analysis (PSA). In PSA, uncertain inputs are allowed to vary simultaneously according to appropriate probability distributions to examine the resulting uncertainty in the outcomes of the analysis. Deterministic sensitivity analyses may also be used to assess the influence of individual parameters or specific assumptions.

Structural and methodological uncertainty concerns uncertainty arising from the structure of the analysis or from methodological choices and assumptions. Examples include alternative model structures, assumptions about relationships between variables, the choice of perspective, costing methods, time horizon or methods used to extrapolate evidence beyond the observed data. Where relevant, alternative assumptions or analytical approaches should be explored through sensitivity or scenario analyses.

Contextual uncertainty concerns uncertainty about the circumstances or wider environment in which an intervention is implemented and in which its effects, resource use or costs may evolve over time.

This may include changes in demographics, epidemiology, technology, clinical or organisational practice, healthcare policy, financing arrangements or other relevant external factors. It may also arise when results need to be extrapolated from the population, setting or period covered by the available evidence to the population, setting or period relevant to the evaluation.

More detailed methodological guidance is provided in KCE Report 400. Table 1 of that report summarises the 16 guidelines that define the reference case for economic evaluations in Belgium.²⁹⁹

7.2.3 *Economic evaluation of complex interventions*

Baghbanian et al. highlight several methodological considerations that are particularly relevant to the economic evaluation of complex interventions.³⁰⁰ Economic evaluation of complex interventions can present additional methodological challenges compared with interventions with more clearly defined and predictable causal pathways. Complex interventions may involve multiple interacting components, stakeholders and contextual factors, with effects that can emerge through dynamic and potentially non-linear pathways. These features can increase uncertainty around both costs and outcomes and make it more difficult to estimate the incremental costs and effects of the intervention with precision. The authors note that complex interventions and the conditions they address may involve high levels of uncertainty, partly because of their interconnected components, diverse stakeholders and contextual dependence.

A particularly important consideration is the **context in which the intervention is implemented**. Complex interventions are described as multi-component and context-sensitive, with initial conditions and contextual variables potentially influencing their effects. From an economic evaluation perspective, this implies that both resource use and outcomes may vary between settings. Factors such as workforce availability, organisational capacity, infrastructure and the way care is organised may influence the resources required to deliver the intervention and the outcomes achieved. Consequently, the transferability of cost-effectiveness estimates between settings may be more limited, and the assumptions used to transfer evidence to the decision-making context should be made explicit and tested where appropriate.

A further consideration is the need to understand the **mechanisms and pathways through which the intervention generates its effects** (see Chapter 1). For economic evaluation, understanding these pathways can inform the identification of relevant costs and outcomes, the selection of intermediate and final outcomes, and the structure and time horizon of an economic model. The economic evaluation should ensure that the relevant causal pathways linking the intervention to resource use and outcomes are adequately reflected in the analysis.

Implementation and sustainability may likewise have direct economic implications. Factors such as resource availability, implementation feasibility and healthcare-system capacity are potentially important determinants of whether a complex intervention can be implemented effectively. These factors may affect the resources required to introduce and maintain the intervention and therefore its longer-term costs and effects. Where relevant, an economic evaluation should consequently consider implementation-related costs, changes in resource use during implementation, and uncertainty about the extent to which the intervention can be delivered as intended and sustained over time.

Overall, **the principles of economic evaluation do not fundamentally differ between complex and non-complex interventions**. Rather, complex interventions may require greater flexibility in how those principles are applied. The methodological literature nevertheless increasingly argues for approaches that better recognise the interactions between the intervention and the system in which it operates.



7.3 Budget impact analysis

The main **output** of a budget impact analysis (BIA) is the change in expenditure associated with introducing an intervention, generally reported separately for each year of the relevant time horizon and, where feasible, until the budget impact reaches a 'steady state'. Steady state refers to the point at which the annual budget impact has stabilised, taking into account the ongoing entry and exit of patients and any longer-term costs or savings associated with the intervention.

The analysis aims to estimate the financial consequences of adopting the intervention for the relevant budget holder, taking into account the expected treatment mix and uptake of the intervention. Whereas a full economic evaluation assesses the value for money of alternative interventions, a BIA addresses **affordability: what additional expenditure or savings can be expected for a specified budget holder when an intervention is introduced?** BIA and economic evaluation share several methodological principles, including those concerning perspective, costing, modelling and the handling of uncertainty, but differ in their research question and in how the target population, comparator, time horizon and results are defined. Table 1 in Sullivan et al. (2014), which sets out principles of good practice for budget impact analyses, identifies key considerations for the design of a BIA.³⁰¹ These include features of the health care system, perspective, use and cost of current and new interventions, impact on other costs, time horizon, time dependencies and discounting, choice of computing framework, uncertainty and scenario analysis, and validation.

A BIA **compares the expected expenditure under the current situation with the expected expenditure after introduction of the intervention**. The analysis should therefore consider not only the cost of the new intervention, but also the costs it replaces and changes in other healthcare resource use resulting from its introduction. Where relevant, savings from avoided treatments, hospitalisations, adverse events or follow-up care should be included.³⁰² The comparator in a BIA may therefore differ from that used in an economic evaluation. While an economic evaluation compares the incremental costs and health consequences of alternative interventions relative to an appropriate comparator, a BIA is intended to estimate the financial consequences of introducing an intervention within the current healthcare system. It therefore considers the current treatment mix and the expected changes in treatment mix following adoption of the intervention. For example, if a new intervention (e.g. a hospital-at-home programme for selected patients with pneumonia) substitutes for an existing treatment (e.g. conventional inpatient hospital care), the BIA should include both the cost of the new intervention and the savings from avoided inpatient hospital care. If some patients continue to receive conventional inpatient care while others receive hospital-at-home care, the analysis should reflect the expected treatment mix and uptake of the new intervention. The budget impact will therefore depend on both the costs of the alternative treatment pathways and the proportion of patients expected to receive each pathway.

The **target population** should reflect the population eligible for and expected to receive the intervention in the relevant healthcare system. In contrast to the closed population commonly used in economic evaluations, a BIA generally considers an **open and changing population**. Changes in the size and characteristics of the target population over time, including changes in incidence, prevalence or disease severity, should therefore be considered. The expected uptake, diffusion or market share of the intervention should also be estimated and justified, as should the extent to which it replaces existing interventions. Where relevant, different subpopulations should be considered separately if they differ in treatment patterns, costs or expected uptake. For example, if a new intervention reduces hospitalisations for patients with a particular chronic condition, the BIA should estimate not only how many patients are currently eligible for the intervention, but also how this population is expected to evolve over the budget horizon. New patients may enter the target population as they are diagnosed, while other patients may leave it because they recover, die or no longer meet the eligibility criteria. The analysis should also consider how quickly the new intervention is expected to be taken up and the extent to which it replaces existing treatment. For instance, if 20% of eligible patients are expected to switch to the new intervention in the first year, increasing to 50% in subsequent years, the BIA should reflect this gradual diffusion rather than assume that all eligible patients receive the intervention

immediately. If the intervention is expected to have different uptake or costs across relevant patient subgroups, these should be modelled separately where appropriate.

The **perspective** should correspond to the relevant budget holder. In the Belgian reference case, the BIA is conducted from the perspective of healthcare payers, including government and patients. The analysis should therefore capture the financial consequences relevant to these payers. However, other perspectives may be appropriate depending on the purpose of the analysis and the decision-maker concerned, and in many cases a BIA may need to consider more than one perspective. For example, from the perspective of a hospital, a hospital-at-home programme may reduce inpatient bed use and associated hospital costs, while from the perspective of the health insurer, it may result in additional expenditure for home-based care. Transfers between different healthcare payers or levels of government should therefore be made explicit, as an intervention may generate savings for one payer while increasing expenditure for another.

The **time horizon** should be sufficiently long to capture the relevant financial consequences of introducing the intervention and should generally extend until the budget impact reaches a steady state, where feasible. KCE Report 400 recommends a minimum time horizon of three years. The budget impact should be reported separately for each year so that changes in expenditure during the implementation and diffusion period can be distinguished from the longer-term financial consequences. Unlike a full economic evaluation, costs are not discounted in a BIA. This is consistent with the purpose of a BIA, which is to describe the expected financial consequences of adopting an intervention for a specific budget holder in each budget period, rather than to express costs occurring at different points in time as a single present value. The recommendation not to discount is therefore primarily a methodological convention reflecting the purpose and use of BIA, rather than a theoretical principle with a universally accepted justification.

A BIA may draw on **information** from an economic evaluation, clinical studies, administrative data, costing analyses and other sources (see Chapter 11). However, information from an economic evaluation may require adaptation before it can be used in a BIA. In particular, the analysis should reflect the real-world target population, expected uptake and diffusion, actual treatment mix and the timing of resource use and costs. Where an economic evaluation and a BIA are conducted for the same intervention, consistency between the analyses should be maintained where appropriate, while recognising their different objectives and potentially different comparators.

Uncertainty should be assessed for assumptions and parameters that may have an important influence on the estimated budget impact. This includes uncertainty around the target population, intervention uptake, treatment patterns, resource use and costs, as well as uncertainty about how these may evolve over time. Particular attention should be paid to structural and contextual uncertainty related to the implementation of the intervention, changes in clinical or organisational practice, technology, policy or financing arrangements. Where relevant, alternative assumptions or scenarios should be explored.

Results should also be disaggregated where relevant, for example by major cost component or healthcare payer, to show where additional expenditure and savings occur. The analysis should make transparent which costs arise from the intervention itself, which result from changes in existing healthcare expenditure, and which arise from subsequent consequences such as adverse events, follow-up treatment or re-hospitalisation.



7.4 Costing

7.4.1 Purpose and role in economic analysis

Costing concerns the identification, measurement and valuation of the resources required to deliver or implement an intervention. It provides information on resource use and associated costs but does not, by itself, assess whether an intervention provides good value for money or what its financial consequences would be for a budget holder. Cost information can subsequently be used in different forms of economic analysis, including an economic evaluation and a BIA. A costing study can also be undertaken independently, for example to estimate the resources required to introduce a new care pathway, identify the main cost drivers of an existing service, or compare the costs of delivering an intervention across organisational settings.³⁰³ Costing may also form part of a process evaluation, where information on resources used to implement and deliver an intervention helps to understand how it was implemented in practice (see Chapters 5 and 8). In such cases, the resource-use information does not necessarily constitute a formal costing study.

The same resource-use data may therefore serve different purposes. For example, suppose an intervention requires nurses to spend 30 minutes per patient.

- In a process evaluation, staff time may be measured to understand how the intervention was implemented in practice, including whether the expected amount of staff time was required and how resource use varied across settings.
- In a costing study, staff time would be systematically identified, measured and valued, together with other relevant resources, with the specific aim of estimating the total cost of implementing and delivering the intervention.
- In a BIA, the additional nursing time would be translated into the financial consequences for the relevant budget holder, taking into account how nursing care is financed, whether the additional workload can be accommodated within existing capacity, and whether additional staffing or other expenditure would be required. These financial consequences would then be combined with the expected number of eligible patients, uptake and other relevant parameters to estimate the overall budget impact.
- In a full economic evaluation, it would first be assessed whether the intervention requires additional nursing time compared with the comparator. If so, the additional 30 minutes would be valued in monetary terms and considered together with other incremental costs and savings. These incremental costs would then be assessed alongside comparative evidence on outcomes to evaluate the cost-effectiveness of the intervention.

These distinctions concern primarily the purpose and analytical use of the data rather than the type of data collected.

A separate full costing study is not necessarily required before undertaking a BIA or an economic evaluation. Where sufficiently reliable and relevant information on (incremental) resource use and unit costs is already available, these data can be incorporated directly into the analysis. Conversely, where such information is lacking or where the objective is specifically to understand the total resource use and costs associated with an intervention or service, a stand-alone costing study may be appropriate.

The objective of a costing study should be defined clearly in advance. In particular, it should be specified whether the analysis aims to estimate the costs of implementing an intervention (i.e. the implementation strategy), delivering the intervention, or both, and whether the intervention is intended to be delivered as an add-on to existing care or to replace an existing intervention or service. This distinction can be important for interventions that require substantial initial investments but have relatively stable delivery costs once established, as well as for interventions that substitute for existing resource use. For example, suppose the costs of introducing a new digital consultation service in hospitals need to be estimated:

- Implementation costs may include staff training, software development, IT integration and project management.
- Delivery costs may include staff time, IT support and infrastructure required for each consultation.

If the analysis aims to estimate the total cost of introducing and operating the service, both implementation and delivery costs should be included, with their timing taken into account.

7.4.2 *Perspective, identification and measurement of resources*

The perspective defined for the analysis determines which resources and costs are relevant for the costing. The identification and measurement of resources should therefore be consistent with the perspective adopted.

The **choice of perspective** should be based on the decision question, the intended use of the results and the stakeholders for whom the information is required. A healthcare payer perspective may be appropriate when the objective is to estimate expenditures borne by public healthcare payers and patients, whereas a provider perspective may be appropriate when the objective is to estimate the resources required to deliver an intervention. A broader healthcare-system or societal perspective may be appropriate when the analysis aims to capture costs borne by multiple stakeholders. Where relevant, costs can be reported separately for different stakeholders rather than combined into a single estimate. Care should be taken to avoid double counting, particularly when a payment represents a cost for one stakeholder but an income or transfer for another. Such transfers should be identified explicitly and should not be counted as costs at the overall system level.

In the Belgian reference case, the perspective is that of public and private healthcare payers, where private payers refer to patients paying out of pocket, and only direct healthcare costs are included (see section 7.4.6). Patient payments include co-payments, fees or other supplements, and payments for non-reimbursed healthcare services, medicines or devices. The distinction between amounts paid by patients and amounts reimbursed by compulsory health insurance should be maintained where relevant, as together they constitute the total expenditure associated with the healthcare resource. Non-healthcare costs borne by patients, such as transport or accommodation, are outside the reference case and require a broader perspective.

For example, suppose the costs of introducing a new home-based care programme need to be estimated.

- If the question is 'How much will home nursing organisations have to spend to provide this service?', a provider perspective is appropriate.
- If the question is 'How much additional expenditure will the public health insurance system and patients incur?', a payer perspective is appropriate.
- If the question is 'What resources does this service require overall, including patient costs, informal caregiver's time and productivity effects?', a broader societal perspective may be appropriate.

The next step is to **identify the resources** relevant to the intervention and the chosen perspective. These may include healthcare services such as consultations, hospitalisations, diagnostic tests, medicines, medical devices and healthcare professional time, as well as relevant non-healthcare resources. Depending on the perspective and research question, resources borne by patients or caregivers and productivity losses may also be relevant.

Resource use should be measured in natural units before being converted into monetary values. For example, hospital resource use may be measured as the number of hospital days, consultations, diagnostic tests or procedures, while staff resources may be measured in hours or full-time equivalents. The corresponding unit costs can then be applied to these quantities. Applying a constant unit cost is, however, a simplifying assumption. In practice, unit costs may vary with the volume of resources used, for example because of economies or diseconomies of scale or capacity constraints.



This may be particularly relevant for complex interventions, where increasing the volume of activity may require additional staff, infrastructure or organisational capacity, thereby changing the cost per unit of activity. Where such variations are expected to be material, they should be considered in the costing approach rather than assuming a constant unit cost throughout.

Resource use should be identified and **measured using appropriate data sources** (see Chapter 11) and the choice of source and methods should be reported transparently. Relevant sources may include prospectively collected data, clinical trials, observational studies, administrative or accounting databases, medical records, registries, surveys and patient diaries. Foreign data should only be used when relevant Belgian data are not available. Whenever possible, observed resource use should be used rather than relying solely on assumptions or expert opinion. Expert opinion may be useful where empirical information is unavailable, but its use should be justified and transparently reported.

7.4.3 Valuation of resources

The costing approach should distinguish between the quantity of resources consumed and the unit cost assigned to those resources, allowing the calculation to be reproduced and facilitating sensitivity analyses.

Costs should be valued at **opportunity cost**. In the reference case, market prices or reimbursement tariffs – if they exist – can be used as practical proxies, while recognising that these may not always reflect true opportunity costs.

Belgian costs and prices should be used wherever possible. Foreign cost estimates should not simply be converted into euros but may be used when reliable Belgian data are unavailable or the cost item has limited importance. When transferring a cost estimate from another country, the cost should first be updated for inflation using the Harmonised Index of Consumer Prices (HICP) for the local currency of the country in which the cost was incurred. The resulting cost should then be converted into its Belgian equivalent using an appropriate purchasing power parity (PPP) index. The impact of using foreign data should be explored in sensitivity analyses where relevant.

Costs should generally be expressed in current or most recent prices. When older cost data are used, they should normally be updated using the HICP, unless this is inappropriate for the specific resource.

For new interventions for which no official price is available, a plausible Belgian cost or price should be estimated using appropriate sources, such as manufacturer information, tariffs for similar interventions, treatment protocols, literature, administrative data, or micro- or macro-costing. The approach used to derive the estimate should be clearly reported.

Appropriate Belgian reimbursement tariffs and co-payments should generally be used. Distinctions between patient groups should be made when relevant to the intervention or the population being evaluated.

KCE Report 400 provides detailed guidance on the valuation of different types of healthcare resources, including healthcare services, pharmaceuticals and other healthcare products. It specifies the recommended approaches and data sources for assigning costs to resource use in economic evaluations and budget impact analyses.

7.4.4 Implementation and delivery costs

For interventions being introduced into practice, it is useful to distinguish between implementation costs and intervention delivery costs. **Implementation costs** are the resources required to introduce, adapt and establish an intervention in the relevant setting, such as training, adaptation of care processes, ICT development, communication and implementation support. **Intervention delivery costs** are the resources required to provide the intervention once implemented, such as healthcare professional time, medicines, equipment and consumables. Implementation costs may arise at different stages of an intervention's development and adoption, including during pilot implementation and wider scale-up. Depending on the purpose of the costing study, implementation and delivery costs may therefore be estimated separately or together.

For example, consider the introduction of a new clinical pathway for patients undergoing hip replacement. Implementation costs may include developing the pathway, adapting clinical protocols, training staff, modifying ICT systems and providing implementation support. Delivery costs would include the resources required to provide the pathway in routine practice, such as healthcare professional time, diagnostic tests, medicines, equipment and hospital stays. The former are primarily associated with establishing the pathway, whereas the latter are incurred as patients receive care through it.

A costing study should provide a clear picture of the resources required to implement and/or deliver an intervention. Depending on the purpose and nature of the study, relevant questions may include:

- Which healthcare professionals are involved, and how much time do they spend on implementation and/or delivery?
- What training, infrastructure, equipment, ICT and other implementation support are required?
- What administrative or coordination activities are required?
- Which components of implementation or delivery are the main cost drivers?
- How do resource use and costs vary across organisations, settings or stages of implementation?

Costs incurred solely for research or evaluation should generally be distinguished from costs that would arise when the intervention is implemented in routine practice. For example, costs of collecting additional data solely for a research study, conducting research-specific assessments or activities undertaken exclusively to evaluate the intervention would generally not be included as implementation costs. However, the distinction may depend on the purpose of the costing and on whether it is still uncertain whether the activity would also be required for routine implementation. Within the evaluation programme, the treatment of such costs should be discussed and agreed with the commissioner at the protocol stage, taking into account the purpose of the evaluation, the available budget, and the implications for the calculation and interpretation of the costs.

7.4.5 Costing approaches

Different approaches can be used to estimate the resources and costs associated with an intervention. The appropriate costing method depends on the purpose of the analysis, the cost object, the level of precision required and the availability of data. The aim is to obtain cost estimates that are sufficiently accurate and transparent for the decision problem, without collecting detail that is unlikely to affect the results. Costing methods can be classified along two complementary dimensions: how costs are assigned to the cost object (top-down or bottom-up) and the level of detail at which resource use is identified (gross-costing or micro-costing).³⁰³⁻³⁰⁵ Although many of the examples below relate to hospital care, these costing approaches can be applied across healthcare settings, including primary care, community care, home care, mental health care and long-term care.

Importantly, micro- versus gross-costing and bottom-up versus top-down costing describe different dimensions of costing. Micro- and gross-costing concern the level of detail at which resources and



costs are identified, whereas bottom-up and top-down costing concern how costs are derived, from individual resource use or from aggregate organisational costs. The approaches can therefore be combined. This distinction should be made explicit when describing the costing methodology.

Table 32 – Framework to classify costing methods

		<i>Level and type of data collected</i>		
<i>Level of identification of resource use items</i>		Expenditure data collected at organisational level (e.g. cost centre)	Resource use data collected for each individual patient and then multiplied by unit cost to estimate the expenditure	
	Highly detailed resource use items are identified	Top-down micro-costing	Bottom-up micro-costing	
	Aggregate resource use items are identified	Top-down gross-costing	Bottom-up gross-costing	

Source: Špacířová et al. (2020)³⁰⁴

7.4.5.1 Cost object and level of precision

The **cost object** is the activity, service, patient, treatment episode, care pathway or other unit for which costs are estimated. It should be defined before selecting the costing method.³⁰⁶

The **level of detail** should reflect the research question and the expected importance of the resources being measured. At the least detailed level, costs may be estimated as an average cost per patient-day or per case. Greater precision can be obtained by estimating costs for more homogeneous patient groups or case-mix categories. At the highest level of detail, micro-costing identifies individual resource items, measures their quantities and assigns a unit cost to each.

Micro-costing can provide more detailed estimates and may be particularly useful when resource use varies substantially between patients, when the intervention is new, or when the costs of specific components are important to the decision. However, it is resource-intensive and is not necessarily required for every cost component. A hybrid approach, in which detailed measurement is applied to important cost drivers while less important or difficult-to-measure costs are estimated using more aggregated methods, may therefore be appropriate.

7.4.5.2 Top-down and bottom-up costing

Two broad approaches can be distinguished for estimating healthcare costs: top-down costing and bottom-up costing. The main difference concerns the level at which resource use and costs are identified and measured. Top-down costing starts from the total costs incurred by an organisation, public authority or cost centre and allocates them to activities, services or patients. A cost centre is a unit within an organisation for which costs are separately identified and recorded, such as a hospital ward, laboratory, operating theatre or administrative department.³⁰⁶ Bottom-up costing starts from the resources used by individual patients or activities and builds up the total cost from the quantities of resources consumed and their corresponding unit costs. In practice, the two approaches can also be combined.³⁰⁴

TOP-DOWN COSTING

Top-down costing starts from the total expenditure or resource use of an organisation or cost centre over a defined period, usually obtained from accounting systems. These costs are subsequently assigned to cost objects, such as services, activities, patient groups or DRGs, using appropriate **allocation keys**. In hospitals, costs may first be grouped into medical cost centres, such as nursing wards, operating theatres, laboratories and radiology, and support or overhead cost centres, such as

administration, ICT and logistics. Allocation keys should reflect, as far as possible, the factors driving resource consumption, such as patient-days, staff time, number of procedures, equipment use or floor space. Overhead costs can be allocated using methods such as direct or step-down allocation. For example, in a hospital, the costs of the ICT department could be allocated directly to nursing wards, operating theatres and radiology based on the number of computers or IT support hours used. With step-down allocation, administration costs could first be allocated to ICT, logistics and clinical departments; ICT and logistics would then allocate their resulting costs to the clinical departments. This ensures that overhead costs are ultimately incorporated into the costs of hospital services or patient care.

A major advantage of top-down costing is that it captures the organisation's expenditure over the relevant period and provides comprehensive cost estimates without requiring detailed measurement of resource use for every patient. It is therefore particularly useful when reliable accounting and activity data are routinely available.

However, top-down costing also has limitations. Accounting costs may not correspond directly to the economic costs relevant to the evaluation, and the treatment of fixed, variable and capital costs requires consideration of the research question, perspective and time horizon. Top-down costing generally produces average costs, and estimates depend on the definition of cost objects and the allocation keys and assumptions used. Where a substantial proportion of costs consists of joint or overhead costs, allocation is necessarily an approximation. Differences in accounting and cost-allocation practices may also limit comparability between organisations, including when using Finhosta data.³⁰⁶

Finally, because top-down costing generally relies on retrospective accounting and activity data, it may be less suitable for estimating the costs of new interventions or services that are not yet routinely delivered. In such situations, bottom-up or micro-costing approaches may provide more appropriate estimates.

Example: In a hospital, total costs may first be allocated to departments such as nursing wards, operating theatres and laboratories. Overhead costs from administration or ICT can then be allocated to these departments using appropriate allocation keys. Departmental costs are subsequently allocated to patients or patient groups using measures such as patient-days, procedure time or relative weights. Costs that can be directly attributed to individual patients, such as specific medicines or implants, can be allocated directly. Thus, overhead costs are ultimately reflected in patient-level costs, but generally through intermediate allocation steps rather than being attributed directly.

BOTTOM-UP COSTING

Bottom-up costing estimates the costs of care by measuring the resources used for individual patients or activities. It is particularly useful when resource use differs substantially between patients or when detailed information is needed for a specific intervention or care pathway. Resources may include staff time, medicines, procedures, equipment and use of facilities.

The approach provides detailed estimates of resource use, but collecting these data can be time-consuming and may require observations or data from a sample of patients. The quality and representativeness of the estimates therefore depend on the available data and sample. Moreover, not all resources can be attributed directly to individual patients. Shared resources, such as facilities, laundry or management, may need to be valued using existing unit costs or complementary top-down approaches. Existing unit costs are costs per unit that are already available from the organisation or other reliable sources, such as the cost per bed-day, laundry item, square metre of facility space or hour of management time. When combining methods, care should be taken to avoid double counting.

Resources are valued by assigning an appropriate unit cost to the quantity consumed. Unit costs may be derived from accounting data, existing cost databases, tariffs or market prices. The appropriate source depends on the resource, perspective and purpose of the analysis. Tariffs or reimbursement



prices may be appropriate from a payer perspective but may not necessarily reflect the economic or opportunity cost of the resources used.

CHOOSING AND COMBINING APPROACHES

Neither approach should automatically be considered superior in all circumstances. The appropriate approach depends on the research question, the level of detail required, the availability and quality of data, and the resources available for the costing study. Although bottom-up costing can provide more detailed estimates of individual resource use, top-down costing may be more appropriate when the objective is to estimate average costs across a population or when comprehensive accounting data are available.

The approaches can also be combined. For example, staff time and medicines may be measured directly at patient level using a bottom-up approach, while shared overheads or the cost of a hospital day may be estimated using top-down data. When combining approaches, the components included in each estimate should be clearly specified and double counting should be avoided. In particular, costs already measured directly at patient level should not be included again in the top-down unit cost of a composite resource.

The choice of approach, the level of detail, the allocation methods and the data sources should be reported transparently so that the resulting cost estimates can be understood, assessed and, where possible, reproduced.

7.4.5.3 *Gross-costing and micro-costing*

Micro-costing and gross-costing refer to the level of detail at which resource use and costs are identified and measured.^{304, 305, 307}

GROSS-COSTING

Gross-costing estimates costs using relatively aggregated resource or cost categories. For example, hospital costs may be estimated using an average cost per inpatient day, outpatient visit, procedure or DRG. This approach is relatively simple and less resource-intensive, and can be appropriate when reliable, aggregated cost estimates are sufficient for the research question. However, aggregated estimates may mask important differences in resource use between patients, procedures or providers and may be inappropriate when the intervention has a substantially different resource profile from the activities underlying the average cost.

MICRO-COSTING

Micro-costing identifies and values individual resource components required to deliver a service or intervention. Depending on the intervention, these may include staff time, medicines, consumables, diagnostic tests, equipment and use of facilities. It provides detailed information on resource use and can be particularly useful for new interventions, when resource use varies substantially between patients, or when specific cost components are important to the decision. However, detailed measurement can be resource-intensive and may require primary data collection. Even micro-costing cannot directly measure all shared resources, which may need to be valued using existing unit costs or complementary costing methods.

CHOOSING AND COMBINING APPROACHES

The choice between gross-costing and micro-costing should be guided by the research question, the level of detail required, the variability in resource use, data availability and the resources available for the study. Micro-costing should not automatically be regarded as superior: its additional detail may not be necessary when reliable, aggregated estimates are sufficient. Conversely, gross-costing may be inappropriate when average costs do not adequately represent the resource use associated with the intervention.

7.4.5.4 Activity-based costing

Activity-based costing (ABC) is a costing approach developed to allocate costs more accurately to specific services, activities or other cost objects. Rather than allocating overhead costs primarily through cost centres and relatively broad allocation keys, ABC first identifies the activities performed and then assigns the resources consumed to those activities. The costs of the activities are subsequently allocated to the relevant cost objects using activity cost drivers, such as the number of procedures, staff time or use of facilities. ABC can therefore provide more detailed information on the resources required to deliver specific processes or services. When it allocates organisational costs to activities and then to cost objects, it can be considered a form of top-down micro-costing.³⁰⁴

ABC requires detailed information on the activities performed and the resources required for each activity. This information may be obtained through administrative data, interviews, observations or other primary data collection. Although ABC can provide a more refined allocation of overhead costs than conventional top-down costing, its implementation can be **time-consuming and data-intensive**, particularly in complex healthcare settings.

Time-driven activity-based costing (TDABC) was developed as a simplified form of ABC. Instead of requiring a separate cost driver for each activity, TDABC primarily uses **time as the cost driver**. It estimates the cost of supplying the capacity of each resource and the amount of time required to perform each activity. The approach typically involves defining the care pathway, mapping the activities involved, estimating the time required for each activity, calculating the cost per unit of available capacity, and then estimating the cost of the complete care cycle.³⁰⁸

An additional advantage of TDABC is that it can identify **practical capacity and unused capacity**, which may be particularly relevant when assessing healthcare services with substantial fixed or capacity-related costs. However, like ABC, TDABC requires detailed information about care processes and resource use and may therefore require considerable data collection and methodological expertise. Its usefulness should consequently be assessed against the additional precision it provides relative to simpler costing approaches.

7.4.6 Types of costs

Costs can be classified according to the type of resources involved and their relationship with the intervention or health condition. A commonly used distinction is between direct healthcare costs, direct non-healthcare costs and productivity costs.³⁰³ **Direct healthcare costs** relate to healthcare resources used for the prevention, diagnosis, treatment or management of a health condition, such as consultations, hospital care, diagnostic tests, medicines and medical devices. **Direct non-healthcare costs** concern resources outside the healthcare sector that are directly associated with the intervention or condition, such as patient or caregiver travel.

Indirect costs refer to costs that are not directly attributable to the provision of healthcare services. A major category of such costs is productivity costs. **Productivity costs** refer to the value of productivity losses resulting from illness, disability, treatment or premature death. They may include time away from paid work as well as, where relevant, losses in unpaid work. Their inclusion and valuation can substantially affect the results of an economic evaluation and depend on the perspective and methodological approach adopted. Because their measurement and valuation are subject to important methodological choices, the approach used should be clearly specified and justified.

The terminology of direct and indirect costs requires some caution. In health economics, 'indirect costs' have traditionally been used to refer primarily to productivity losses. In accounting and costing, however, indirect costs refer to costs that cannot be directly attributed to a particular cost object, such as overhead costs. To avoid this ambiguity, the term productivity costs is increasingly preferred in economic evaluation when referring to productivity losses. **Future unrelated healthcare costs**, such as healthcare consumed during additional life years gained as a result of an intervention but unrelated to this intervention or the initial condition it addressed, represent another category that does not fit neatly into the traditional direct/indirect classification. These costs are neither productivity costs nor



necessarily costs directly related to the intervention or condition being evaluated, and their inclusion depends on the costing framework, perspective and methodological conventions adopted.

For these reasons, costs in an economic evaluation should preferably be described according to the type of resource involved (e.g. healthcare, transport, informal care, productivity loss, etc.), the sector in which it occurs, and its relationship to the intervention or health condition (directly related or future unrelated), rather than relying solely on the labels 'direct' and 'indirect'. In the Belgian guidelines for economic evaluation KCE Report 400), future unrelated healthcare costs are referred to as 'indirect healthcare costs'.

In the Belgian guidelines, the reference case includes only direct healthcare costs (see Table 33, adopting the perspective of public and private healthcare payers (see section 7.2.2)).²⁹⁹

Table 33 – Classification of types of costs in the Belgian guidelines

	Healthcare costs	Non-healthcare costs
Direct costs	Included e.g. health services, medicines, devices, hospitalisations, outpatient care, examinations, consultations, etc.	Not included* e.g. travel expenses to and from hospital, informal care, invalidity/incapacity allowances, social services, etc.
Indirect costs	Not included* e.g. healthcare costs in life years gained (unrelated healthcare costs**)	Not included* e.g. productivity losses patient or informal caregiver

*Not included in the reference case; ** These are healthcare costs that occur during additional years of life not related to the disease or intervention being evaluated.

Source: Neyt et al. (2025)²⁹⁹



8 IMPLEMENTATION RESEARCH AND THE KCE EVALUATION PROGRAMME

KEY POINTS

- Implementation science concerns the systematic study of methods to promote the **uptake of evidence-based interventions** into **routine practice** and to improve the quality and effectiveness of healthcare.
- A clear distinction should be made between the **evidence-based intervention** (e.g. a care pathway or policy measure) and the **implementation strategy** used to support its adoption.
- In an **ideal situation** (when the focus of the KCE evaluation is on the implementation strategy), the **intervention** to be implemented is **already evidence based** and not the subject of the evaluation.
- **In practice**, particularly in the evaluation of complex interventions, this **distinction** between development of evidence on effectiveness of interventions and implementation strategies **can be blurred**, with effectiveness and implementation aspects being examined in parallel. It is also possible, within the KCE evaluation programme, that the implementation strategy (e.g. a new payment mechanism) aims to implement a consensus-based innovation.
- Implementation research typically starts with **describing the problem** (e.g. unexplained variation, lack of uptake in practice). Next, a **diagnostic phase** has to take place. This phase aims to identify contextual factors (barriers and facilitators) influencing uptake across individual, organisational and system levels.
- The use of **frameworks** such as CFIR, TICD or other structured approaches can support a comprehensive and systematic analysis of these determinants.
- Based on this diagnostic phase, a tailored **implementation strategy** should be developed. There is no single 'best' strategy applicable across contexts.
- Ideally, implementation strategies can be **theory-informed** (as supported by TBCM or TTT), though non-theory informed overviews of strategies (such as EPOC or ERIC) can also be consulted.
- Applying **generic implementation approaches by default** (e.g. audit and feedback, e-learning, electronic record modifications) without linking them to identified barriers and facilitators in the relevant context is **not recommended**.
- Implementation strategies should be **evaluated rigorously**, including impact, process and economic evaluations, using experimental, quasi-experimental and mixed methods designs. Hybrid designs may examine implementation and intervention effects simultaneously.



8.1 Introduction to implementation strategies and implementation research

In this chapter we describe the use of ‘implementation research’ in the context of the KCE evaluation programme. What is understood under ‘implementation’ and ‘implementation research’?

8.1.1 Implementation strategies

Wensing et al. (2021) describe **implementation strategies** as ‘*interventions that aim to enhance the uptake, sustainment, and scale-up of practices and those that aim to remove, reduce, replace, or restrict the use of ineffective interventions or practices*’.^{309, 310} These strategies can include actions targeting individuals (e.g. educational activities), organisational changes (e.g. additional staffing, task redistribution, or appointing an implementation champion), financial incentives (e.g. rewards for adherence to the intervention), technological tools (e.g. decision support systems), etc. Implementation strategies are often multi-faceted.³¹¹

8.1.2 The intervention or innovation

It is important to **distinguish implementation strategies from the intervention (or innovation) itself**. For example, when a guideline recommends a clinical treatment (e.g. a pharmaceutical intervention or counselling technique), this constitutes the intervention. Often, the guideline recommends its implementation based on evidence, consensus, patient preferences, and/or clinician perspectives. Implementation strategies, in contrast, focus on how to adopt the recommendation (or de-implement practices that deviate from it), for instance through education of healthcare professionals or organisational changes such as providing longer consultation times. While this example relates to the work of [KCE Practice](#), the same distinction applies to the KCE evaluation programme. If evidence supports an innovative care model (e.g. hospital at home services), these are the interventions/innovations that will need to be implemented. The activities used to introduce them (e.g. training or financial incentives) are the implementation strategies. Implementation strategies specifically aim to support their uptake in practice.

In practice, however, this distinction is not always made explicitly. Evaluations of complex interventions often combine the intervention itself with multiple implementation strategy components. It is therefore essential to disentangle these elements in order to clarify what is being evaluated.³¹² Implementation strategies may differ in their target groups (e.g. healthcare professionals), mechanisms, and components, and should ideally be considered separately.

8.1.3 Implementation research

Implementation research is defined by Eccles et al. (2006)³¹³ as ‘*The scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice, and, hence, to improve the quality and effectiveness of health services. It includes the study of influences on healthcare professional and organisational behaviour*’.³¹³ Implementation research thus studies the reasons why evidence, guidelines, innovations, best practices, etc. are not applied in daily practice, which strategies are needed to improve this as well as evaluating the effectiveness of these strategies.

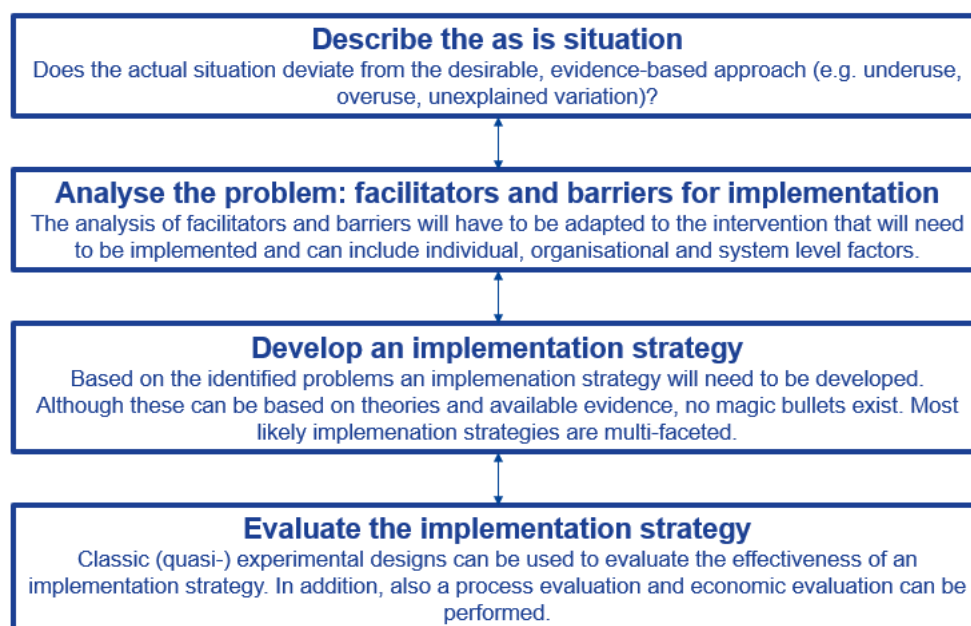
For topics submitted to the KCE evaluation programme, it is therefore important to clarify whether the objective is to evaluate the innovation, best practice or evidence-based intervention itself (e.g. an innovative care model), or the implementation strategy. This chapter focuses primarily on the latter. This implies that the innovation, best practice, or evidence-based intervention is not under evaluation. Rather, it is assumed to be effective and the objective is to implement it. While implementation processes can also influence outcomes in impact evaluations of complex interventions (see Chapter 5 on process evaluation), the focus here is distinct. As a consequence, the approach of implementation science, in the context of the KCE evaluation programme, is particularly relevant for policy initiatives that are already evidence-based. This can include, for example, interventions following a KCE report or based on international best practices to be implemented in the Belgian context. When effectiveness

is still uncertain, activities are closer to experimentation than implementation.³¹⁰ Nevertheless, this does not mean that during the development of the intervention itself implementation issues are not important. On the contrary, it is already during the development phase important to anticipate on implementation issues (e.g. stakeholder engagement to enhance acceptability and feasibility of the intervention). In addition, while the innovation/intervention submitted by the commissioner is not evidence-based, it might still be relevant to evaluate the implementation strategy (e.g. a new payment system that targets a certain behaviour).

8.2 Stages in implementation research

Already in 1997, Grol³¹⁴ described a ‘**model for implementation of change**’. Building on this work, these steps were consolidated into a model for planning change (see Figure 33).³¹⁰

Figure 33 – Stages of implementation change



Source: Adapted from Wensing et al. (2023)³¹⁰

A first step is to **analyse the ‘as is situation’** and how this deviates from the desirable ‘evidence-based’ situation, based on evidence from effectiveness, process, or economic evaluations. This can be done via quality indicators (see Chapter 9). A second step is to identify the **stimuli or problems for implementation**. This includes an analysis of facilitators and barriers that influence the introduction of the innovation. It is important that this analysis includes not only individual factors but also team, organisational and system level factors. The analysis will have to be adjusted to the innovation/intervention studied. In the context of the KCE evaluation programme, the interventions or policy innovations to be implemented can vary substantially in nature, as do the associated implementation challenges. For example, implementing the centralisation of complex surgical interventions differs fundamentally from the introduction of a case manager to coordinate care for persons with chronic conditions. Although such interventions may be supported by evidence, the determinants influencing their implementation and outcomes are likely to differ considerably.³¹⁵

Based on the diagnostic phase, an **implementation strategy** can be developed. Within the KCE evaluation programme, it is essential to align implementation strategies with the type of intervention and the specific implementation stimuli and challenges. There is no single best approach. Different contexts and target groups require tailored strategies. A systematic analysis of barriers and facilitators



should therefore guide the selection of a coherent set of strategies. This will often be a combination of actions at the individual (e.g. continuing education, feedback, practice support), team (e.g. leadership), organisational (e.g. revision of professional roles, redesign of healthcare delivery processes) and system level (e.g. financial incentives, adapted legislation).³¹²

Given the absence of a 'magic bullet'³¹⁶, it is particularly important to evaluate the effectiveness of the implementation strategy (using the designs described in Chapters 2 and 3). In this type of evaluation, the implementation strategy is the focus of the evaluation, not the innovation itself. This implies that the innovation may also be present in comparator groups, but without the additional implementation strategy.³¹²

8.2.1 'As is situation'

A first step is the description of a problem (e.g. the conservative first line treatment option in case of knee osteoarthritis is heterogeneously applied; complex cancer surgery for which concentration is recommended is fragmented across many hospitals). This can, among other methods, be done by means of quality indicators (see Chapter 9).

8.2.2 Identify facilitators and barriers

A next step is to understand the factors that facilitate or hinder the uptake of an innovation (e.g. a clinical guideline, a best-practice care pathway, or a new professional role). The approach typically begins with an **exploration of the existing scientific literature**³¹⁷ (see Textbox 23 for an example of barriers). A rapid ('quick') scan can establish what is already known about implementation determinants in comparable settings. Based on this quick scan, a decision on the need for further evidence synthesis can be made. Where sufficient primary research (or systematic reviews) exists (most commonly for guideline implementation, thus more relevant for KCE Practice), a dedicated systematic review of facilitators and barriers may be indicated. Yet, this will not always be necessary. The quick scan of the literature can also be a good starting point to guide and structure subsequent data collection efforts (e.g. interview guides, survey development).

Textbox 23 – Example of barriers to implementation of evidence

Practice environment (organisational context)

- Financial disincentives: e.g. lack of reimbursement
- Organisational constraints: e.g. lack of time
- Perception of liability: e.g. risk of formal complaint
- Patient's expectations: e.g. expressed wishes related to prescription

Prevailing opinion (social context)

- Standards of practice: e.g. usual routines
- Opinion leaders: e.g. key persons not agreeing with evidence
- Medical training: e.g. obsolete knowledge
- Advocacy: e.g. by pharmaceutical companies

Knowledge and attitudes (professional context)

- Clinical uncertainty: e.g. unnecessary test for vague symptoms
- Sense of competence: e.g. self confidence in skills
- Compulsion to act: e.g. need to do something
- Information overload: e.g. inability to appraise evidence

Source: Grol and Grimshaw 2003³¹⁸

Primary data collection can be used to complement and contextualise the literature.

A survey can be set up. This can either include the use of validated instruments or tailor-made questionnaires informed by prior evidence, field visits and stakeholder consultations.

Less intuitively, **routinely collected data** can also play a role at this stage by identifying patterns of implementation across populations (e.g. patient groups, regions, or provider types), thereby highlighting where barriers may be concentrated.

Qualitative methods⁹ are particularly valuable and may include focus groups and in-depth interviews with relevant stakeholders. These can be organised in a general manner (e.g. purposive sampling of healthcare professionals) or in a more targeted manner. The latter, for instance, could include the identification of implementers and non-implementers through a survey or routinely collected data (such as administrative datasets). In these distinct groups, the reasons for (not) implementing can then be explored via qualitative research methods.

Direct observation is also a possibility to consider. Although it is resource-intensive and not often used at KCE, it can yield rich, practice-based insights that are not readily accessible through self-report methods.

To ensure a comprehensive assessment, it is recommended to draw on **theories, models and frameworks (TMFs)** to guide data collection. Frameworks such as the Consolidated Framework for Implementation Research (CFIR) help to ensure that attention extends beyond individual behavioural factors to organisational, contextual and system-level determinants.³²⁰ While it is not always necessary to operationalise an entire framework, the selection (and omission) of components should be explicitly discussed, justified and documented within the KCE team (see Chapter 1).

As an alternative or complement to TMFs, structured **checklists such as the TICD** (Tailored Implementation for Chronic Diseases) are widely used. TICD distinguishes seven domains: (1) guideline or innovation factors (e.g. quality of evidence); (2) individual health professional factors (e.g. knowledge, skills, behaviour); (3) patient factors (e.g. preferences, motivation); (4) professional interactions (e.g. teamwork and referral pathways); (5) incentives and resources (e.g. financial incentives, budget availability); (6) capacity for organisational change (e.g. leadership, mandates, internal support); and (7) social, political and legal factors (e.g. regulation and contracts). Such frameworks support a systematic and transparent identification of barriers and facilitators, thereby strengthening the basis for subsequent implementation strategies.

Next, the barriers/facilitators and problems need to be organised and ranked according to priority. Wensing (2023)³¹⁷ describes tools such as Pareto (80/20 rule: 20% of the barriers explain 80% of the problem), Fishbone diagram and flow charts.³¹⁷ Within this exercise it is important to not only look to barriers but also consider the facilitators. Often facilitators (e.g. available infrastructure, motivated clinical leaders, routinely planned interdisciplinary meetings, patient advocates pulling for innovation, etc.) exist to build on during the implementation process. When the problems are described and the most important facilitating and blocking factors are identified, it is important to define the goals for the implementation programme. The implementation strategy needs to build on existing facilitators. In addition, they should focus on the most important obstacles (that can be changed) for the most important shortcomings in the care or healthcare system. Techniques like a Delphi-approach³¹⁷ or nominal group technique can support this exercise. It is also important to note that new blocking factors might emerge or change during the implementation process.³²¹

⁹ Established KCE methods for qualitative data collection can be applied³, alongside guidance on qualitative research and implementation-focused reviews (e.g. Holtrop et al., 2025)³¹⁹.



8.2.3 Development of an implementation strategy

Based on the diagnostic phase, an implementation strategy will have to be developed to overcome the main obstacles. The link between identified obstacles/facilitators and implementation strategy is essential (see Table 34 for examples of linking determinants to implementation strategies). It is bad practice to launch implementation interventions (e.g. audit & feedback, decision support tools, training) without understanding what the problem is. This jumping to conclusions should be avoided.

Table 34 – Examples of linking implementation strategies to determinants

Determinant	Implementation strategy	Mechanism	Implementation outcome
Provider knowledge deficit	Education (provision of information)	Awareness-building, knowledge-acquisition	Awareness; knowledge-acquisition
Provider skill deficit	Training (teaching and practice with corrective feedback)	Skill acquisition, refinement, mastery	Fidelity to EBP
Provider views EBP unfavourably	Audit and feedback Provision of descriptive social norms indicating peer use of EBP	Social pressure/norms	Adoption

EBP = evidence-based practice

Source: Lewis et al. (2021)³²² – extraction of some examples

In this respect, it is good practice to select implementation strategies that are **theory-informed** and explicitly linked to the determinants of behaviour. As highlighted by Kok et al. (2015),³²³ behaviour change methods are only likely to be effective if three conditions are met:

- they must target determinants that actually predict the behaviour of interest,
- they must be capable of influencing these determinants, and
- they must be translated into practical applications in a way that preserves their effectiveness while fitting the target population, culture and context.³²³

This underscores the importance of aligning identified barriers and clear objectives for implementation change with appropriate, theory-informed and evidence-based intervention components, rather than applying generic or off-the-shelf solutions. The theory explains how the mechanism of the implementation strategy works (assumed causal link between application of the method and behaviour change).³²³

It is beyond the scope of this process note to list or summarise common theories used in implementation science. The KCE team will have to perform a search and explore which theory can be helpful in the process of the development of implementation strategy. In any case, it is important that the theory is well-studied and understood and translated in practical applications. This will require a combination of desk research, brainstorming sessions and stakeholder involvement. While there is no such thing as 'the theory' to guide implementation change, tools exist to support the team (e.g. the Taxonomy of Behavior Change Methods (TBCM) by Kok et al 2015³²³; and the 'Theories and Techniques Tool' (TTT; <https://theoryandtechniquetool.humanbehaviourchange.org/>)) in making a choice. In addition, theories can also be applied in a 'common sense' manner: for each implementation improvement objective one or more implementation strategies are selected, informed by theories.³²⁴

A recent review by Vis et al. (2025)³²⁵ describes several methods matching implementation strategies to behavioural determinants. Yet, they indicate that practical guidance is often lacking. The **EPOC (Effective Practice and Organisation of Care Group) taxonomy**, as developed within the Cochrane Collaboration (but which is no longer kept up to date) is a well-known example in which implementation

strategies and measures are categorised in a logical way (but not linked to theory)³²⁶. It consists of four main domains:

- Delivery arrangements: changes in 'how' (e.g. group versus individual care; coordination of care amongst different providers), 'when' and 'where' healthcare is organised and delivered, and 'who' (e.g. self-management, task shifting) delivers healthcare.
- Financial arrangements: changes in the way how funds are allocated (e.g. extra funding; payment methods for healthcare professionals; the use of targeted financial incentives or disincentive (pay-for-performance), etc.
- Governance arrangements: rules or processes that affect the way in which powers are exercised (e.g. stakeholder involvement in decision making), particularly with regard to authority (e.g. ownership, accreditation), accountability of healthcare professionals (e.g. scope of practice), etc.
- Implementation strategies: interventions designed to bring about changes in healthcare organisations (e.g. organisational culture), the behaviour of healthcare professionals (e.g. audit & feedback, communities of practice, interprofessional education, opinion leaders) or the use of health services by healthcare recipients.^{324, 327}

Another example is the ERIC-list (Expert Recommendations for Implementing Change) which contains 73 implementation strategies (e.g. access new funding, conduct local consensus discussions, identify and prepare champions, shift and revise roles among professionals).³²⁸ Waltz et al. (2019)³²⁹ developed the CFIR-ERIC matching tool: <https://www.nccmt.ca/organizational-change/results/77..> However, it should be noted that the matching tool is based on largely lacking consensus between experts. Whereas cut-offs for consensus are usually set at 67 to 75%, the matching tool uses $\geq 50\%$ agreement between experts as the bar for a 'Level 1 matching strategy', and $\geq 20\%$ consensus as the bar for a 'Level 2 matching strategy'. Therefore, consulting the ERIC list as a list of options and working with implementation scientists to reasons towards potential matches is likely more fruitful. Given the low level of consensus, variability in expert recommendations and the importance of contextual factors, the CFIR-ERIC matching tool's outputs should not be applied in isolation (see Figure 33).

A recent scoping review by Meng et al. (2026)³³⁰ highlighted the current use of the tool. It should be noted that its application remains largely confined to clinical intervention improvement. Moreover, an important limitation of this classification is its reliance on (largely lacking) expert consensus, which introduces subjectivity. Also the relatively generic nature of suggested strategies may limit local applicability. In addition, the absence of a standardised update mechanism raises the risk that some components may become outdated over time.

The effectiveness of tailoring implementation strategies was subject of a recent Cochrane review.³³¹ Tailored strategies are specifically designed to address identified barriers and facilitators in a given setting: e.g. if nurses in a busy clinic find a guideline too long to use, a tailored strategy might provide a short visual summary or electronic prompts to remind the nurse about which actions to take.³³¹ The tailoring of implementation strategies implies not only the linkage of interventions with identified barriers. It also includes matching strategies to the different stages of change:³²⁴

- Orientation: e.g. no interest or not aware → dissemination via different communication channels, and actively involving key actors;
- Insight: e.g. overestimation of its own behaviour → self-evaluation methods, feedback;
- Acceptance: e.g. negative attitude (focus on disadvantages, not believing in the feasibility) → demonstrate the applicability, adapt the innovation to the concerns of the target population;
- Change: e.g. not adopting the change due to time constraints → provide additional funding to have more time;
- Maintenance: e.g. not part of routine care (fallback on old habits) → monitoring & feedback, part of care pathways.³²⁴



The Cochrane review³³¹ found that, in the context of guideline implementation, tailored implementation strategies likely lead only to small improvements in healthcare professionals' adherence to guidelines compared with non-tailored approaches or no intervention. This suggests that evidence supports the value of tailoring, but that the expected effects are modest rather than substantial. While this research is mainly focused on guideline implementation, clear parallels can be drawn with the KCE evaluation programme.

8.2.4 Evaluation

The evaluation of implementation strategies can include an impact, process, and economic evaluation. We refer the reader to the respective chapters for guidance on study design. For impact evaluation, specific designs exist that combine effectiveness and implementation elements: 'effectiveness-implementation trials'. These hybrid trials^{5, 332} aim to increase the speed of the translation of scientific discoveries into practice.³³³ More information on these hybrid trials can be found in Chapter 8 of the book edited by Richards et al. (2026).^{7, 333} In summary, three types are described:

- Assessment of an intervention's effectiveness in routine practice (with primary outcomes typically at the patient level, such as functioning or symptoms), with an exploration of implementation issues as a secondary aim (with secondary outcomes such as feasibility, acceptability, or facilitators and barriers to implementation). As a result, the design is often a pragmatic trial combined with a mixed-methods approach (guided by a theoretical model framework such as CFIR; see Chapter 5) to study the implementation issues. Such a trial can be considered when the 'intervention' that is the subject of the KCE evaluation programme is either introduced without convincing evidence of its effectiveness, or when the intervention has been substantially modified compared to what was tested in the literature. A limitation is that implementation issues are only studied at the study sites, which may provide only a partial view of real-world implementation, as these sites might be more motivated.
- A second type is the inverse: the primary focus is on implementation issues (e.g. adoption, fidelity, and reach at the patient, clinician and site level), while health outcomes (symptoms, functioning, etc.) are studied only as a secondary objective to assess whether the implementation strategy contributes to effectiveness. These are most often cluster randomised clinical trials, as the implementation strategy typically needs to be tested in 'clusters' to avoid contamination. This type of design, where the 'implementation strategy' is in fact the 'intervention arm', is suitable when there is already evidence on the effectiveness of the intervention being implemented (e.g. a care pathway or payment model). However, when there is substantial evidence on both the effectiveness and safety of the intervention, it may be sufficient to focus solely on the implementation strategy; in that case, a hybrid trial is not required. To make the study as lean as possible, it should be assessed whether the secondary outcomes can be derived from electronic patient records or routinely collected data. In addition, it may be considered to work with proxies (e.g. health outcomes inferred from utilisation patterns, such as identifying diabetes based on medication use).
- The third group consists of designs in which the outcomes of interventions/innovations and implementation strategies are studied simultaneously. This dual focus requires dual randomisation: at the patient level for health outcomes and at the site level for implementation outcomes. Importantly, studies of this kind also require solid sample size calculations for both the primary intervention outcome and the primary implementation outcome, and for which clustering of data should be taken into account (see Chapter 4). This group can use a combination of experimental and quasi-experimental designs. For instance, implementation can be evaluated using a stepped-wedge approach (or non-randomised alternatives) at the site level, while effectiveness is assessed through classical patient-level randomisation. This type of design can be applied in the KCE evaluation programme for topics that have some evidence of effectiveness but require substantial adaptation to the Belgian context.³³³

Hybrid study designs are only of interest when both effectiveness and implementation research questions are pertinent. The choice between the three hybrid types depends on the priority given to the outcomes and should involve discussion within the KCE team, based on an assessment of the existing evidence, consultation with stakeholders, and, of course, the commissioners' main interests.

A classic implementation trial (for innovations for which there is sufficient evidence on its effectiveness) is a pragmatic trial in which both in the intervention as well as in the control arm the intervention is available. Yet, in the control arm the implementation strategy (which is being evaluated) is absent, different, or minimised.³¹²

8.3 Concluding remarks regarding the use of implementation research in the KCE evaluation programme

For the KCE evaluation programme, it will be important to clearly identify what the question of the commissioner is: is the focus of the evaluation on the innovation/intervention, the implementation strategy, or both? While this step is important, it will not always be clear-cut.

A new payment strategy, for instance, can be considered as the intervention while it can also be considered as an implementation strategy to change the behaviour of healthcare professionals in a desired (evidence-based) direction. While the innovation is ideally evidence-based before implementation research starts, this will in practice not always be the case (e.g. a policy negotiation that results in an intervention that is implemented). In such cases, the research team will have to study the submission and discuss with the commissioner to understand what the question is and what the KCE evaluation could contribute.

It is certainly not necessary to always go through all stages of implementation research. It can be very relevant to policymakers that KCE studies first the determinants for implementation and develops and recommends an implementation strategy which can be evaluated in follow-up research. It is also possible that, even in an observational study (e.g. in case a policy intervention is already being implemented), also (intermediate) effects are measured. Examples are:

- Does the strategy that is implemented change the behaviour of patients and/or healthcare professionals?
- Do sites (e.g. hospitals or GP practices) with higher uptake have better results?

Also these observations could result, next to implementation outcomes (e.g. uptake), in meaningful recommendations.

This chapter can be used to situate the questions submitted for the KCE evaluation programme. It can help to decide whether it concerns the evaluation of an intervention and/or implementation strategy. It can also help to decide on which of the stages are relevant to study as to enable the formulation of policy relevant recommendations.



9 QUALITY INDICATORS

KEY POINTS

- Donabedian proposed the triad of **structure**, **process** and **outcome** for the description and evaluation of the quality of healthcare.
- Structure and process indicators should be supported by **valid evidence** that relates these indicators to the health outcomes of interest.
- Quality indicators should provide: 1) a quality **goal**, i.e. a clear statement about the intended goal or objective; 2) a **measurement concept**, i.e. a specified method for data collection and calculation of the indicator; and 3) an **appraisal concept**, i.e. a description of how a measure is expected to be used to assess and/or improve quality.
- Structure, process and outcome indicators present a **mix of strengths and weaknesses**. With respect to outcome indicators, attributing patient health outcomes to providers (e.g. practitioners, hospitals) should be done with caution. In addition, potential confounding variables, including case-mix variables, should be identified and appropriately accounted for by risk adjustment.
- QI development comprises various steps: 1) **topic selection**, 2) **guideline selection**, 3) extraction of **recommendations**, 4) **QI selection**, 5) **practice test**, and 6) **implementation**. Every phase of the development process should be transparently documented within the KCE report or its appendices.
- A principal element of any quantitative assessment is **data quality**. More precisely, it has to be assessed whether the data are **complete**, **valid**, **reliable** and **comprehensive**, whether they have been collected in a timely manner and whether applied definitions are consistent.
- With regard to the **number of indicators** that should be selected, a **balance** should be made between on the one hand **comprehensiveness** of the set and on the other hand the **resources required** to develop and calculate them as well as the interpretation of a large amount of QI results.
- Quality assessment teams should **update and adapt** QIs when indicated; in some instances, they may be retired.
- The **GRADE working group** developed a framework for the integration of guideline and QI development.

9.1 Introduction

Quality measurement and improvement are critical components in the delivery of healthcare services.³³⁴ To achieve this goal, quality indicators (QIs) can be employed.

Detailed documentation of administered care is required to identify **areas for improvement** and to raise awareness and motivation among stakeholders.³³⁵ Clear and unbiased data are vital, as many healthcare professionals overestimate the quality of their performance.³³⁶ Data on the actual care delivered can create a sense of urgency that stimulates the adoption of recommended practices or the de-implementation of redundant procedures.³³⁵ For instance, KCE report 305 revealed that the recommended use of FDG-PET(/CT)^r in patients with head and neck squamous cell cancer (HNSCC)

^r FDG-PET(/CT): Positron-emission tomography (/Computed tomography) with the tracer molecule fluorodeoxyglucose merges functional PET data with high-resolution anatomical CT maps to identify

was below par.³³⁷ Only 47.6% of patients with stage III-IV head and neck squamous cell cancer (HNSCC) underwent FDG-PET(/CT) before start of treatment, while the target was set at $\geq 90\%$. On the other hand, 22.9% of patients with stage I-II HNSCC had a FDG-PET(/CT), while this diagnostic procedure is not recommended in this subgroup of patients (the target was set at $\leq 5\%$).³³⁷

Yet, evaluating the standard of care is inherently complex because it is a theoretical concept that can encompass different aspects depending on the exact definition and the context of measurement.³³⁸

Before turning our attention to QIs, we need to think about what good healthcare means^s. Initial frameworks of healthcare quality derived primarily from clinical experts and researchers. The last decades, however, a consensus has emerged regarding the vital contribution of patient and public preferences.³³⁹ The framework issued by the National Academy of Medicine (NAM, known until 2015 as Institute of Medicine (IOM)) in 2001 distinguishes **six domains of quality of care**.³⁴⁰

- **Effectiveness**: providing evidence-based health care services to those who need them;
- **Safety**: avoiding harm to people for whom the care is intended;
- **People-centredness**: providing care that responds to individual preferences, needs and values;
- **Timeliness**: reducing waits and sometimes harmful delays for both those who receive and those who give care;
- **Efficiency**: avoiding waste, including waste of equipment, supplies, ideas, and energy;
- **Equity**: providing care that does not vary in quality because of personal characteristics such as gender, ethnicity, geographic location, and socioeconomic status.

These are considered the core domains of quality of health care. When evaluating the quality of care, it is essential to determine which specific domains require assessment. Different quality indicators will be required to measure the various aspects of care. For some evaluation studies, it may be indicated to assess other “cross-cutting system properties”, such as sustainability, adaptability and resilience.

For several years, KCE has been engaged in a quality system in oncology, that started with the development of clinical practice guidelines, followed by the development of a set of indicators together with the Belgian Cancer Registry (BCR), who provided individual feedback to all Belgian hospitals.^{337, 341-346} Within the **KCE evaluation programme**, quality indicators can be utilised for multiple purposes. They may be part of a comprehensive evaluation study (e.g. in the process evaluation of an intervention, see Chapter 5), yet, they can also be used to describe the ‘status as is’ (i.e. before the implementation of change) as well as to assess the effect of an implementation strategy (e.g. in the fourth step of Grol and Wensing’s framework for implementation of change (see Chapter 8, Figure 33).^{310, 314}

Lastly, there is quite some **inconsistency** in the way **terminology and definitions** are used in quality assessment/assurance.³⁴⁷ For instance, a quality indicator was defined by some scholars as a “measurable element of practice performance for which there is evidence or consensus that it can be used to assess the quality, and hence change in the quality, of care provided”,³⁴⁸ while others used the same definition to describe performance indicators.^{349, 350} Within the Grading of Recommendations Assessment, Development and Evaluation (GRADE) initiative on the integration of guideline and QI development (see section 9.6), the following definitions of terms have been proposed:³⁵⁰

precise spatial locations of metabolic activity; it is used in oncological staging as well as treatment monitoring.

^s For more in-depth background on healthcare quality, the reader is referred to “An introduction to healthcare quality: defining and explaining its role in health systems” by Busse et al., 2019.³³⁹



- **Quality indicators** (may also be conceptualised as quality markers/quality constructs/quality parameters) are constructs used as a guide to monitor, evaluate, and improve the quality of services;
- **Performance measures** (may also be conceptualised as performance tools or performance instruments) are tools that quantify or describe measurable elements of practice performance;
- **Performance indicators** (may also be conceptualised as performance estimates) are quantifiable and measurable units or scores of practice.

This chapter focuses on quality indicators, as described in the above definition of terms.

9.2 From quality goals to quality measurements

9.2.1 Purposes of quality assessment

Quentin et al. (2019) distinguish two main purposes of quality measurement^t.³³⁸

- as a summative mechanism for external accountability and verification in quality assurance systems;
- as a formative mechanism for quality improvement.

Braspenning and colleagues (2020) add a third: scientific research.³³⁵

Although the distinction between quality assurance and quality improvement may be helpful, they are not always easy to separate in practice. From a societal perspective, quality assurance aims at eliminating poor-quality care and thus contributes to raising the overall standard of care. In addition, various strategies categorised as quality assurance methods (e.g. public reporting), contribute to improving quality of care and assuring public trust in healthcare services.³³⁸ Audit and feedback can be used both as a formative tool (preferably) and as a summative tool (often without the feedback).

Within the **KCE evaluation programme**, we will most often be asked to develop QIs for summative purposes, e.g. to identify problems with quality of care and assure a sufficient level of quality across health care providers. Consequently, a high level of precision is indicated and advanced statistical methods may need to be employed to make sure that observed differences between providers are “real” and attributable to provider performance (cf. infra).³³⁸

9.2.2 Advancing from quality standards to quality measurement

In order to achieve improvements of the quality of care that is offered, **setting standards** through the development of guidelines and clinical pathways is **not sufficient**.³⁵¹

A survey on the awareness of the NICE guideline on Endometriosis and the NICE quality standards (QS172) on endometriosis amongst healthcare professionals illustrates this.³⁵² While staff from endometriosis specialist centres were all aware of the NICE guideline and had full knowledge of it, this was not the case among non-specialist gynaecologists. They were generally aware of the guideline, but in some cases did not feel fully up to date or knowledgeable about its content.³⁵² An assessment of the impact of four ESHRE guidelines, among which the guideline on the management of endometriosis, came to comparable conclusions.³⁵³ Overall, about 90% of the clinician-respondents (N=260) knew and downloaded the guideline on endometriosis, while 72% reported to use the guideline in their routine practice. About one in two of the clinician-respondents reported that the ESHRE guideline resulted in changes in their practice, with one in three who made changes perceiving

^t For a more comprehensive overview of the comparative characteristics of quality measurement for quality assurance vs. quality improvement purposes, the reader is referred to Table 3.1 in Quentin et al. (2019).³³⁸

*that their patients benefitted from it. Only 12% referred patients to the patient-friendly version of the ESHRE guideline on endometriosis.*³⁵³

Setting standards by itself does not lead to change. First of all, standards must be taught to healthcare providers, which can be achieved through accreditation and certification. In a second step, monitoring (e.g. through audit and feedback) can be used to assess the adherence to the guidelines/standards (this can be done with the assistance of quality indicators) and necessary **actions** can be taken to implement change (e.g. pay for quality^u, public reporting).³⁵¹ Based on the recommendations adopted in clinical practice guidelines (or other sources of (preferably high level) scientific evidence), quality indicators can be developed to aid the monitoring of care.

Usually, quality indicators are expressed with a **numerator**, which describes what is desirable, and a **denominator**, which describes the target population.

For example, 'Proportion of patients with head and neck squamous cell carcinoma [denominator] who underwent FDG-PET(/CT)' within 6 weeks before start of treatment [numerator]'.

Most often the quotient is expressed as percentage.

It is important to note, that elements of the care process do not **signify quality** until their **relationship to desirable outcomes** has been proven.^{355, 356}

9.2.3 Components of quality indicators

In essence, quality indicators function as **quantitative measures that provide information about any of the six domains of quality of care**. Therefore, quality indicators should provide:³³⁸

1. **A quality goal**, i.e. a clear statement about the intended goal or objective; for example, when performing surgery in women with ovarian endometrioma, clinicians should perform cystectomy instead of drainage and coagulation, as cystectomy reduces recurrence of endometrioma and endometriosis-associated pain;³⁵⁷
2. **A measurement concept**, i.e. a specified method for data collection and calculation of the indicator; for example, 'the proportion of women with ovarian endometrioma, in whom cystectomy, and not drainage and coagulation is performed during surgery'; and
3. **An appraisal concept**, more precisely a **target or threshold**; for example, if in 90% of women with ovarian endometrioma, cystectomy is performed during surgery, and not drainage and coagulation, this is considered to be good quality.

Especially in quality assurance contexts (i.e. summative evaluation), measures need an appraisal concept (i.e. a defined threshold), because otherwise they are unable to indicate whether measured values denote good or bad quality of care.³³⁸ Importantly, this threshold should not be set at 0% or 100%, because valid clinical reasons (such as severe illness or rapid death after diagnosis) may prevent certain patient groups from following standard guidelines.

The Achievable Benchmark of Care (ABCTM) approach^v has been developed to provide an objective, reproducible and attainable target that is based on data derived from assessing actual performance rather than on subjective opinions as to what 'should' be done.³⁵⁸ The 'pared mean' is used to operationalize the benchmark: it summarises the performance of top-ranked providers whereby at least 10% of the patient pool across all providers is included.³⁵⁸

^u For more in-depth background on pay for quality, the reader is referred to KCE Report 424.³⁵⁴

^v For more information on the Achievable Benchmark of Care (ABCTM) approach, the reader is referred to Weissman et al. (1999).³⁵⁸



9.3 Types of quality indicators

9.3.1 Structure, process and outcome indicators

Various frameworks exist for the categorization of quality indicators (see Textbox 24),^{338, 359} yet, the most frequently used is Donabedian's classification. Based on the concept of 'input – process - output' applied in industrial manufacturing, Donabedian proposed the triad of **structure**, **process** and **outcome** indicators for the description and evaluation of the quality of healthcare.³⁵⁵ Translating Donabedian's position that 'good structure increases the likelihood of good processes, and good processes increase the likelihood of good outcome',³⁶⁰ to for instance endometriosis care: the availability of the right mix of qualified professionals at a hospital increases the likelihood that infertility treatment or chronic pelvic pain management will be performed following current professional standards, and this in turn increases the likelihood of desirable outcomes, such as quality of life.

Textbox 24 – Typologies of quality indicators

Proportion-based indicators vs. simple count-based indicators:³³⁸

- Most indicators are expressed as **proportions** (or rates) with explicit numerators and denominators;

e.g. proportion of women with IBC who received mammography and breast sonography before any oncological treatment

- **Count-based indicators** typically measure rare sentinel events, serving to operationalise patient safety within quality frameworks and identify intrinsically adverse occurrences;

e.g. counting retained surgical objects

Generic vs. disease-specific indicators:³³⁸

- **Generic** indicators measure aspects of care that apply to patient populations;

e.g. the proportion of patients who waited more than one week for a general practitioner appointment

- **Disease-specific indicators** are exclusively applicable to patients presenting with a designated diagnosis;

e.g. the proportion of patients with head and neck cancer who are alive 30 days after surgery

Donabedian's classification of structure, process and outcome indicators:^{335, 338, 355, 359, 360}

- **Structure indicators** assess the setting of care; they relate to material resources (*e.g. facilities, capital, equipment, drugs*), intellectual resources (*e.g. medical knowledge, information systems*) and human resources (*e.g. staffing ratios, qualifications of medical staff*). They are generally reported as a frequency per 1000 patients or as a dichotomous numeric value.

- **Process indicators** are used to assess whether high-quality care actions occur during service delivery. Ideally, they rely on robust scientific evidence linking compliance to improved patient outcomes. Process indicators comprise the predominant share of most quality measurement frameworks. They can be expressed as a numerator and a denominator;

e.g. proportion of women with invasive breast cancer who received mammography and breast sonography before any oncological treatment

but also in a dichotomous form, which can then be translated in a percentage:

e.g. staff performed hand hygiene measures before care (yes) or skipped it (no); % of staff who performed hand hygiene measures before care.

Patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs) can also be integrated into process indicators;

e.g. proportion of women who had breast reconstruction, who felt content with the shared decision-making process

- **Outcome indicators** specify the effects of care processes on the health status and wellbeing of patients (and populations), but they may describe also patient evaluations of care. Like process indicators, they can be expressed as a numerator and a denominator;

e.g. proportion of women with epithelial ovarian cancer who died within 30 days of surgery.

Note: This selection of typologies is intended to be illustrative rather than exhaustive.

9.3.2 Strengths and weaknesses of structure, process and outcome indicators

All three indicator types present a mix of **strengths and weaknesses**^w. **Structure indicators** offer the advantage of high accessibility, clear traceability, and straightforward verification. However, their primary limitation is that the link between structures and clinical processes or outcomes is often indirect and dependent upon healthcare providers' behaviour^x.³³⁸

Process indicators offer advantages including straightforward measurement, uncomplicated interpretation because there is often no need for risk-adjustment (cf. infra). In addition, since poor performance on process indicators can be directly attributed to the actions of healthcare providers, there is clear indication for improvement, e.g. by better adherence to clinical guidelines.³⁶¹ However, assessors should realise that healthcare remains inherently multifaceted, whereas process indicators frequently restrict their focus to isolated interventions targeting (sometimes) 'narrow' patient cohorts.³³⁸

Outcome indicators are usually more relevant to patients and policymakers. They may also encourage innovations in healthcare provision if these lead to better outcomes than following old routines. Yet, outcomes may require a long time before they manifest themselves, thereby complicating the utilization of outcome measures for quality assessment.³³⁸

Comparing patient health outcomes between providers (e.g. practitioners, hospitals) should be done with caution. **Potential confounding variables**, such as the influence of other providers or the meso and/or macro level of the healthcare system on the outcome of interest, should be carefully explored in concertation with relevant stakeholders.³⁶² Accurate attribution of causality is essential to ensure that providers are not unjustly held accountable for external factors beyond their control.

For instance, certain outcomes related to surgical interventions (e.g. 30-day mortality rates) may not only depend on the expertise and experience of the individual surgeon, but also on the appropriateness of the care plan developed by the multidisciplinary team, the quality of care provided by the anaesthesiologist as well as the quality of care provided in the intensive care unit.

Case-mix variables, such as sex, age, severity of illness/cancer stage, socioeconomic status, adherence to treatment, health literacy, are another type of potentially **confounding** external factors

^w For a comprehensive overview of the strengths and weaknesses of structure, process and outcome indicators, the reader is referred to Table 3.3 in Quentin et al., 2019.³³⁸

^x Actually, structure, process and outcome indicators are all dependent upon healthcare providers behaviour.



which are beyond provider influence. These factors should be identified and appropriately accounted for by **risk adjustment**^y.

Assessors should realise that in cases where responsibility is attributed incorrectly, quality improvement measures will be in vain, as they will miss the appropriate target.³³⁸ Furthermore, this may yield counterproductive results, as unfair judgments risk diminishing providers' motivation and trust in quality assurance.

9.4 Operationalising the theoretical concept of quality into a set of quality indicators

Operationalising the theoretical concept of quality by translating into a set of quality indicators needs a clear understanding of the **purpose** of the measurement (cf. section 9.2.1) as well as the **context**.³³⁸ With regard to the context, a simple example: while acute care interventions (e.g. rapid triage protocols) generate immediate, observable effects, the effectiveness of primary or secondary prevention (e.g. vaccination, screening) is typically measurable across an extended time horizon.³³⁸

In the following sections we reflect on the various steps in QI development, data for quality assessment, the recommended set of QIs as well as composite indicators, and sources of available QIs.

9.4.1 Development of quality indicators

A wide variety of methodological approaches for guideline-based QI development have been described in the literature. Yet, it is unclear which method leads to the best QIs.³⁴⁷

It is very important that every phase of the development process is transparently documented within the KCE report or its appendices.

As for all KCE reports, the KCE team should first of all assemble a **multidisciplinary panel** of clinical experts, ensuring comprehensive representation across all professional disciplines engaged in the patient care path. To prevent a narrow viewpoint, active inclusion of paramedics (e.g. physiotherapists, dieticians), (specialised) nurses and patient representatives should be considered.³⁶⁴ This panel will support the KCE team throughout the development process.

Kötter et al. (2012) specified the following steps in guideline-based QI development^z:³⁴⁷

1. Topic selection
2. Guideline selection
3. Extraction of recommendations
4. QI selection
5. Practice test
6. Implementation

^y For more information on risk adjustment for measuring health care outcomes, the reader is referred to Iezzoni (2003).³⁶³

^z For a comprehensive overview of all methodological approaches identified for various steps in guideline-based QI development, the reader is referred to Figure 3 in Kötter et al. (2012).³⁴⁷ CEQUID, the Cebam Quality Indicator Development guidance document, specifies 8 steps in QI development.³⁶⁵

TOPIC SELECTION

At KCE, topic selection is dependent on the research questions defined in the selected TPF. Yet, it may be wise to explore in the scoping phase of the study whether e.g. the scope can be narrowed, and/or whether certain areas merit specific attention as they present areas for quality improvement (e.g. sentinel lymph node biopsy only in invasive breast cancer with clinically negative lymph nodes).³⁶⁶ In this phase the purpose of the QI should also be defined (cf. section 9.2.1), just like the intended users (e.g. policy makers, GPs, hospital staff) and the quality domains involved (cf. section 9.1).³⁶⁵

GUIDELINE SELECTION

With regard to guideline selection, we refer to the KCE process note 'Research Methods for guidelines – update 2025', where you can find – among others – guidance on search strategies for evidence-based guidelines (and other sources for guidelines, e.g. dedicated websites) and on the quality assessment of published guidelines (with the Appraisal of Guidelines Research and Evaluation in Europe (AGREE) II instrument).³⁶⁷

EXTRACTION OF RECOMMENDATIONS

In the following step, the KCE team will extract the recommendations from the selected guidelines in a database, structured by e.g. clinical domain, phase of the care pathway. Not all recommendations should be selected. They should be relevant for the selected scope and quality domains defined and must be sufficiently specific and measurable to be translated into a QI with a numerator and a denominator.³⁶⁵ The research team should further define whether e.g. only strong recommendations are selected, and whether Good Practice Statements and recommendations based on consensus are included.

Existing QIs (see section 9.4.6) are added to the database, after quality assessment (e.g. with the AIRE instrument, see Textbox 25).

A transparent description of the respective sources of the recommendations/QIs is essential.

QI SELECTION

After compiling a long list of QIs, a definitive subset must be chosen. This process can be facilitated by the RAND/UCLA^{aa} Appropriateness Method.³⁶⁸ This approach combines scientific evidence from a review with expert judgment, reaching a consensus from a group with individual opinions aggregated into a refined aggregated opinion following face-to-face discussions.³³⁵

In the **first round**, experts receive the literature review and a survey containing the long list of recommendations and (existing) indicators^{bb}. They are asked to read the review and subsequently rate^{cc} the indicators based on certain criteria,³³⁵ which may include impact on patient health,

^{aa} RAND/UCLA stands for a partnership between the RAND Corporation and the University of California, Los Angeles (UCLA).

^{bb} In some projects, it may be necessary to shorten the long list of recommendations that the panel needs to assess since it might be too long.

^{cc} The panellists rate the suggested QIs on a point scale, for example, from 1 to 9. After calculating the average score of all experts, each QI is selected for the second round according to a previously determined weighting (e.g. average score above 80%), considered again and discussed during the second round (e.g. average score between 60% and 80%), or rejected (e.g. average score below 60%). Alternatively, an indicator is selected for the second round in case of a median score of 7 or higher, unless there is disagreement (at least 30% of the scores in both the lowest tertile – score 1, 2, or 3 – and in the highest tertile – score 7, 8, or 9). Indicators with a median score of 1–3 or 4–6, respectively, are not selected or introduced in the second round for discussion.³³⁵ Last, if a strong selection is required based on relevance, panellists may also be requested to submit a ranked top-five list.³³⁵



contribution to high quality care, improvability, relevance to obtaining value for money, feasibility, measurability.^{335, 347} Of note, panellists are also invited to adapt the recommendations/indicators where indicated and to suggest new indicators.

The **second round** entails (a) face-to-face meeting(s) during which initial scores are reviewed, panel members discuss all QIs and then re-rate them. Only second round ratings are used to derive indicators.³³⁵

The results/consensus derived during the meeting(s) are summarized and in a **third and final step** presented (e.g. by mail) to the panel members for their final approval. The final set of recommendations is translated into QI, i.e. a measurement concept with a well-defined numerator and denominator (see Table 35 and section 9.2.3), and where indicated exclusions.

Table 35 – Rephrasing recommendations into QIs

Recommendation	QI
Time interval from surgery to the start of adjuvant chemotherapy ideally up to 6 weeks or maximum within 8 weeks*	Proportion of patients with non-metastatic invasive breast cancer [denominator] that start adjuvant chemotherapy within 6 weeks interval from the date of surgery to the start of adjuvant chemotherapy [numerator]
SLNB, rather than full nodal clearance, is the standard of care for axillary staging in early, clinically node-negative breast cancer	Proportion of patients with invasive BC and clinically negative axilla (i.e. negative at clinical examination by palpation and axillary ultrasound, cN0) [denominator] who underwent SLNB, [numerator] (excluding patients who received NAST) [exclusion]
Post-operative RT is strongly recommended after BCS	Proportion of patients with invasive BC and node negative disease [denominator] who received postoperative WBI after BCS [numerator]; Patients with low - risk BC (grade 1, >65 years, ER positive, HER2 negative <2 cm, no lymphovascular invasion, unicentric, non - lobular) are excluded as they are considered possible candidates for no post-operative RT [exclusion]

BC: breast cancer; SLNB: sentinel lymph node biopsy; WBI: whole-breast irradiation; BCS: breast conservative surgery; ER: oestrogen receptor; HER2: Human epidermal growth factor receptor 2; *: the aspect 'maximum 8 weeks' can be assessed through supplementary sensitivity analyses.

Sources: Cardoso et al. (2019)³⁷¹; Rubio et al. (2024)³⁷²

OPERATIONALISATION OF QIS

In the following step, technical sheets are developed for each selected QI. Topics that can be elaborated in these sheets comprise:

- **Rationale:** the scientific evidence underlying the QI;
- **Type of QI:** structure, process or outcome;
- **Calculation:** concrete description of the numerator and denominator and time frames^{dd};
- **Target:** the target can be derived from the international literature (e.g. set by EUSOMA for breast cancer care, see Table 35) or through panel discussion and consensus procedures;

^{dd} The aspect 'time frames' can be illustrated with the following example: a QI on the multidisciplinary approach of breast cancer can be expressed as 'Proportion of women with breast cancer discussed at the multidisciplinary team meeting (MDT)'. Yet, for the operationalisation of the QI a time frame for the MDTs has to be defined. The numerator can e.g. be expressed as 'Number of women with breast cancer who were discussed during an MDT within 1 month before and 2 months after incidence date'.

- **Data sources** and time frames: e.g. IMA – AIM data 2017-2026, survey(s) developed for the study
- **Technical definitions**: provides details on – among others – disease classification, definition of incidence date, the codes of the fee schedule to select certain procedures (e.g. imaging, surgery, systemic treatment)
- **Risk adjustment**: which case-mix variables will be included in the analyses (especially for outcome indicators, rarely for process indicators);
- **Limitations**: e.g. proven under-registration of certain aspects of care may result in underestimation of related indicators;
- **Subgroup analyses**: e.g. by sex, age at diagnosis, incidence year;
- **Sensitivity analyses**: e.g. patients younger than a certain age.

All details with regard to operationalisation of the QIs must be thoroughly reviewed and approved by the involved clinical experts prior to initiating any quality indicator calculations.

PRACTICE TEST

QIs need to be **valid** and **reliable**, as well as **acceptable** and **feasible**.³³⁵ In order to test whether QIs fulfil these quality requirements, they should be tested in a pilot phase, before being used in practice. Especially measurability is frequently difficult to estimate in advance and often proves to be disappointing in practice.

When several QIs turn out as not measurable, one should realise that relying only on a few process indicators (e.g. those for which data are readily available in administrative databases) carries the risk of distorting care provision towards a focus on measured/measurable areas of care while disregarding other (potentially more) important tasks that are harder to monitor.³³⁸

Again, it is important to document transparently on the changes made (and the rationale for doing so), as well as on the QIs that were adopted in the final list but eventually excluded due to measurability issues.

In section 9.4.2 we elaborate further on data for quality measurements.

IMPLEMENTATION

Although a wide variety of implementation strategies have been applied in quality measurement, they are not always integrated in the QI development process.³⁴⁷

At KCE, the final step in QI development is the calculation of the QIs.

9.4.2 Data for quality measurement and improvement

Operationalizing quality constructs through quality indicators requires good data. This requirement is further elaborated in this section.

9.4.2.1 Quality of the data

A principal element of any quantitative assessment is **data quality**. Indeed, the reliability of quality measures relates to the quality of the data on which they are based and the robustness of the methods used.³³⁸ More precisely, it has to be assessed whether the data are complete, valid, reliable and comprehensive, whether they have been collected in a timely manner and whether applied definitions are consistent (Textbox 25).



Textbox 25 – Challenges with regard to data quality

The challenges with regard to data quality can be classified into six key aspects:

- **Completeness**, i.e. ideally, the data should include all patients, however, this is rarely the case. If data are missing, this should occur randomly and only occasionally. If missing data are too common, the data become unfit for assessing quality;
- **Validity**, i.e. the data should accurately reflect reality and should be free of bias and other errors;
- **Reliability**, i.e. repeated measurements by different data collectors, judges, or instruments, at different times and places, provide similar results;
- **Comprehensiveness**, the used database(s) should contain all relevant variables needed for analysis, such as diagnosis and therapeutic codes, pharmaceuticals, details from pathology and surgery reports, etc;
- **Timeliness**, i.e. the data are available for use without considerable delay;
- **Consistency**, i.e. the data are based on agreed definitions, and described exhaustively and exclusively.

Sources: Mainz, 2003;³⁵⁹ Gurol-Urganci et al., 2017;³⁷³ Quentin et al., 2019.³³⁸

Regardless of the database that is used, when quality of care among hospitals, regions, or countries, is compared or when evolutions over time are assessed, **classification systems** and **registration routines** should be **uniform**.³⁷⁴ This challenge affects both the development of oncological quality indicators, given the frequent updates to the TNM classification^{ee}, and the development of quality indicators for conditions like endometriosis, which still lack a validated staging system.³⁷⁵

But also, as discussed supra, when outcomes (and certain processes) are compared between providers (e.g. hospitals, health care professionals) important **confounders** (in particular 'case-mix variables') should be taken into account. For instance, the severity of disease (e.g. stage information in oncology, superficial vs. deep endometriosis), but also other coexisting conditions (e.g. diabetes) and/or lifestyle factors (e.g. smoking, alcohol abuse) and/or socio-economic background may influence the outcomes of care. Therefore, these factors should be documented so that they can be included in the analyses (i.e. 'risk adjustment' or 'case-mix adjustment'). Lastly, not all case-mix variables are documented and/or easy retrievable in databases (e.g. consumption of tobacco and alcohol). This aspect should be incorporated into the discussion of limitations.

9.4.2.2 Data sources

Please note that data requests require a substantial processing duration (see Chapter 11); therefore, initiating the procedure early is strongly advised.

ADMINISTRATIVE DATABASES

Various data sources can be used for quality measurement. **Administrative data**, which are routinely collected, such as **health insurance data** (in Belgium e.g. data from the Intermutualistic Agency, IMA – AIM) and **hospital discharge data** (MZG – RHM) can provide valuable data for process indicators. They have the advantage that they provide a good coverage of the population (i.e.

^{ee} The TNM classification, which stands for Tumour, Node, and Metastasis, is a globally recognised system to describe the size and spread of solid cancer tumours.

limited selection bias), they are readily available and easily accessible in electronic form, and no additional efforts are required to collect them.

Yet, fee schedules may be vague, as they are primarily intended for administrative and management purposes and not for quality assessment.^{337, 346} In addition, codes may be 'used' by healthcare professionals for other procedures than the ones intended ('optimisation'), which may result in the over- or underestimation of some procedures.³⁷⁶ But also, administrative databases often lack important information for quality measurement, such as patient reported outcomes (PROMs) and experiences (PREMs), important case-mix variables (cf. supra) or the extent of the multidisciplinary approach.^{337, 346}

Sometimes the original quality indicator needs to be **adapted** (slightly) so that the result can be calculated with administrative data.

For instance, for the Belgian outpatient setting, there are only data publicly available on the pharmaceuticals that are purchased, not on the number of prescriptions nor on the effective intake by the patient. As a consequence, a process indicator like 'proportion of adolescents with severe dysmenorrhea and/or endometriosis-associated pain, who are prescribed hormonal contraceptives or progestogens as first line hormone therapy' will need to be re-phrased as 'proportion of adolescents with severe dysmenorrhea and/or endometriosis-associated pain, who purchased hormonal contraceptives or progestogens as first line hormone therapy'. But then the question raises: does this rephrased indicator still denote quality? The rephrased quality indicator does not only assess the knowledge and prescription behaviour of the healthcare provider, but also, to some extent, the patient's (and parents') preferences and initiative to purchase hormone therapy.³⁷⁷

REGISTRIES

Disease-specific registries may also provide useful information for the calculation of quality indicator results. For example, to calculate the indicator "Proportion of patients with head and neck squamous cell carcinoma who underwent FDG-PET/(CT)^r before start of treatment", data for the denominator (i.e. the number of patients with head and neck squamous cell carcinoma) can be derived from the Belgian Cancer Registry, which is a population-based registry. However, registries are not available for all pathologies. They require significant investment in establishment and maintenance.³³⁸ Still, it has been suggested that it is possible to generate and feed disease-specific population registries from electronic health records,³⁷⁸ which may significantly reduce the costs of data collection for registries.

Before using data collected in a registry for quality assessment, its **completeness** should be assessed. For instance, if data collection is performed on a voluntary basis, some healthcare professionals may opt to only register a subgroup of patients, which may have a major impact on the calculation of certain quality indicators, e.g. those on patient outcomes or on adverse events.³⁷⁹ Likewise, if registries are limited to inpatient data (such as the Swedish National Inpatient Register),³⁸⁰ they will for instance not include patients (e.g. with endometriosis) whose treatment was (so far) restricted to the outpatient setting (e.g. patients treated solely with analgesia or hormone treatment), hence precluding the calculation of certain quality indicators.

For Flanders, for instance, QIs have been developed based on the Intego database, a **primary care registry** based on which disease trends in general practice can be mapped.³⁸¹

OTHER SOURCES

Other data sources which can (potentially) be used for quality measurement include **medical records** (which requires – among others - ethical approval, patient consent or legal waivers, and data de-identification), **surveys** among patients (e.g. to collect PROMS and PREMS) and healthcare providers, **interviews** and direct **observation** of healthcare providers' activities (e.g. hand hygiene).



9.4.3 Quality requirements for QIs

There are several guidelines and criteria available for evaluating the quality of quality indicators (see Textbox 26),³³⁸ yet there is no international consensus on characteristics of good quality indicators.³⁸²

Textbox 26 – Guidance and criteria for assessing quality indicators' quality

Organisation/institution/authors (date)	Content
OECD (2006)	Health Care Quality Indicators Project: Conceptual framework paper, which includes a list of criteria for the selection of quality indicators. ³⁸³
Federal Office for Quality Assurance (BQS, Germany; 2007)	QUALIFY – a tool for assessing quality indicators. ³⁸⁴
de Koning, et al. (2007)	The Appraisal of Indicators through Research and Evaluation (AIRE) instrument, assesses whether key methodological components and supporting information are clearly documented to evaluate the methodological quality of QIs. It comprises 20 items across 4 domains: (1) purpose, relevance, and organizational context; (2) stakeholder involvement; (3) scientific evidence; and (4) additional evidence, formulation, and usage. ³⁸⁵
Agency for Healthcare Research and Quality (AHRQ, 2011)	Key questions to inform the development of quality indicators. ³⁸⁶
Guidelines International Network (GIN, 2016)	Reporting standards for guideline-based performance measures. ³⁸⁷
International Society for Quality in Health Care External Evaluation Association (ISQua EEA, 2022)	Six-steps guidance on QI development. ³⁸⁸

Note: This selection of guidance documents is not meant to be exhaustive.

Kötter et al. (2012) further add that QIs also need to be **reliable**, **easily understandable** for providers and patients, **changeable** by behaviour (or sensitive to change), **achievable**, and **measurable** with high validity.³⁴⁷ Becker and colleagues (2019) further add that QIs should be **strongly correlated** with the quality of care provided.³³⁴ In addition, QIs should be able to **highlight relevant differences** in the quality of care between healthcare providers, organisations/hospitals or time periods^{ff}. QIs for which virtually all providers achieve maximum scores can still be useful for monitoring purposes, but often offer limited added value for quality improvement.³⁶⁵

9.4.4 Recommended set of indicators

Healthcare provision is complex, while process indicators usually focus only on narrow procedures in well-defined (sub)groups of patients.³³⁸ Therefore, hundreds of indicators are needed to enable a comprehensive assessment of the quality of care provided by a hospital or multidisciplinary team. Yet, reality forces us to scale that number back to realistic proportions.

But **how many QIs** should we include? Piggott and colleagues (2023) propose '*no more than ten quality indicators as reasonable for a moderate-sized guideline/QA scheme*',³⁸² while Wollersheim et

^{ff} Based on the literature, it is unclear at what level the distinctive character is deemed sufficient.

al. (2007) recommend based on their experience ‘*selecting about 12 clinical indicators for a care process to achieve a good balance*’.³⁶⁴ We should realise that if we put too much emphasis on minimising the recording efforts⁹⁹, and hence only a few QIs are selected, this restricts the evaluation to limited facets of the care process.³⁶⁴ In addition, the ideal balance between structure, process and outcome indicators in quality measurement is yet to be determined.^{389, 390}

Although guidance is lacking on desirable properties of indicator sets,³⁹⁰ in the QI development phase, we should keep the following important **threats** to the content validity of an indicator set in mind:

- omission of relevant QIs,
- overrepresentation of QIs for some aspects of care and
- inclusion of irrelevant QIs.³⁹⁰

9.4.5 Composite indicators

And how should patients, policymakers, purchasers and the general public handle the abundance of information when the results of a set of indicators are available? How should they select the best providers (e.g. the best hospitals)? How should they identify good-quality providers when the results from different indicators are conflicting? How should they estimate whether improvements have been made in the quality of care provided?

It may be tempting to base one’s decision on only a few selected indicators, e.g. those that you understand, although these may not be the most important ones, while crucial insights from other (potentially more) relevant indicators will be discounted. To address these challenges, numerous quality assessment frameworks have developed methodologies for aggregating different indicators into composite indicators or composite scores.³⁹¹ They allow the aggregation of different aspects of quality into one measure to give a clearer picture of the overall quality of healthcare providers.

The literature offers no standardised protocol for the proper construction of a composite indicator.³³⁸ However, what is important is that indicator construction is transparent, with clear explanations of methodological choices and their underlying rationales to facilitate understanding. In addition, as varying choices will generate distinct incentives for improvement, these must be carefully evaluated during composite construction.

The interested reader is referred to Quentin et al. (2019), section 3.6, where the methodological choices to be made when developing composite indicators, as well as their strengths and weaknesses are further elaborated.³³⁸

9.4.6 Sources of (validated) quality indicators

As explained supra, evidence-based clinical practice guidelines are relevant sources for generating QIs. ‘Guideline based QIs’ can either be generated from already available guidelines or developed during the guideline development process.^{334, 347}

In addition, a wide range of QIs exists that can serve as basis for the development of new quality measurement initiatives.³³⁸ In Table 36 a non-exhaustive list is provided for inspiration.

Importantly, existing QIs should be tested for relevance, validity, and reliability and there may be a need to adjust them to the Belgian context.³⁶⁴

⁹⁹ Of note, recording efforts can be reduced to a minimum when data extraction is automated (e.g. by extracting data automatically from electronic medical records), but this should not lead to an unmanageable or excessive proliferation of QIs.

**Table 36 – International databases and sources of existing QIs**

Organisation/institution	www
Agency for Healthcare Research and Quality (AHRQ)	https://www.ahrq.gov/
Australian government. Department of Health, Disability and Ageing	https://www.health.gov.au
Australian Commission on Safety and Quality in Health Care (ACSQHC)	https://www.safetyandquality.gov.au/
Canadian Institute for Health Information	https://www.cihi.ca/en/access-data-and-reports/indicator-library
International Consortium for Health Outcomes Measurement	https://www.ichom.org/
Institute for Quality Assurance and Transparency in Healthcare (Institut für Qualitätssicherung und Transparenz im Gesundheitswesen, IQTIG)	https://iqtig.org/
National Institute for Health and Care Excellence (NICE)	https://www.nice.org.uk/what-nice-does/standards-and-indicators
National Quality Forum (NQF)	https://psnet.ahrq.gov/issue/national-quality-forum
RAND	http://www.rand.org

Note: This selection is intended to be illustrative rather than exhaustive.

9.5 Adaptation and retirement of QIs

QIs may not be used for ever. A QI should be adapted or even retired if, for instance, the original quality gap has disappeared, in case new evidence emerges that would change the recommendation on the use of a quality indicator, but also when measurement of the QI is associated with unintended consequences or harm.^{387, 392, 393}

9.6 GRADE initiative on the integration of guideline and QI development

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group observed that while QI development is a high priority, progress in research methods is very slow and their development is rarely linked to guideline development.^{350, 393} Furthermore, they noted considerable inconsistency in the way terminology and definitions are used in quality assurance (QA).³⁵⁰

At the invitation of the European Commission Joint Research Centre (JRC), a multidisciplinary group of experts from key organizations representing the fields of guideline and QA development, has built a methodological framework that integrates guideline recommendations and QA.³⁵⁰ To that end, the Guidelines International Network (GIN)–McMaster guideline development checklist^{hh} (GDC) has been amended with forty additional items that will prompt guideline developers to consider factors important to quality assurance and improvement reporting and development standards from the outset of guideline planning.³⁹⁴

The interested reader is referred to <https://macgrade.mcmaster.ca/resources/gin-mcmaster-checklist-extension-for-quality-assurance-and-quality-improvement/> and to Piggott et al. (2022 & 2023).^{382, 394}

^{hh} Available at <https://heigrade.mcmaster.ca/guidelinedevelopment/using-checklist>.

10 SYSTEMS THINKING AND COMPLEXITY SCIENCE (STCS) INFORMED EVALUATION

KEY POINTS

- Systems thinking is an **approach to complexity** that enables us to identify the root causes of chronic or persistent problems and formulate more satisfying, sustainable and long-term solutions.
- Whilst complex interventions are often considered to have multiple objectives, strategies and components, implemented across multiple sites by multiple actors, the use of complexity linked to interventions could also be understood as interventions which are implemented in **complex systems**.
- A system is '***an interconnected set of elements that is coherently organised in a way that achieves something***'. It consists of elements, interconnections and a function or purpose. A system as a whole cannot be understood by analysis of its separate parts. Likewise, a (complex) intervention cannot be reduced to its individual parts.
- Several authors came up with **checklists**, guides and frameworks to summarise how systems thinking and complexity science (STCS) can be integrated in evaluation studies.
- The **system evaluation theory** (SET) is the most recent and most detailed attempt to provide a step-by-step approach to evaluate (complex) interventions. We propose the development of an STCS-informed programme theory as amendment to the SET.

10.1 Introduction

Evaluating complex interventions often will require approaches that adequately capture the dynamic and interrelated nature of the systems in which these interventions operate. In this chapter, systems thinking and complexity science (STCS) are introduced as complementary perspectives that can enrich evaluation by explicitly incorporating a system's lens. Rather than replacing established evaluation approaches, STCS helps to broaden the analytical perspective, particularly in situations where interdependencies, context and adaptation over time play an important role.

STCS is not a single coherent body of thought, although there is a common set of concepts and principles about how complex systems behave and evolve over time.^{397, 398} An increasing literature has considered the implications of STCS for evaluation research.

The underlying logic of the flow of this chapter builds on the sequence WHY – WHAT – HOW: the chapter starts with systems thinking as a way to approach complexity, linked to several types of systems thinking (hard and soft), and the added value of systems thinking (WHY). This is followed by a short introduction into the basic concepts and principles of STCS and its integration in evaluation research (WHAT). Next, we present a method for STCS-informed evaluation at KCE (HOW). For the interested reader who wants to delve somewhat deeper into systems thinking, we present three systems thinking approaches at the end of the chapter: system dynamics (SD), the Viable System Model (VSM) and Soft Systems Methodology (SSM). However, this trio is not an exhaustive representation of all systems approaches. Other interesting systems thinking methodologies, not described in this chapter, are for example critical systems thinking, complexity science and developmental evaluation.



10.2 Systems thinking as a way to approach complexity

In the course of intellectual history, there has been a paradigm shift regarding the management of complexity. The conviction that there is a pattern or order hidden in apparent chaos or disorder was challenged by the idea that complexity is a web of interwoven events, interactions, feedbacks and coincidences, inevitably involving uncertainty.³⁹⁹

The origins of contemporary systems thinking can be situated around the turn of the 20th century, when people started thinking that systems could not be understood by analysing the component parts. By the end of the 20th century, complexity science provided new insights regarding the non-linearity and self-organising abilities of complex physical, biological and social systems.^{400, 401}

Horizontal complexity, represented by networks of cause and effect relationships, became the domain of system dynamics (see section 10.6.1). System dynamics understand the behaviour of a system as the result of cause and effect relationships, feedback loops and delays between elements of a system.

In addition to the horizontal axis of complexity, the vertical axis introduces hierarchy as an organising principle. One of the metaphors that is often used in systems thinking is the iceberg, illustrating that observable events are the manifestation of the patterns and structures that are below the surface of the water and cannot be observed directly.^{399, 402} The viable systems model (VSM) (see section 10.6.2) is an approach incorporating the hierarchy principle, together with the idea of recursiveness, as we can find in fractals (e.g. blood vessels, Romanesco broccoli, lightning bolts), i.e. geometric shapes with repeating patterns that look the same at different scales.

While the previous systems approaches are labelled as 'hard' systems thinking, 'soft' systems thinking introduced the idea that our observations are always filtered by our world view, hence we only have a partial picture of reality. Soft systems methodology (SSM) (see section 10.6.3) tries to align the partial pictures that people have about a problem to be able to take action and improve problematic situations.⁴⁰³

Finally, relevant for the understanding of the effect interventions have or refrain from having is the concept of emergence, which refers to the macro level manifestation of micro level behaviour.⁴⁰¹ This is a characteristic for complex adaptive systems (CAS), the study subject of complexity science. An example of CAS is a flock of starlings: the synchronised movement of thousands of birds that is made possible by a limited number of simple rules each bird obeys to. Hence the macro movement of the entire flock emerges from the micro level behaviour of each individual bird. Interestingly, emergence can be created in a strategically desired direction in social systems.³⁹⁹ The introduction of an intervention could be a way to do so.

10.3 The added value of systems thinking

10.3.1 Systems thinking as a way of looking

Systems thinking is not a tool or method, not even a methodology, but an approach, a way of looking that can be applied to every domain, including the evaluation of interventions. STCS is the grouping of various approaches (e.g. system dynamics, viable systems model, soft systems methodology) that developed over time. Each perspective offers a specific methodology, which can be added to the toolbox of evaluation research, depending on the type of intervention, context variables and the degree of complexity.

10.3.2 Systems thinking as a diagnostic tool

Systems thinking is also a diagnostic tool for examining problems more completely and accurately before acting, because it allows us to ask better questions. It involves the identification of patterns and underlying structures – so called 'root causes' – that drive events or behaviour. This can lead to the revealing of hidden links, prevent unintended consequences and more satisfying, long-term solutions

to chronic or persistent problems. By helping us to articulate problems in new ways, systems thinking expands the range of solutions to a problem.⁴⁰⁴

10.4 Unpacking systems thinking and complexity science

10.4.1 What is complexity?

In section 10.2, we already introduced the concept of complexity. We referred to its horizontal (cause-and effect relationships) and vertical (hierarchy) axis, and related concepts such as emergence. In this part we will explain 'complex' by contrasting it to 'simple' and 'complicated'.

Systems approaches aim to simplify the process of our thinking about and managing of complex realities that have been described as messes (cf. Russell Ackoff), complex or wicked problems (cf. Horst Rittel).⁴⁰⁵ Hence, systems thinking is closely linked to the paradigm of complexity. In Table 37, we compare three degrees of complexity relative to social problems to clarify the difference between simple, complicated and complex problems.

Table 37 – Degrees of complexity

Simple problem	Complicated problem	Complex or wicked problem
Clear cause-effect relationships	Clear cause-effect relationships	No clear cause-effect relationships. Instead: • Disproportionality • Time delay • Emergence
Standardised solutions (e.g. recipe or procedure)	No straightforward solutions	Unclear what solutions could be or what success means
Basic skills to tackle the problem	Certain degree of expertise needed to solve the problem	High uncertainty about the outcome + one shot
High degree of predictability	Predictable if a number of conditions are fulfilled	Unpredictable: no rules to guarantee success
Example: to bake a cake	Example: to launch a rocket	Example: to raise a child

For complex problems (e.g. raising a child) there are no clear rules to follow to guarantee success. It is not even agreed upon what success means, since there is no definitive formulation of a complex problem. There are no cookbook recipes to solve complex problems and uncertainty about the outcomes is high, predictability is low. Solutions to complex problems are not true or false, but rather better or worse. Every solution to a complex problem is a one-shot operation, because there is no opportunity to learn by trial and error. Every complex problem can be considered a symptom of another problem.⁴⁰⁶

There is **no complete agreement** on all aspects and definitions of complex systems. It is a matter of degree or perspective, rather than a clear distinction, as to what is complex and what is not.⁴⁰⁷

10.4.2 What is a system?

To understand systems thinking an essential first step is to know what systems are. Meadows defines a system as '*an interconnected set of elements that is coherently organised in a way that achieves something*' (p. 11).³⁹⁵ There are, however, many definitions of a system. Those definitions are consistent in their reference to elements, interconnections and a function or purpose, which are three basic characteristics of systems.

- **Elements** are parts or units, mostly visible tangible things and therefore the easiest to notice. A system part is in fact anything needed to make a system work.⁴⁰⁸
- **Interconnections** are the relationships that hold the elements together. System elements don't just influence each other, they need each other to function, they are interdependent.⁴⁰⁹



- **Purposes** are deduced from the behaviour of a system. '*System elements work together to create 'something' which they cannot achieve independently*' (p. 36).⁴⁰⁸ This 'something' can be different things (e.g. an unintended consequence), but one of them is the purpose of the system.

A definition that more accurately reflects the importance of interdependence and emergence is '*an integrated whole whose essential properties arise from the relationship [interdependence] between its parts*' (p. 142).⁴¹⁰

Intervention programmes consist of interrelated parts that together produce an effect that they would not be able to produce on their own and can therefore be considered as a system.

10.4.3 Premisses of systems thinking and complexity science

In addition to the system characteristics outlined above, there are several premisses or basic rules which are typically true for systems.

▪ The whole is greater than the sum of the parts

Central to a systems perspective is holistic thinking, as opposite of reductionist thinkingⁱⁱ. A key understanding within systems thinking is that a system as a whole cannot be understood by analysis of its separate parts.³⁹⁶ The functions and meanings of the parts are lost when separated from the whole. Reynolds and Holwell explained that '*if one considers a situation as a whole, rather than focusing on its component parts, then there are properties which can be observed which cannot be found simply from the properties of the component parts*' (p.8).⁴⁰⁵

An intervention programme cannot be reduced to individual parts. A systems framework invites us to understand the intervention programme in relation to other programmes and as part of a larger context.

▪ Parts are interdependent and the focus is on interconnected relationships

The elements of a system are often the easiest parts to notice, because many of them are visible, tangible things. You can divide elements in sub elements and then in sub-sub-elements, but soon you lose sight of the system. It is a good idea to stop dissecting elements and to start looking for the interconnections, the relationships that hold the elements together.

The change in perspective that comes with systems thinking focuses our attention on how the web of relationships function together rather than as a linear chain of causes and effects. A change in one part has implications for all parts and their interrelationships.

▪ Every system has its own system purposes

Information-based relationships are hard to see, but functions or purposes even harder. A system's function or purpose is not necessarily spoken, written or expressed explicitly, except through the operation of the system. The best way to deduce the system's purpose is to observe it for a while to see how the system behaves. Important: purposes are deduced from behaviour, not from rhetoric or stated goals.

Changing elements usually has the least effect on the system (e.g. our body replaces cells constantly, it remains the same body), as long as its interconnections and purposes remain intact. If the interconnections change, the system may be greatly altered (e.g. change the rules of a football game and you get a different game). The same accounts for changing the purpose (e.g. changing the purpose of the football game from winning into losing). A change in purpose changes a system profoundly, even if every element and interconnection remains the same.

ⁱⁱ In reductionist thinking the whole is understood through the careful study of its parts.

- **Systems are made up of subsystems and function within larger systems**

Systems can be embedded in systems which are embedded in yet other systems and out of one system other completely new, never-before-imagined systems can arise. Systems can be nested within systems and at the same time be part of a larger whole. This is a recurrent structure of repeated shapes that are similar in shape across different scales. The same structure becomes smaller and smaller or larger and larger each time it is repeated.

How subsystems function within larger systems and how larger systems connect to and are influenced by subsystems can be part of a systems inquiry into understanding a programme and its effects. E.g. the content and processes of a prenatal programme for pregnant teenagers will be affected by larger societal norms. That's why the programme in France, Burkina Faso and Mexico will be different, even if based on the same model.

- **Systems cause to a large extent their own behaviour and can be self-organising**

A system is a set of things, like people, cells or molecules, that are interconnected in such a way that they produce their own pattern of behaviour over time. This means that the system may be driven by outside forces, but the system's response to these forces is characteristic of the system itself. To explain the principle that a system causes its own behaviour, Meadows³⁹⁵ uses a slinky as a metaphor in comparison to a box. If you hold a slinky in one hand, it drops down if you pull away the hand that supports it. If you do the same thing with a box, it does not drop down but keeps its shape. The pulling away of the hand is the same outside force, but the reaction of the slinky and the box are different. So, the systems (box and slinky) cause their own behaviour in reaction to the same outside force. An outside event may unleash the behaviour, but the same outside event applied to a different system is likely to produce a different result. Often systems are self-organising and self-repairing over at least some range of disruptions.

For intervention programmes this means that their response to an outside force is characteristic for each intervention programme. In other words, different programmes will respond differently to the same outside force, for example financial cutbacks.

- **System boundaries are necessary and inevitably arbitrary**

Systems exist within an environment and are separated from its environment by means of a boundary. The boundary of a system refers to which parts to include and which to leave out. This decision is a matter of judgement and experience. Manageability is set out against completeness. Usually there is not one right answer, but a whole range of possible answers depending on the aim and context variables.⁴¹¹ Systems are social constructions. What to include in a system and where to draw the boundaries are matters of utility. Including too much makes the system overwhelming. Including too little risks missing important elements that affect programme processes and outcomes. The dilemma is how much to simplify reality. The challenge is to construct simplifications that pass the dual tests of usefulness and accuracy. Defining boundaries is an essential part of systems thinking.

10.5 Applying systems thinking and complexity science to evaluation research

10.5.1 How to recognise whether systems thinking is the best fit for evaluating an intervention?

Emily Gates claims that STCS '*is relevant to and has value for any evaluation purpose and context*'.⁴¹² Still, we think it is useful to disentangle where exactly STCS can be of added value in evaluating (complex) interventions. In what follows we will first argue that a complex intervention can be defined as a system and we will situate STCS in the MRC framework to clarify the parts for which STCS can be particularly valuable even if the intervention under study cannot be considered a system.



10.5.1.1 Complex interventions acting as a system

About interventions, two questions arise:

- Are they complex?
- Do they act as a system?

About complexity, Skivington et al. argue that an *'intervention might be considered complex because of properties of the intervention itself, such as the number of components involved, the range of behaviours targeted, the expertise and skills required by those delivering and receiving the intervention, the number of groups, settings, or levels targeted, or the permitted level of flexibility of the intervention or its component'* (see Introduction Chapter).⁹ However, whilst complex interventions are often considered to have multiple objectives, strategies and components, implemented across multiple sites by multiple actors, the use of complexity linked to interventions could also be understood as interventions which are implemented in complex systems (see section 10.5.1.2). By consequence, a focus on the complexity of systems implies that simple interventions, as well as complicated interventions, may be subject to STCS-informed evaluation.³⁹⁸ In other words, the complexity of the intervention as such is not a necessary prerequisite to look at it from a STCS perspective.

In addition to the complexity of the intervention, we could determine whether it acts as a system. Arnold and Wade claim it should have a clear purpose (emergent property) and the parts should work toward that purpose (interdependence).⁴¹³ Meadows proposes three questions to differentiate whether you are dealing with a system:³⁹⁵

- Can you identify parts?
- Do the parts affect each other?
- Do the parts together produce an effect different from the effect of each part on its own?

Renger points to the 5 C's i.e. cooperation, coalition, collaboration, collectiveness and coordination that signal you might be dealing with a system, because these keywords often indicate many interdependent parts which work together to achieve something greater than they could independently. When looking at complex interventions through the STCS lens, it is likely that they can be defined as systems, act as systems and should be treated as systems.⁴⁰⁸

If the intervention is likely to be a complex intervention acting as a system, the system evaluation theory (SET) of Ralf Renger provides a step by step approach to evaluate the intervention.⁴⁰⁸ We outline SET in section 10.5.2.2.

When designing an evaluation study, it is useful to explore the complexity and 'system-ness' of the interventions under study to know how to proceed with the evaluation informed by STCS. If this mapping of the intervention concludes that it is not a complex intervention that acts as a system, STCS might still be valuable. In the next part, we refer to the MRC framework to clarify this and to show where else STCS comes in.

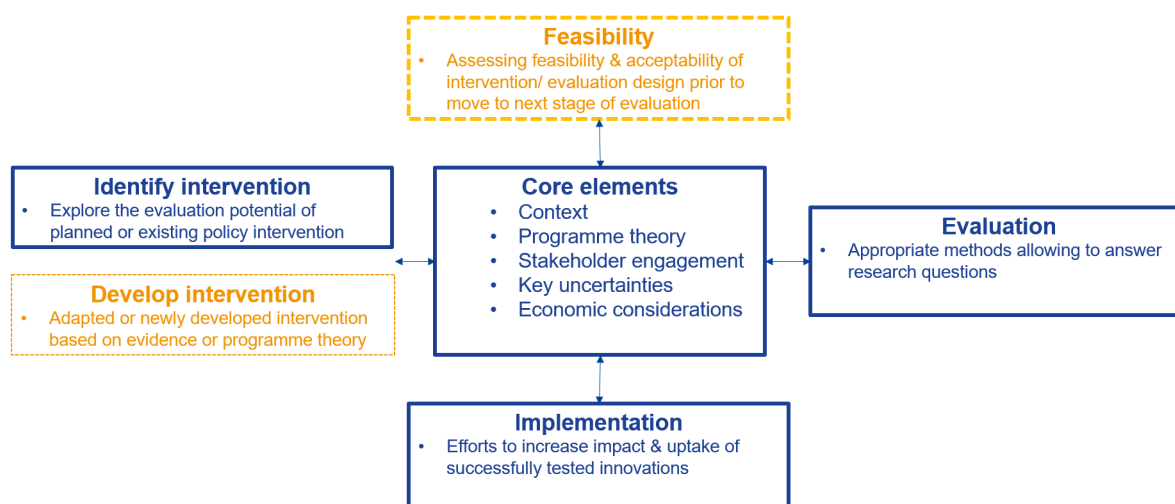
10.5.1.2 The MRC framework seen through a system lens

In addition to the complexity of the intervention and of the problem the intervention wants to tackle (e.g. antibiotics consumption, eating disorders), with reference to the MRC framework, STCS offers a particular lens to look at the **core elements** (see Figure 34), such as context and programme theory. The **context** in which the intervention is implemented can consist of one or more complex systems. In addition, complexity may arise from the interaction between the intervention and its context (see Chapter 5). The mechanisms of impact may include feedback loops and time delays, as reflected in the **programme theory**.

In addition, STCS is mostly suited to be integrated in the **evaluation phase** of the MRC framework (Figure 34). The binary effectiveness question (i.e. whether the intervention works or not) is transcended and system level outcomes including system change are involved.

In the following paragraphs, we elaborate on each of the areas in which complexity may come into play, starting with the context in which an intervention is implemented.

Figure 34 – MRC model seen through an STCS lens



Source: Adapted from Skivington et al. (2021; 2024)^{9, 10}; Blue: included in the current process note; orange: not part of the process note.

10.5.1.3 The context

The context may consist of one or more complex systems and/or complexity may arise through interactions between the intervention and its context.

An intervention, simple or complex, may be embedded in the context of a complex system (e.g. healthcare system, educational system, social welfare system), with which it interacts or by which the intervention is influenced. Properties of the system may determine whether the intervention has the intended effects.

Thinking about the context as a system helps to understand the interaction between an intervention and its context.⁹ STCS-informed evaluation takes into account the complexity that arises both from the intervention's components and from its interaction with the context in which it is being implemented.⁹

10.5.1.4 Programme theory

Mechanisms of impact describe how the delivered intervention produces change. Exploring these mechanisms is crucial to understanding both how the effects of the specific intervention occurred and how these effects might be replicated by similar future interventions.²⁵ These effects are hypothesised in the programme theory of the intervention, which describes how an intervention is expected to lead to its effects and under what conditions (see Chapter 1).⁹

The mechanisms of impact alike the programme theory may – instead of being a linear logic – contain system elements such as feedback loops and time delays. The mechanisms of impact may appear to be more complex than the programme theory hypothesised. They may for example act as a system archetype, i.e. a common pattern of (problematic) system behaviour, such as addiction, escalation, drift to low performance, fixes that backfire. More information on system archetypes can be found in the subsection on system dynamics (section 10.6.1) and more extensively in the books Thinking in



Systems by Meadows³⁹⁵ and the Fifth Discipline Fieldbook by Peter Senge et al.⁴⁰² Recognising a system archetype can be a powerful diagnostic tool to understand the mechanisms of impact or lack of impact, hence build a programme theory.

10.5.1.5 *Interest in system-level changes in the evaluation phase*

Interventions may aim at outcomes **transcending the binary effectiveness question** (i.e. whether the intervention works) and involving system level outcomes including system change. Outcome measures could capture changes to a system rather than changes in individuals, such as changes in relationships between actors within the healthcare system, or the normalisation or discouragement of practices (e.g. prescription of antibiotics). System level outcomes include how changing the dynamics of one part of a system alters behaviours in other parts (e.g. displacement of smoking into the home after a public smoking ban).⁹

Note that it is a misconception that systems approaches are more appropriate for large-scale, longer-term evaluations. Like evaluation, **systems approaches are scalable**. They can be applied to big and small issues, over short and long timeframes. The key is the nature of the situation, not the scale of the situation.^{408, 414}

10.5.2 *How to apply systems thinking to evaluate (complex) interventions?*

So far, we identified where STCS can be of added value in evaluating (complex) interventions. We described how to recognise complex interventions acting as systems and linked STCS with the MRC framework.

In the next step we describe how STCS can be applied to evaluate (complex) interventions.

10.5.2.1 *Frameworks to help evaluators integrate systems thinking into evaluation studies*

Plenty of authors have reviewed the literature to summarise how STCS can be integrated in evaluation studies. They came up with checklists, guides and frameworks. Examples are Westhorp's complexity checklist⁴¹⁵, a review of Walton³⁹⁷ providing 8 guiding themes for applying a complexity frame of reference to evaluation, a framework for qualitative complex system process evaluations developed by McGill et al.⁴¹⁶, the STCS-informed and principles-based evaluation framework developed by Bustamente et al.⁴¹⁷, the propositions for evaluating complexity from Preskill and Gopal and the principles for effective use of systems thinking in evaluation developed by the Systems in Evaluation Topical Interest Group of the American Evaluation Association (SETIG).⁴¹⁸

Each framework is explicitly or implicitly built around the four core principles of systems thinking:

- interrelationships between elements of a system,
- perspectives from which a situation or system can be understood,
- boundaries between the system and its environment,
- the dynamics that influence and are influenced by the behaviour of the system over time and at different scales.

Although these principles can function as general guidelines to keep in mind when starting an STCS-informed evaluation, we searched for more concrete hand-on instructions. The most recent and most detailed guidebook is System Evaluation Theory (SET). This is a blueprint for practitioners evaluating complex interventions operating and functioning as systems, written by Ralf Renger.⁴⁰⁸

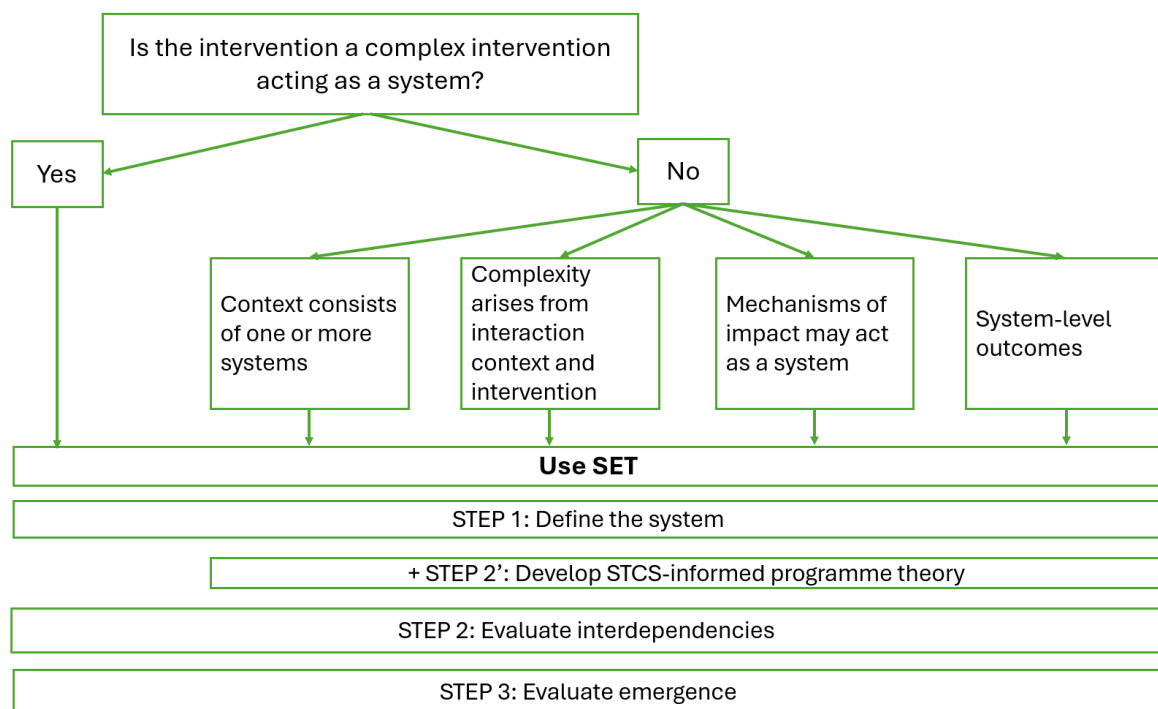
10.5.2.2 *System evaluation theory (SET)*

SET has been developed for complex interventions acting as a system. However, we propose implementing it also in case complexity and 'system-ness' are situated in other parts of the MRC framework.

Note also that STCS should be integrated at the beginning of an evaluation design in order to inform the evaluation focus, sampling strategy, data collection methods, analysis and interpretation of findings.

SET is a three-step evaluation framework (see Figure 35) that examines how interdependent parts of a system work together, how efficiently they interact and whether those interactions produce effective system-level outcomes. In what follows we summarise SET, but for a more detailed presentation we refer to Renger's guidebook.⁴⁰⁸

Figure 35 – How to integrate STCS in evaluation



STEP 1: Define the complex intervention acting as a system, more specifically its boundaries, system parts, their interdependencies and the emergent system property. You need to know what to evaluate and how it is supposed to operate before you can evaluate whether it is operating as it should.⁴⁰⁸

This first step, the description of the system, results in a snapshot, a static system description aimed at gaining an initial understanding of the system, including the system structure, the boundaries, the constituent elements and the relationships between these at a given time point.⁴¹⁶ This description of the system should be developed in co-creation with stakeholders and inform the evaluation design, including the research questions.

It may be helpful to create a map of the system, based on stakeholder input. The output can be a qualitative description of the system structure, elements, boundaries, and relationships. Evaluators may begin to make sense of the local rules that govern the intervention in the system, including how different system elements relate to each other and how the intervention operates.⁴¹⁶

This first step is further specified in **seven operational steps**. We refer to the Renger's book for more detailed information on each of the steps.⁴⁰⁸



1. **Get leaders to the table.** The leaders are the ultimate decision-makers who control the money and decide whether to implement the evaluation recommendations. However, bear in mind where, when and by whom boundary judgements around a system of interest are made, whilst taking power dynamics into account. Power relations should be a focus of attention and form an explicit part of the next steps, particularly in situations where there is little common interest shared between stakeholders, or in the event of fundamental conflicts or unequal access to power resources. We refer to critical systems thinking for the readers who want to delve deeper into the issue of power in systems.⁴¹⁹
2. **Define the emergent system property.** The successful operation of a complex intervention depends on consensus among leaders and actors of what they are trying to accomplish (If everything in the system is working as it should, what would it produce?).
3. **Define the system parts.** Leaders should be asked in an open discussion to define the parts that are needed for the essential system property to emerge. Subject matter experts (SMEs) are needed who have lived experiences in the day-to-day system operations to list the system parts. The nominal group technique, i.e. a structured small group discussion to reach consensus, could be useful at this stage.
4. **Connect the parts.** This comes down to drawing the arrows between the parts and draw a dotted line around all the connected parts to depict the system boundary created for evaluation purposes. This way the boundaries evolve naturally by defining the system parts.
5. **Unpack the arrows** or define the system operating procedures (SOPs) or workflows. We want to know how an individual part operates in concert with other parts. This forms the standard against which to evaluate success. You need to know how system parts are intended to operate together, before you can evaluate whether they are in fact working as intended. If the SOPs are not predefined, process flow mapping (PFM) (i.e. a visual technique to represent a workflow from start to finish) can be used to map them (together with the SMEs for example through interviews). Note that the mere facilitation of the mapping process may lead to an improved SOP.
6. **Bring groups together to define joint system operating procedures.** At this point individual system parts gain an appreciation for each other and the impact their decisions have on the other parts and the emergent system property.
7. **Push the boundary envelope.** Findings are presented to the leaders to close the loop. As a result of defining the joint system operating procedures (JSOPs), it may be necessary to adjust the boundaries, either to include a missing system part and/or remove a redundancy.

Bear in mind that part of systems thinking involves making conscious decisions about timing. This is particularly important when findings need to be shared, as this is how information is fed into the system. By closing the loop, evaluation becomes an integral part of the system.

STEP 2: Evaluate the interdependencies or system efficiency, i.e. how effectively the components work together regarding information flow and feedback mechanisms, coordination and communication, resource sharing and bottlenecks and process weaknesses.

To evaluate system efficiency, compare how the system interdependencies are supposed to work (defined by the SOPs) with how they are actually executed. This can be done by means of direct observation (site visit), surveying system actors or analysing system data. Look for where information is exchanged (i.e. feedback loops), for possible processes that can lead to a domino effect of problems (i.e. cascading failure), and for places where efficiencies might be achieved by creating a bypass (i.e. reflex arc). To evaluate system interdependencies Renger focuses on these three system principles: feedback loops, cascading failures and reflex arcs.⁴⁰⁸

- **Feedback loops:** They can occur anywhere and anytime information is exchanged between system parts. We want to know whether the feedback information is credible, otherwise it will not be used for decision-making, whether the feedback is relevant, whether the feedback is timely, at

which rate it is provided and whether it is specific enough. Feedback loops can be mapped in causal loop diagrams.

- **Cascading failures:** These are processes in a system in which the failure of one or a few parts can trigger the failure of other parts.
- **Reflex arc:** This is a bypass in the system intended to improve efficiency. Note that an observed departure from the stated SOP might be an organically derived bypass to improve efficiency. For example, staff at the operational level may encounter implementation challenges for which they developed their own workaround. This is a reflex arc that has not yet been documented in the SOPs.

Regardless of the scale, level, or boundaries of the complex intervention acting as a system, the process for defining the system and evaluating efficiency is the same.

STEP 3: Evaluate system effectiveness, i.e. whether the system as a whole achieves its intended purpose, or what system actors at all levels are trying to accomplish together that they cannot do individually (i.e. emergent properties). Renger takes three considerations into account for evaluating a system's effectiveness:⁴⁰⁸

- **Avoid reductionist thinking, instead think synthesis.**
- **Consider the multidimensionality of the emergent system properties.**
- **Ascertain who has responsibility for evaluating the system's emergent properties**

In addition, Renger differentiates between weak and strong emergence.⁴⁰⁸ The first occurs when you can reason what might be expected when interdependent system parts interact. Strong emergence, on the contrary, is when you can't predict the emergent property from the parts. Unexpected emergent properties can easily escape the attention of researchers and require an open and critical approach on their part.

There can be more than one emergent property, but perhaps resources to evaluate them all will be lacking. Prioritisation might be necessary.

10.5.2.3 Development of an STCS-informed programme theory

The development of an STCS-informed programme theory is not included in SET. However, when an intervention is developed by others and no programme theory is specified, evaluators still need to theorise the intervention before evaluating it.⁹ An STCS-informed programme theory generates hypotheses about the ways a complex intervention that acts as a system may lead to change or about ways a simple intervention may lead to change within a complex system.

Models which leave out the implementation context, concurrent programmes, or characteristics of clients, may overstate the causal contribution of the intervention.⁴²⁰ In addition, they are more likely to present a single theory of change, rather than representing different stakeholders' views about what are desirable outcomes and how these might be achieved.⁴²⁰ The range of interventions suited for simple logic models (see Chapter 1) is described by Wholey as those where goals can be agreed and precisely quantified, where progress towards them can be reliably measured, and where both staff activities and the results of those activities can be readily observed.⁴²¹ Rogers argued also that in all other cases simple logic models can best be understood to be heuristic simplifications instead of accurate models.⁴²⁰

Therefore, we propose the development of an STCS-informed programme theory as amendment to SET. More specifically, we would add it as a second step after the description of the system (see Figure 35).

An STCS-informed programme theory should be sensitive for causal loops, tipping points (i.e. after having achieved a certain critical level a small additional effort can result in a disproportionately large effect), emergence (i.e. when specific outcomes emerge during implementation of an intervention),



non-linearity, interdependence and context. Thus, for complex interventions acting as a system or in case complexity and 'system-ness' are situated in other parts of the MRC framework, the theory of change needs to:

- be able to accommodate local adaptation and change⁴²⁰
- view the situation holistically as a set of interacting elements within an environment
- recognise that the relationships between elements are more important than the elements themselves in determining the behaviour of the system
- recognise a hierarchy of system levels
- take into account differing viewpoints of actors/stakeholders.⁴²²

Note that programme theories can change and evolve over the life of an intervention programme. After evaluation, the programme theory might be adapted in function of the findings, as is the case in the realist evaluation research cycle (see Chapter 6). The more volatile the environment of a programme, the more likely it is that the programme theory will be affected by that volatility.⁴²³

10.5.3 STCS-informed methodology

From a complex systems perspective, the system is taken as the starting point for the analysis in a process evaluation. Consideration is given to the ways in which the intervention may have both direct and more indirect effects, and to the ways in which that intervention may influence the behaviour of the system's components. An STCS-informed evaluation approach must, by its very nature, allow for adaptation and change, both in response to early evaluation findings and in response to changes in the intervention and the system.⁴¹⁶

A purely quantitative or qualitative approach is rarely adequate for complex intervention research. Mixed-methods designs, where quantitative and qualitative methods are combined, are necessary to answer questions beyond effectiveness. When systems thinking is the main lens through which a complex situation is approached, the evaluation toolbox should adapt accordingly.

STCS-informed evaluation includes the involvement of stakeholders with different viewpoints in the selection of adequate outcomes. In line with its pluralist perspective STCS-informed evaluation also favours study designs with multiple methods. This should lead to contextualised understandings of how an intervention induces change as well as the enablers and constraints across different contexts.⁹

10.6 A brief introduction of systems thinking approaches

We will present in what follows three systems thinking perspectives: system dynamics and the viable systems model, both from the hard systems tradition, and soft systems methodology (SSM) from the soft systems tradition. Each perspective has its own objectives and represents a different way to approach complexity.

10.6.1 System dynamics

System dynamics are situated within the hard systems thinking wave which was popular in 1950s-1960s and referred to as the first wave of systems thinking. With foundations in economic sciences, it assumed objectivity and therefore quantitative modelling was most used. Agreement was supposed among individuals within a system about a predetermined and unproblematic goal.⁴²⁴

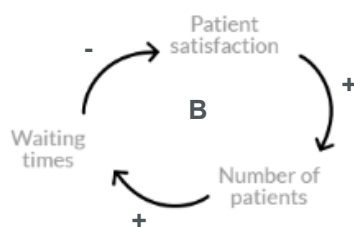
System dynamics are a toolbox to model the dynamics of complex systems, to solve complex problems, and are concerned with how and why systems change over time.^{425, 426} The main goal is to increase the understanding of the problem under study in order to – in a next step – intervene in a system and change it in the desired direction.⁴²⁶ Hence, it is not only a diagnostic tool, but it is also action-oriented. System dynamics can be used to evaluate complex intervention programmes and diagnose what goes well or wrong.

System conceptualisation encompasses what parts to include in the system, based on a hypothesis of what creates the observed behaviour. Often the modelling starts simple and small, and complexity is gradually added. The model's purpose defines what is inside the system, and what belongs to the environment. This model boundary should include all the components that are involved in generating the behaviour of interest. Of most interest is capturing the feedback loops, rather than the detailed specification of the individual parts. The focus should be on explaining the dynamics, not on explaining the components of the system.⁴²⁶

A model is most presented graphically by means of a causal-loop diagram. In causal loop diagrams, causal links are used to clarify how one variable influences another and are indicated with a single headed arrow with a sign (- or +) added. '+' means that all else equal, the two variables will follow the same trajectory over time, i.e. increases or decreases in the variable at the beginning of the arrow will result in increases or decreases in the variable at the end of the arrow. For '-' the opposite is true: increases in the variable at the beginning of the arrow will result in decreases in the variable at the end of the arrow.

When two variables are interconnected with causal links, it is called a causal loop. More precisely, a causal loop is a closed causal connection that forms a circular relationship among variables. In complex systems it is common to have multi-loop diagrams representing complex feedback structures. In a feedback loop every element is both 'cause' and 'effect'. For every variable you can trace links that represent influence on another element. This way cycles are revealed that repeat themselves. Causal loop diagrams can be made by means of Kumu.io or Vensim.com. Both are widely used tools in system dynamics and Vensim is specifically used to quantify causal loop diagrams. Figure 36 presents an example with increasing numbers of patients increasing waiting times in a clinic, and increasing waiting times leading to decreasing numbers of patients, leading to decreasing waiting times again, and so on.⁴⁰²

Figure 36 – Example of a feedback loop



There are basically two types of causal loops (also called loop polarities):

- **A reinforcing loop or positive feedback loop (R):** reinforcing cycles can act as engines of growth or engines of deterioration, in which the growth or collapse continues at an ever-increasing rate. A small change builds on itself, resulting in big changes after some time. There can be any number of elements in a reinforcing loop, all propelling each other's growth. Reinforcing loop situations generally 'snowball' into highly amplified growth or decline. Somewhere sometime the reinforcing loop will run up against at least one balancing mechanism that limits it. The letter R is used to mark a reinforcing loop.
- **Negative feedback loops or balancing loops (B)** tend to counter the original tendency and balance the result. As illustrated in Figure 36, they generate stability. Balancing processes generate the forces of resistance, which eventually limit growth. Balancing loops are found in situations which seem to be self-correcting and self-regulating. The letter B is used to mark a balancing loop.

In addition to feedback loops also time (II) needs to be considered. Both in reinforcing and balancing loops delays may occur. Delays are the points where the link takes a particularly long time to play out. Delays can have enormous influence in a system, frequently accentuating the impact of other forces.



When unacknowledged delays occur, people tend to react impatiently, usually redoubling their efforts to get what they want. This results in unnecessarily violent oscillations. One of the purposes of drawing systems diagrams is to flag the delays which you might otherwise miss.⁴⁰²

Of note, whether the loop goes in clockwise or counterclockwise direction is not important; only the loop polarity and causal signs matter.

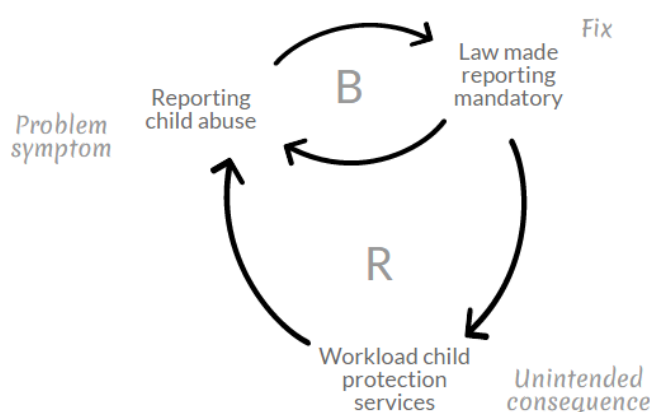
There are several causal loop templates describing commonly occurring system dynamics, called **system archetypes**. Once a specific archetype is recognised to be at work, the theory of that archetype can be used to expel a particular problem and work toward an intervention. To find out which archetype applies, a good strategy is to look at the problem situation through the lens of several different archetypes. Two or three may fit together, each highlighting a different aspect.

In what follows, one archetype (fixes that backfire) is briefly presented as an example. However, many more archetypes are described in Senge, P. et al. 1994, *The fifth discipline field book*, p. 125-150, Meadows, D. 2008, *Thinking in Systems*, p. 110 – 141, or Daniel H. Kim, 1992, *System archetypes I*, p 6-7.^{395, 402, 427}

The central theme of the **fixes-that-backfire archetype** is that almost any decision carries long-term and short-term consequences, and the two are often diametrically opposed (see Figure 37). A problem symptom cries out for resolution. A solution is quickly implemented (the fix) which alleviates the symptom (balancing loop), but the unintended consequences of the fix (reinforcing loop) actually worsen the performance or condition which we are attempting to correct.⁴⁰²

The fixes-that-backfire archetype can be illustrated with the following example: child abuse is underreported to authorities. In the US, they made reporting mandatory. However, child protection services were not reinforced, hence were overwhelmed by the number of reports, and could only investigate a small part of all reports. By consequence they got the reputation of being untrustworthy. In response, people decided not to report (although mandatory) and tried to find solutions themselves or did not do anything. Number of reports decreased again, hence the problem of under-detection was reinforced.⁴⁰²

Figure 37 – System dynamics model for “Fixes that backfire” – example



Causal loop diagrams can be quantified by means of stock-and-flow diagrams. Mathematical modelling can simulate the behaviour of complex systems and is therefore a vital research tool for exploring complex systems, such as the healthcare system.⁴²⁸ A stock and flow diagram describes the relationships between system variables using stocks, rates and influencing factors. A stock and flow diagram can be interpreted as mimicking the flow of water in and out of a bathtub. The rates control the quantity of water that enters or leaves the tub (i.e. the stock). The rate changes over time depending on the constraints of conditions that are placed on the system.^{395, 428}

Further reading:

Meadows, D. H. (2008). Thinking in systems. A primer. Chelsea Green Publishing, Vermont.³⁹⁵

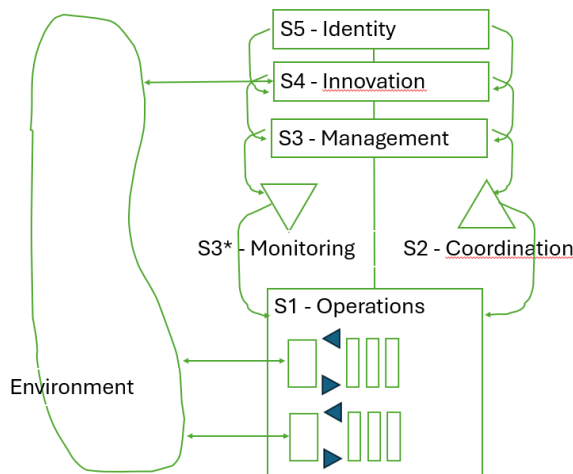
Interesting links related to system archetypes are:

- <https://thesystemsthinker.com/a-pocket-guide-to-using-the-archetypes/>
- <http://www.systems-thinking.org/arch/arch.htm>
- <http://blog.iseesystems.com/tag/archetypes/>

10.6.2 Viable System Model (VSM)

It is possible to evaluate the viability of an organisation or system, i.e. how well it performs as an open system seeking to survive in a turbulent environment, and whether it exhibits the necessary characteristics to thrive. The model can be used as an ideal type against which an actual organisation can be compared to identify weaknesses, mismatches or missing elements. It can also be used as a framework for organisation design.⁴²⁹

Figure 38 – Schematic representation of the Viable System Model



The VSM is presented in a graphical way (see Figure 38), as a picture of all the critical components that are connected in a particular way and are needed for an organisation to be viable. The VSM architecture consists of 5 subsystems and an environment:

- System 1 (S1): the primary operations or the set of activities performed by the system to create value.
- System 2 (S2): represents the conflict resolution function. It includes a set of activities that reconcile conflicting interests as parts of S1 interact.
- System 3 (S3): refers to the management function, which comprises activities that regulate S1, allocating resources to operations and ensuring that they achieve the desired level of performance.
- System 4 (S4): represents the innovation function, which encompasses activities that are oriented towards understanding the outside world in function of planning and change. Signals are picked up, interpreted and turned into meaningful information that may impact the activities of S1 and S3.
- System 5 (S5): represents the identity function, which are management activities that provide the ground rules and reflect on the system's governing purpose in the light of a changing environment and the different interests of the parties affected by the operation of the system.



The environment is everything that is outside the boundaries of the system. Some authors also refer to System 3* (S3*) as the audit function of the system, which includes activities that monitor the activities of S1 to help S3 perform its control function.⁴³⁰

The subsystems do not necessarily have different teams or departments. One person can perform several of the functions and can be active in several areas of the model. So the model is fundamentally different to a conventional organisation chart.⁴²⁹

Note that this graphical model is a simple, static systemic structure broken down by type of activity and connections between those activities, then how can it deal with complexity. VSM does this by being a recursive or fractal model. Within S1, there will be a set of sub-activities, each of which has the same system structure as S1 as a whole. So, we have viable systems made up of viable systems which are made of the same system architecture.⁴²⁹ Recursive means that the structure has the same pattern and properties at each level.⁴²⁹

Further reading:

Espejo and Harnden (eds.) (1989) *The Viable System Model: interpretations and applications of Stafford Beer's VSM*, Wiley.

Espinosa, A., Walker, J, Martinez-Lozada, A.C. (2023) *The Viable System Model: an introduction to theory and practice*, Journal of Systems Thinking 3 (1).

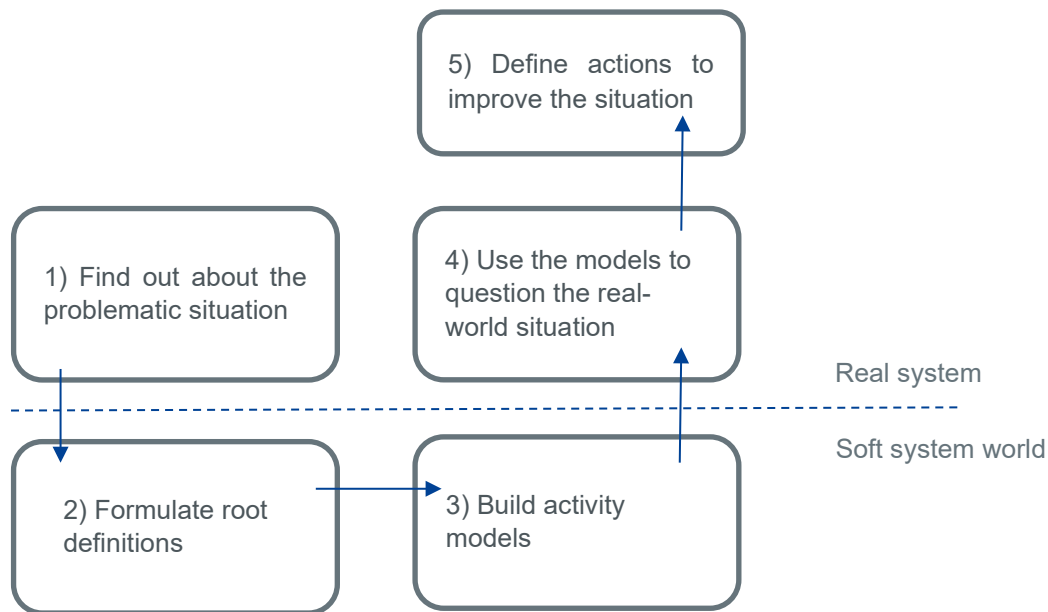
10.6.3 Soft systems methodology (SSM)

Soft systems methodology (SSM) is a subdiscipline, founded by Checkland and Poulter, that is most often referred to as representative of the second wave. Whereas quantitative modelling was most used in the first wave, in the second wave qualitative modelling is favoured, because intersubjective understanding is sought instead of objectivity. Soft systems approach explicitly integrate the assumption that an objective representation of reality does not exist: our perspective is always directed and filtered by our world view. It became important to consider individuals' perspectives on the goals to be achieved.⁴²⁴

Note that SSM uses a specific vocabulary. We will link it to concepts commonly used in evaluation (e.g. programme theory, intervention, context) where appropriate.

SSM is an approach which in a systematic way tries to establish and structure a debate concerning actions for improving the problem situation.⁴³¹ The SSM process takes the form of a cycle. It is a cycle of learning which goes from finding out about a problematical situation to defining/taking action to improve it. The steps in the learning cycle are described below (Figure 39).

Figure 39 – The SSM's learning cycle



FIND OUT ABOUT THE PROBLEMATIC SITUATION

The starting point is a problematical situation characterised by multiple interacting actors with their own world view and acting purposefully, with intention. This first step could be compared to describing the (complex) system in the system evaluation theory (SET) described earlier (see section 10.5.2.2).

In the language of SSM four ways of finding out about a problematical situation are described.

- Making rich pictures of interdependencies between actors

Rich pictures are created to show multiple interacting relationships. Knowledge about a situation can be assembled by means of interviews, reading documents, attending meetings etc. and be summarised afterwards in a rich picture. The pictures become richer as inquiry proceeds. In making a rich picture, the aim is to capture, informally, the main entities, structures and viewpoints in the situation, the processes going on, the issues currently recognised and any potential ones. Qualitative research techniques (such as observations, interviews, focus groups) are particularly suited to build rich pictures.

- Analysing the intervention

Identify who are in the roles of 'client' (i.e. the person or group of persons who caused the intervention to happen) and 'practitioner' (i.e. the person or group of persons who were conducting the investigation), and who are people who are concerned about or affected by the situation and the outcome of the effort to improve it.

- Analysing the social context

If we want to know whether a practical action could improve a situation, the changes involved in 'improvement' must be not only desirable but also culturally feasible. They need to be possible for particular people, with their particular history and their particular world views.

Every time we interact with the situation by talking to people, reading documents, sitting in a meeting, conducting an interview, we learn about the roles, norms and values characterising a particular group. Document them by writing down notes or memos.



- Analysing the political context

The political context is about the disposition of power in a situation and the processes for containing it. This is a powerful element in determining what is culturally feasible. Politics is also about accommodating different interests. In this analysis it is asked: 'how is power expressed in this situation?' What are the commodities (e.g. personal charisma, membership of various committees, reputation, access to information, etc.) which signal that power is possessed in this situation? What are the processes, by which these commodities are obtained, used, protected, defended, passed on, relinquished, etc.

FORMULATE ROOT DEFINITIONS

In order to construct an activity model, we need a statement describing the activity system to be modelled. This description is the root definition (RD), i.e. the description of what the system does, how and why. This is known as the PQR formula: do P (what), by Q (how), in order to help achieve R (why). The root definition is written out as a statement modelling a transformation process.

Although the PQR formula helps to define the root definition, which is the basis for the activity model, it can be further enriched by the use of the mnemonic CATWOE, which stands for Customer, Actor, Transformation, Worldview, Owner, and Environment. The idea is that purposeful activity, defined by a transformation process (Transformation) and a worldview (Worldview) will require people (Actors) to do the activities which make up T (Transformation). It will affect people (Customers) outside itself who are its beneficiaries or victims. It will take as given various constraints from the environment outside itself (Environment). It could be changed or stopped by persons (Owners) who are regarded as owning it.

In evaluation language a root definition could be described as the formulation of the intervention's theory of change in one sentence.

BUILD ACTIVITY MODELS

Building activity models means putting together the activities needed to describe the transforming process, in other words defining and linking the activities needed to achieve the transformation process. It is about the activities which do the transforming. Every activity in the model must be linkable to something in the root definition.

The activity model can be compared with the theory of change of an intervention, as an expression of how the intervention will lead to the desired results.

USING THE MODELS TO QUESTION THE REAL-WORLD SITUATION

As already explained, the activity models are the devices or tools which enable that discussion is a structured rather than a random one. The models are sources of 'good' questions to ask about the real situation, enabling it to be explored richly. For example: here is an activity in this model, does it exist in the real situation? Who does it? How? When? Who else could do it? The questions resulting from the comparison between the activity model(s) and the real world could be addressed in a focus group or even an individual face-to-face interview. An informal approach is to have a discussion about improving the situation in the presence of the models. We could ask whether we would like activity in the situation to be more, or less, like that in the model. Such questioning organises and structures a discussion/debate about the real-world situation. The purpose of the discussion is to surface different worldviews and to seek possible ways of changing the problematical situation for the better.

DEFINE/TAKE ACTION TO IMPROVE THE SITUATION, SEEK ACCOMMODATION

Identifying different world views and seeking ways for improvement, means finding an accommodation, this is '*a version of the situation which different people with different worldviews could nevertheless live with*' (p. 55).⁴⁰³

Questions which can inspire discussions leading to accommodation are for example, what combination of structural, process and attitudinal change is needed, how can it be achieved, what criteria will judge success/lack of success?

These questions represent things to think about when considering changes which are both desirable and feasible. The question about “enabling action” refers to which actions are needed to make a potential change accepted. This recognises the social context in which any change is embedded. Because of this context, introducing the change may require enabling action, which is not directly part of the change itself.

Further reading:

Checkland, Peter and Poultnier, John (2007) Learning for action. A short definitive account of soft systems methodology and its use for practitioners, teachers and students, Wiley.⁴⁰³

11 DATA REQUIREMENTS AND SOURCES

KEY POINTS

- The **identification of data requirements** is a core component of every evaluation study and should start at the earliest stages of the project. Data needs, data sources, data gaps and planned data collection activities should be explicitly described in the study protocol.
- Data requirements should be **driven by the research questions** and study design, not by the availability of existing datasets. Researchers should first determine what information is needed and then assess whether these data already exist or need to be collected specifically for the study.
- **Different types of data** may be required **for different components of an evaluation**, including problem description, target population identification, effect evaluation, implementation evaluation, contextual analysis, quality measurement, costing, budget impact assessment and economic evaluation. Information required for KCE evaluations commonly include sociodemographic and socioeconomic characteristics, clinical and epidemiological information, healthcare utilisation data, process-of-care information, quality indicators, and expenditure and cost data.
- **Administrative databases** constitute an important source of information for evaluation studies. This chapter provides a non-exhaustive overview of major Belgian administrative data sources collected routinely, such as reimbursement databases and the hospital discharge dataset. Administrative data offer broad population coverage, large sample sizes, relatively low acquisition costs, standardised coding and opportunities for longitudinal analyses, subgroup investigations and data linkage. They may nevertheless be limited by their original financing or regulatory purpose, incomplete or insufficiently detailed variables, coding practices that may affect validity, and delays in data availability.
- **Clinical registries** can provide more detailed disease-specific information than administrative databases. This chapter focuses on legally mandated registries with high population coverage, such as the Belgian Cancer Registry, as voluntary registries may be affected by selection bias.
- **Recurrent population surveys** complement administrative data by providing information that is typically unavailable in routine healthcare databases, such as self-reported health status, quality of life, health behaviours and non-reimbursed healthcare use. National surveys such as the Health Interview Survey are therefore important sources for evaluation studies.



- **Understanding each data source is essential** before using it in an evaluation. Researchers should familiarise themselves with the primary purpose of the data source, population coverage, variable definitions, coding systems, data collection procedures and known limitations.
- Close **collaboration** within the research team and with the KCE data manager, data analysts, etc. is crucial. Input from data holders, healthcare professionals and coding experts may be necessary to ensure appropriate data selection, interpretation and validation. In addition, study findings derived from routinely collected data should be critically assessed against evidence from the literature and, where appropriate, validated with stakeholders from the field. Unexpected findings may reflect data limitations, coding practices or contextual factors rather than true effects.
- Access to health data generally requires **formal approval procedures** and should be initiated as early as possible. Securing access often involves substantial preparatory work, legal and governance processes, and close interaction with the KCE data manager throughout the data cycle. The data process cycle should be considered from study conception to data destruction. Data identification, access, acquisition, linkage, preparation, analysis, reporting, governance and end-of-life procedures are interdependent steps that require early planning and continuous oversight.

11.1 Role of data in the evaluation programme

Data are a fundamental component of studies conducted within the evaluation programme and are essential for assessing whether an intervention achieves its intended objectives, for whom, under which conditions, and at what cost. The data requirements should therefore be considered from the outset of the study and should be directly linked to the evaluation questions, the chosen evaluation design and the outcomes of interest. It is also important to distinguish between data needed to plan the study and data collected or used to conduct the evaluation. Rather than starting from the data that happen to be available, the evaluation should first identify the information required to answer the relevant questions and subsequently determine whether these data are available, fit for purpose, or need to be collected specifically for the evaluation.

Data may be required at different stages of the evaluation. At the planning stage, data can be used to describe and characterise the target population, establish baseline levels of relevant outcomes and resource use, identify potential comparators, and assess the feasibility of the proposed evaluation design. During implementation, data may be needed to document intervention exposure, reach, fidelity and adaptations, as well as contextual factors that may influence implementation. During the outcome evaluation, data are required to assess effectiveness, safety, unintended consequences and, where relevant, differences in effects between population subgroups. Data on resource use, costs and outcomes are additionally required when an economic evaluation or broader assessment of economic impact is undertaken.

The study should make a clear distinction between routinely collected data and data collected specifically for evaluation purposes. Routine administrative data can provide important opportunities for efficient and large-scale evaluation, but their availability does not necessarily imply that they are suitable for every evaluation question. Variables may be incompletely recorded, definitions may differ between data sources or over time, and information on important outcomes, contextual factors or potential confounders may be absent. Conversely, primary data collection can provide information that is not available in routine sources but may increase the cost and burden of the evaluation. The choice between existing and newly collected data should therefore be based on their relevance, validity, completeness, timeliness, feasibility and associated costs.

Data quality, accessibility and governance should be considered from the outset of the study, alongside the evaluation design. The suitability of each data source has to be assessed in relation to its intended use, taking into account legal and ethical conditions and the rules and procedures

governing the use, protection and sharing of data. These considerations should inform the selection of data sources and, where relevant, the interpretation of study findings.

It is important that in the **data analysis part of the study protocol** these principles are translated into study-specific data requirements. For each evaluation question, it should specify the data required, how the relevant variables and outcomes will be defined and measured, the data sources to be used, the timing and frequency of data collection, and any requirements for data linkage. The protocol should also document the rationale for selecting the data sources, anticipated limitations in data availability or quality, and how these limitations will be addressed. Where data are collected specifically for the study, the protocol should describe the data collection procedures and responsibilities. Where routinely collected or data collected for a different purpose are used, it should describe the relevant extraction, linkage and validation procedures.

It is good practice to discuss a draft protocol with relevant KCE domain experts and, where appropriate, with other colleagues with expertise in specific data sources or data linkage. The draft should also be discussed with the KCE data manager (see also section 11.5).

11.2 Data sources

A wide range of data sources may be relevant, depending on the evaluation question. This section focuses primarily on **routinely collected administrative data**, as these provide an important source of information for many evaluations. However, registries, surveys, clinical records and data collected specifically for an evaluation may also be relevant, particularly when they provide information that is not available, or is insufficiently detailed, in administrative data. A brief overview of **surveys and registries** is therefore also included.

Administrative databases contain data collected primarily for administrative, regulatory or financial purposes, rather than specifically for research or evaluation. In healthcare, these include, for example, data generated through the reimbursement of healthcare services, hospital activity or health insurance eligibility. Although their primary purpose is administrative, such databases may also contain information relevant to clinical or other evaluation questions, such as diagnoses, procedures, treatments, healthcare utilisation and patient characteristics.

The distinction between administrative and other types of data is not always clear-cut. The same data may serve multiple purposes and may be combined with or linked to other data sources. For example, administrative healthcare data may be linked to clinical registries to obtain more detailed information on clinical characteristics and outcomes. Conversely, some databases that are primarily intended for clinical or monitoring purposes may also contain administrative information, such as the number of beds per service recorded in the Minimal Hospital Data (MZG – RHM). **The classification of a database should therefore not be considered a sufficient basis for assessing its relevance to an evaluation.** Data sources differ in their purpose, coverage, definitions, level of detail, quality and accessibility, and **no single source will generally provide all the information required for an evaluation.** Data sources should therefore be selected based on the evaluation question and the information needed to answer it. Depending on the evaluation, combining different sources may be necessary to obtain the required information and provide a sufficiently complete basis for the analysis.

The overview of routinely collected administrative data presented here is not exhaustive, and other data sources may therefore need to be considered for a specific evaluation. An overview of databases containing health- and healthcare-related data is available on the Health Data Agency's (HDA) platform (<https://catalog.hda.belgium.be>). For more detailed information on the administrative databases described below, we refer to the [HDA catalogue](#) for up-to-date information.



11.2.1 Administrative databases

STRENGTHS AND LIMITATIONS

Belgium has a wide range of administrative data that can be used for health and healthcare research. These data provide an important resource for evaluations, particularly when information on healthcare use, expenditure, population characteristics or care processes is required. However, administrative data are collected primarily for administrative, financial or policy purposes rather than research. Their usefulness therefore depends on the research question and on the content, quality and accessibility of the data. Administrative data should not be assumed to be suitable simply because they are available; their strengths and limitations should be assessed for each evaluation.

Administrative data have several important advantages. They are generally less costly to use than collecting new data and do not impose an additional burden on respondents. Because they are recorded as part of routine administrative processes, they may avoid some sources of measurement error associated with surveys, such as recall and social desirability bias. They reflect real-world practice and have the advantage that they include the heterogeneity of the covered population, including difficult to reach groups. They can provide detailed longitudinal information on healthcare use or expenditure over long periods, often for very large populations. Regular or continuous updates can also make them useful for monitoring changes over time. Systematic registration procedures and quality checks may further support data quality, while broad population coverage can allow analyses of population groups or geographical areas that would be difficult to study using surveys. They offer the potential to be linked with other data sources.⁴³²

Administrative data also have important limitations. They are rarely sufficient on their own to answer a research question, as variables relevant to the evaluation may not have been collected for administrative purposes. Researchers generally have no control over which variables are recorded, how they are defined or when definitions change. Information such as education, household composition, health behaviours, preferences, experiences or attitudes may therefore be absent. They also often lack clinical details (e.g. absence of lab tests).⁴³³ Even when a relevant variable is available, its definition, reference period, coverage or level of detail may not correspond to the research question. Variables that are not central to the administrative process may also be less complete or subject to changes in recording practices. Over- or under-registration may also occur when data collection is linked to financial incentives or reimbursement, as the way information is recorded may affect or be affected by the reimbursement process. Documentation may be limited, and changes in coding systems or definitions over time can affect comparability. The large samples entail the risk of giving an illusion of great power and precision. Finally, data protection and bureaucratic requirements may make access, linkage and publication more complex.⁴³²

Hereunder we describe the most common administrative databases that can be relevant to use in a KCE evaluation study without trying to be exhaustive in the data sources nor in the description of what they entail and how they are set up. It is the responsibility of each research team to further explore this more in detail when data sources are selected to be used in a KCE study.

IMA – AIM DATABASE

The InterMutualistic Agency (IMA – AIM) database is a partnership of the seven Belgian sickness funds that collects and combines detailed administrative health insurance data in a common format. It covers healthcare reimbursed by the compulsory health insurance and healthcare used by residents insured under this system. The database combines three main types of data: (1) population data, including sociodemographic and social security characteristics; (2) reimbursed inpatient and outpatient healthcare expenditure, including procedures, services, hospital admissions and pharmaceuticals supplied by hospital pharmacies; and (3) reimbursed pharmaceutical expenditure for medicines and other products supplied through public pharmacies (Pharmanet).

The database is available at individual patient level and allows longitudinal follow-up of healthcare use, including inpatient and outpatient care, before and after an intervention. It does not contain

patients' diagnoses or other clinical information as such. Proxy variables for selected health conditions can partly address this limitation. For example, IMA – AIM calculates pharmaceutical cost groups that indicate the potential presence of a chronic pathology based on reimbursed pharmaceutical use (see https://metadata.ima-aim.be/nl/app/vars/FKG_XXX_Pa (NL); https://metadata.ima-aim.be/fr/app/vars/FKG_XXX_Pa (FR)). It is important to carefully consider the use of such proxy variables as they can underestimate or overestimate the presence of a health condition depending on the specificity and sensitivity of their definitions.

The most common way to select data from the IMA – AIM database is through the RIZIV – INAMI nomenclature of healthcare (NL; FR) or by identifying pharmaceutical products of interest (NL; FR). For the latter, the [Anatomical Therapeutic Chemical \(ATC\) classification system](#) is of particular interest (see also IMA – AIM metadata NL; FR).

Healthcare not covered by compulsory health insurance is not captured in the IMA – AIM data. For example, certain pharmaceuticals like benzodiazepines are prescription-only but not reimbursed and are therefore not present in the data. These limitations should be considered when defining the study population and interpreting healthcare use and expenditure data.

The IMA – AIM data also give rise to several derived databases for specific purposes. Further information is available at <https://ima-aim.be/-Onze-databanken-?lang=nl> (NL) and <https://ima-aim.be/-Nos-banques-de-donnees-?lang=fr> (FR). Aggregated data can be consulted in the publicly accessible [IMA Atlas](#).

PERMANENT SAMPLE (EPS)

The InterMutualistic Agency also compiles and manages the Permanent Sample (EPS), a representative longitudinal sample of the IMA – AIM population containing healthcare consumption data from 2002 onwards. The sample consists of randomly selected members of the sickness funds, stratified by age and sex. Approximately 1 in 40 members under the age of 65 are included, compared with 1 in 20 among those aged 65 and older. The higher sampling rate among older people improves the precision of estimates for this group, which generally incurs higher healthcare costs and would otherwise be underrepresented.

The EPS can be useful for analyses of sufficiently large populations, particularly for common conditions. Extrapolation based on the stratified age and sex sampling design can allow for conclusions at the level of the Belgian compulsory insured population. However, additional checks using other information are necessary to verify the potential representativeness of the study population, e.g. by verifying the expected occurrence of selection criteria like RIZIV – INAMI nomenclature or pharmaceutical products codes. For rare conditions or small target populations, however, the sample may not provide a sufficient number of observations. Further information is available at <https://metadata.ima-aim.be/nl/app/bdds/Ps> (NL) and <https://metadata.ima-aim.be/fr/app/bdds/Ps> (FR).

RIZIV – INAMI DOC. N

The N documents (Doc. N) contain monthly aggregated data on reimbursed amounts, number of cases and, where applicable, days, by year of realisation (i.e. the date of the intervention), accounting year (i.e. the year in which the reimbursed amount was paid), and RIZIV – INAMI (pseudo-) nomenclature code.

Doc. N do not contain information on the prescriber or healthcare provider. Data are available by patient social category (for example, entitled to increased reimbursement or not, self-employed or salaried employee), district (*arrondissement*), age and sex, but not at individual patient level. As the data are aggregated and do not contain individual patient records, they cannot be linked to another database to follow individual patients over time, nor can they generally be used to estimate the number of unique patients. An exception is when the data concern a convention that is registered once per patient per year, in which case the number of records can be used to estimate the number of patients.



Two versions of Doc. N are available: at 18 months and at 24 months. The 24-month version is considered exhaustive, as healthcare attestations (documents certifying the provision of care) must legally be submitted to the sickness funds within two years of the date of care delivery to qualify for reimbursement. The 18-month version, which is available more rapidly, is already more than 95% complete. Every six months, the sickness funds complete the data by adding information on age, sex, social category and patient district. RIZIV – INAMI subsequently makes aggregated versions available to subscribers, such as KCE, on a quarterly basis.

The main advantage of Doc. N is their broad coverage of reimbursed healthcare activity, while their aggregated nature limits analyses to population-level volumes and expenditures. They are therefore not suitable for patient-level analyses, patient follow-up or analyses requiring linkage with other databases.

RIZIV – INAMI DOC. P

RIZIV – INAMI [Doc. P](#) consists of semi-annual statistical tables by (pseudo-) nomenclature code, healthcare provider, prescribing physician, place of intervention and patient residence. Codes that are not linked to a healthcare provider (e.g. a hospital) are not included.

Doc. P enables RIZIV – INAMI to monitor reimbursements received by the same patient, but only within the framework of its statutory missions. This level of detail is not available to third parties, such as KCE.

HOSPITAL BILLING DATA

The Hospital Billing Data comprise reimbursement and billing information for hospital stays covered by compulsory health insurance. Hospitals bill the patient's sickness fund for the services provided, based on the RIZIV – INAMI nomenclature. The data cover a broad range of hospital services, including fees charged by physicians and other healthcare providers, pharmaceuticals and implants or invasive medical devices, as well as lump-sum payments per admission and per diem. A subset of these transactions is transmitted annually by the sickness funds to RIZIV – INAMI in two datasets: Anonymous Hospital Stays (*Anonieme Ziekenhuisverblijven / Séjours hospitaliers anonymes*; AZV – SHA) for inpatient stays and Anonymous Day-care Stays (*Anonieme DagHospitalisaties / Hospitalisations de jour anonymes*; ADH – HJA) for day-care stays.

The AZV – SHA and ADH – HJA data are available at patient level and can be used to analyse hospital utilisation over time, including repeated admissions, and to study healthcare pathways when linked to other data sources. For registration year Y, the data cover stays from 1 July Y–1 to 30 June Y+1 and are transmitted by the sickness funds to RIZIV – INAMI in January Y+2. Since registration year 2017, patient supplements are included in the AZV – SHA and ADH – HJA data. Since registration year 2019, non-reimbursed implants and long-term invasive medical devices can be identified at patient level through their notification codes, together with the amounts paid by patients.³⁰²

MINIMAL HOSPITAL DATA

All general (including university) hospitals are required to submit twice a year a comprehensive set of data on inpatient and day-care hospital stays and emergency department contacts to the Federal Public Service Public Health (FOD – SPF). Day-care data include surgical and non-surgical day-care stays that meet the applicable registration criteria. In general, all stays for which the patient was discharged in the registration semester are in scope, with some additional stays based on long duration and VG-MZG – DI-RHM registration requirements (see below).

These Minimal Hospital Data (*Minimale Ziekenhuis Gegevens / Résumé Hospitalier Minimum*; MZG – RHM) contain demographic, clinical and hospitalisation-related information, including hospital and site identifiers, patient characteristics, admission and discharge information, diagnoses, procedures, nursing units, length of stay and discharge status. They also contain the information required to classify hospital stays using the 3M™ All Patient Refined Diagnosis Related Group (APR-

DRG) system, including the Severity of Illness (SOI) and Risk of Mortality (ROM) subclasses. APR-DRG classification is based on information such as the principal and secondary diagnoses, procedures, age and sex, with additional variables used for specific groups, such as birth weight for neonates.

The data are compulsory and pseudonymised and cover hospital episodes regardless of the patient's insurance status. The MZG – RHM can be used, among others, to analyse hospital activity, patient characteristics, case mix and care pathways. For example, the data can be used to identify admissions and readmissions within the same hospital, although they do not directly identify readmissions to another hospital. The TCT data, which link hospital and reimbursement data, can be used to identify readmissions across hospitals (see below).

The most common way to select data from the MZG – RHM is by diagnose or procedure as classified in the ICD-10-BE classification system ([NL](#); [FR](#)). Like with all code selection lists, it is important to have the code selection validated internally as well as externally, by subject-matter experts and data experts from the FOD – SPF, to assure accurate coverage.

The MZG – RHM also includes a structural data module describing the organisation and physical structure of hospitals. It provides information on hospitals and their sites, nursing units and bed indices, including the number of beds and associated cost centres. It also contains information on hospital associations. The research team should also determine the most appropriate unit of analysis for the study, i.e. the hospital as a whole or an individual hospital site, depending on the research question. The site level is important when specific services or activities are licensed at site level rather than at hospital level, for example when only one site of a hospital is licensed to operate a breast clinic.

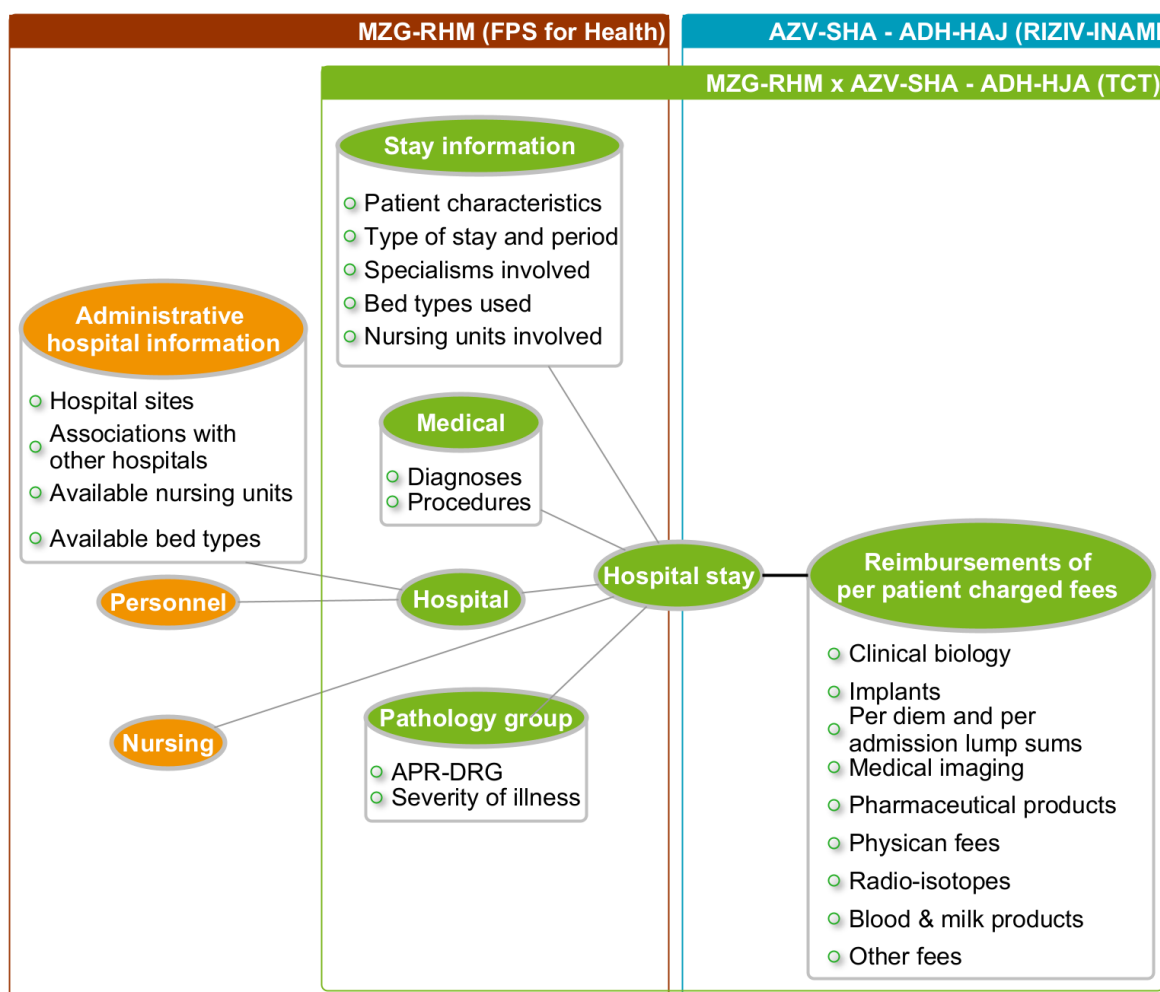
More detailed information on the data and variables is available on the [FPS Public Health](#) website and in the [HDA catalogue](#).

TECHNICAL CELL (TCT) DATA

The Technical Cell (*Technische Cel / Cellule Technique*; TCT) is a joint service of RIZIV – INAMI and the FOD – SPF. Its role is to collect, link, validate and anonymise hospital-related data and, where appropriate, to act as a trusted third party for linking data sources. Each year, the TCT links the Minimal Hospital Data (MZG – RHM) with hospital reimbursement data from the sickness funds (AZV – SHA and ADH – HJA) to support analyses of the relationship between healthcare expenditure and the conditions treated, as well as the development of hospital financing rules, accreditation standards and quality requirements. Figure 40 provides an overview of the structure of the information available in the TCT data.



Figure 40 – Structure of the information available to KCE in the TCT data



The resulting linked database combines information on hospital episodes with information on reimbursed hospital care. As the linkage relies on sickness fund reimbursement data, the TCT data cover patients insured under the compulsory health insurance scheme. Patients without such insurance coverage, as well as information on occupational diseases and occupational accidents, are therefore not included. Also, the linkage is quite comprehensive but not exhaustive as the linkage depends on matching stays and patients between the two sources.

Access to TCT data by KCE is provided for in the legislation establishing the TCT. The data available to KCE are subject to the scope of the data-delivery agreement and therefore do not include all variables present in the underlying datasets; in particular, the implant notification codes and patient supplements available in AZV – SHA / ADH – HJA are not included in the TCT data accessible to KCE and the reimbursements of clinical biology are grouped. More information on the MZG – RHM variables included in the linkage is available on the [FPS Public Health](https://www.fps.fgov.be/en/public-health) website.

NURSING MINIMAL DATA SET

The Nursing Minimal Data Set (*Verpleegkundige Gegevens van de Minimale Ziekenhuis Gegevens / Données Infirmières du Résumé Hospitalier Minimum*; VG-MZG – DI-RHM) provides standardised information on nursing activities in Belgian acute hospitals and is used, among other purposes, as a basis for hospital financing. Data collection is mandatory and takes place during four registration periods each year, covering the first 15 days of March, June, September and December. Registration covers nursing units in acute hospitals, excluding psychiatric nursing units, operating rooms and emergency departments.⁴³³

The VG-MZG / DI-RHM records nursing care at the level of a care episode, defined as a period beginning or ending with a transfer between nursing units or at midnight. Consequently, a patient may have several care episodes during a single day, for example when transferred between units or before and after surgery. The data provide information on the nursing activities actually performed, covering domains such as basic (e.g. hygiene care), and complex physiological care (e.g. administration IV medication), behaviour (e.g. cognitive therapy), safety (e.g. monitoring vital observations), family and relatives (e.g. post-partum care), and healthcare management (e.g. multidisciplinary concertation).

The data describe care actually delivered rather than care required. They can therefore be used to examine patterns and intensity of nursing activities, but do not by themselves allow assessment of whether the observed level of care was appropriate or sufficient. The VG-MZG / DI-RHM is a separate module within the MZG – RHM, implying that the nursing data can be complemented with patient-level and hospital-level information. For example, the MZG – RHM can provide information on age, comorbidities and hospital characteristics.

MINIMAL PSYCHIATRIC DATA

The Minimal Psychiatric Data (*Minimale Psychiatrische Gegevens / Résumé Psychiatrique Minimum*; MPG – RPM) are a mandatory administrative registration for psychiatric hospitals and psychiatric departments of acute hospitals in Belgium. The registration provides information on patients receiving psychiatric care and is used for healthcare policy, planning, financing and monitoring of psychiatric services.

The MPG – RPM comprises three types of data: structure data, continuous data and sample-based data. Structure data provide information on the organisation and characteristics of the psychiatric service, such as the type of service and its capacity. Continuous data are recorded for patients throughout their medical-psychiatric stay and include information on patient characteristics, admission and discharge, diagnoses, treatment and the end of the stay. Sample-based data are collected for a sample of patients during maximum two registration weeks per calendar year, with each registration week covering seven consecutive days. They provide more detailed information on patients' presence, functioning and care activities during these observation periods and therefore do not provide continuous information throughout the psychiatric stay.

The different types of MPG – RPM data can be used to analyse psychiatric care, including the characteristics of psychiatric services, patient profiles, diagnoses, admissions and discharges, length of stay and care activities. The scope and frequency of the information therefore differ between the three types of data. Individual-level MPG – RPM data can be requested from the FOD – SPF, subject to the applicable data-access and information-security procedures.

More detailed information on the data and variables is available on the [FPS Public Health](#) website and in the [HDA catalogue](#).

FINHOSTA

Belgian hospitals are required to submit financial, accounting, personnel and operational data to the FOD – SPF through [Finhosta](#), a mandatory data collection system.

The Finhosta data cover four main domains: budget, annual data, financial auditor and social agreements. The information includes hospital activities, accounting and financial results, personnel, investments and data required for the calculation of the hospital budget or Budget of Financial Means (BFM). Finhosta comprises multiple tables within these domains. The FOD – SPF provides detailed manuals specifying the variables, formats and reporting requirements, which are regularly updated and publicly available in Dutch and French.

Finhosta provides data at both hospital and hospital-service level, including information from general and analytical accounting and the allocation keys used in analytical accounting. It is therefore an important source for analysing hospital costs and resource use. However, caution is warranted when using these data for cost analyses. A KCE study based on 2019 Finhosta data found substantial



differences between hospitals in how costs were recorded, partly reflecting organisational differences and insufficiently detailed recording guidelines.³⁰⁶

More detailed information on the data and variables is available on the [FPS Public Health](#) website and in the [HDA catalogue](#).

11.2.2 Registries

Health registries collect systematically recorded information on individuals, health conditions, healthcare interventions or outcomes for monitoring, surveillance, quality improvement, research or other purposes. They vary considerably in their scope, coverage and purpose. Some registries have broad population coverage and can provide information relevant to population-level analyses and evaluations, while others focus on a specific disease, condition, treatment or patient group and may have more limited coverage or a highly specialised purpose. These categories are not mutually exclusive: for example, the Belgian Cancer Registry has national coverage but is disease-specific (see Chapter 9). This section therefore focuses on registries with broad population coverage or a wider relevance for health and healthcare evaluations. Highly specialised or disease-specific registries are not discussed in detail, as their relevance depends on the specific evaluation question and the condition concerned.

BELGIAN CANCER REGISTRY

The Belgian Cancer Registry (BCR) is a national, population-based registry that systematically records cancer diagnoses in Belgium.⁴³⁴ It was established in 2005 and has reliable cancer incidence data for the Flemish Region from 1999 onwards and for the whole of Belgium from 2004 onwards. To ensure that every newly diagnosed cancer is included in the registry, two legally obliged registrations exist:

- Hospitals with oncological care programmes and basic cancer care are required to register cancer diagnoses.
- Pathology, clinical biology and haematology laboratories are also legally required to register relevant test results.

The registry contains information on the patient and the tumour, including demographic characteristics, diagnosis and date of diagnosis, tumour characteristics and stage, pathology findings, treatment, follow-up and treatment outcomes. Information on survival and date of death is also available. The registry therefore provides a basis for studying cancer incidence, stage at diagnosis, treatment patterns, survival and other aspects of the cancer burden in Belgium. More information can be found at <https://kankerregister.org/nl> and in the [HDA catalogue](#).

The BCR data can be used for epidemiological research, monitoring and evaluation of cancer care and screening, and analyses of quality of care. For example, the registry supports analyses of differences in cancer incidence and survival between population groups and regions, monitoring of cancer trends, and evaluation of cancer screening programmes. The BCR also develops quality indicators for cancer care and conducts research using patient-reported outcomes and experiences.

An important strength of the BCR is the possibility of linking cancer registration data with other data sources. Subject to the applicable legal and governance requirements, BCR data are enriched with population and mortality data, health insurance reimbursement data and hospital data, such as admissions and length of stay. Such linkages can provide information that is not contained in the cancer registry itself. Two possibilities exist: 1) acquire access to the linked data that are available with the BCR (the data from the linked data sources such as MZG – RHM or IMA – AMI data are restricted in scope); 2) acquire permission to link the BCR data with the primary data sources such as the MZG – RHM and the IMA – AMI data. For both options, the data manager will have to be contacted to assess which route is the most adequate.

Publicly available data include aggregated figures by cancer type, sex, region and age in a user adaptable [dashboard](#), while more detailed anonymised or pseudonymised data can be requested from the BCR subject to its data-access procedures and applicable privacy and legal requirements.

Several KCE cancer reports have identified limitations, including incomplete or missing information on clinical stage, recurrence, cause of death and other clinical characteristics. The completeness and quality of the available information should therefore be assessed for the specific evaluation question and cancer type.

INTEGO

Intego is an electronic medical record-based primary care database coordinated by the Academic Centre for General Practice at KU Leuven.⁴³⁵ Established in 1994, it contains longitudinal information routinely recorded by participating general practices in Flanders, including patient characteristics, diagnoses, prescriptions, laboratory test results and other clinical information. Intego-II^{jj} comprises data from 180 contributing practices, including 135 active practices, and 973 444 unique individuals as of December 2024. Active patients represented 7.46% of the Flemish population. The database is organised into modules covering patient information, medical history and clinical encounters. Data are made available through a secure research environment, with quarterly releases since 2025.

Intego is particularly relevant for analyses of morbidity, primary care, treatment patterns and quality of care. Its data can also be linked to other sources, including IMA – AIM, to combine clinical information with data on healthcare utilisation. As the data are derived from routine clinical recording and participation is not universal, the completeness and representativeness of individual variables should be considered when interpreting results.

11.2.3 Surveys

Many other data sources may be relevant depending on the evaluation question. In addition to routinely collected administrative data and registries, this section provides a brief overview of surveys, which can complement these sources by capturing information that is generally not available in administrative databases or registries, such as patients' experiences, health status, quality of life, preferences and unmet needs. Surveys may also be conducted among healthcare professionals to collect information on their experiences, practices or perceptions.

The surveys presented here were selected based on their national relevance, established use and coverage of health or healthcare information that complements routinely collected administrative data. Other surveys are also conducted in Belgium for specific topics or population groups, but these are not discussed here because they have a more specific scope and are not intended as general sources of health or healthcare information.

HEALTH INTERVIEW SURVEY (HIS)

The Belgian Health Interview Survey (HIS), conducted by Sciensano, collects information on a broad range of health and healthcare domains, including health status and quality of life, lifestyle and health-related behaviours, preventive practices, mental health and well-being, healthcare use and experiences, and oral health. It also covers aspects of health and society, such as the impact of health conditions on daily and social life and collects a wide range of sociodemographic and socioeconomic characteristics. The survey therefore provides information that is generally not available in administrative data, including self-reported health, pain or discomfort, anxiety and depression, mobility, ability to perform usual activities, involvement in treatment decisions and selected measures of chronic disease prevalence and impact.

^{jj} The current database is referred to as Intego-II. Substantial improvements in data quality, database structure, operations and accessibility led to the transition from earlier versions to Intego-II.



The HIS is conducted among a representative sample of the Belgian population, including people living in nursing homes. Data are collected through face-to-face interviews and self-administered questionnaires. The standard sample comprises approximately 10 000 participants, with the possibility of oversampling specific population groups to obtain more precise estimates. Weighting procedures are used to produce estimates representative of the Belgian population. The survey has been conducted in 1997, 2001, 2004, 2008, 2013, 2018 and 2023. In 2018, a health survey was conducted among some of the HIS participants ([Health Examination Survey – HES](#)), while in 2023, an oral health survey ([Oral Health Data Registration and Evaluation System – OHDRES](#)) was conducted in selected HIS participants.

The HIS uses a repeated cross-sectional design, meaning that different samples of the population are surveyed in each wave. Therefore, longitudinal analyses following the same individuals over time are not possible. Since 2013, the Belgian HIS has also been part of the European Health Interview Survey (EHIS).

More information can be found at <https://www.sciensano.be/en/projects/health-interview-survey> and in the [HDA catalogue](#).

EUROPEAN UNION STATISTICS ON INCOME AND LIVING CONDITIONS (EU-SILC)

The EU-SILC is a harmonised European survey providing comparable information on income, living conditions, social exclusion and health. In Belgium, the survey is conducted by Statbel. The annual sample contains approximately 10 500 households, of which approximately 7 000 private households take part each year. All household members aged 16 years and over are individually interviewed. Collective households, such as nursing homes, monasteries and prisons, are not included in the sample.

The EU-SILC uses a panel design and therefore provides both cross-sectional and longitudinal data. Since the transition to a six-year panel was completed in 2021, households can participate for up to six consecutive years, with part of the sample being replaced each year. This makes it possible to analyse changes over time at the individual or household level, in addition to providing annual cross-sectional estimates.

The EU-SILC provides representative individual- and household-level information on sociodemographic and socioeconomic characteristics, including age, sex, education, employment status, income and sources of income, housing and living conditions, household composition and health status. In addition to the core questionnaire, ad hoc modules provide more detailed information on specific topics. A health-related module covering health, lifestyle and healthcare is included every three years, while a module on quality of life is included every six years.⁴³⁶ EU-SILC is therefore particularly useful for analysing socioeconomic inequalities in health and living conditions and for examining associations between socioeconomic characteristics and health. From 2019 onwards, EU-SILC underwent major content and methodological reforms, including greater use of administrative income data and revised weighting procedures, meaning that results from 2019 onwards are not directly comparable with those up to 2018.⁴³⁷

The EU-SILC can also be combined with or linked to other data sources to obtain information that is not collected in the survey, for example on healthcare utilisation, healthcare expenditure or healthcare supply. Its suitability for a specific evaluation should therefore be assessed in relation to the information required and the variables available in the relevant survey wave.

More information can be found at <https://statbel.fgov.be/en/themes/households/poverty-and-living-conditions/faq>.

HOUSEHOLD BUDGET SURVEY

The Belgian Household Budget Survey (HBS), conducted every two years by Statbel, collects detailed information on household consumption and expenditure. It combines interview data with household expenditure diaries and covers a representative sample of approximately 6 000 households. The

survey includes detailed expenditure data classified according to COICOP^{kk}, as well as sociodemographic and socioeconomic characteristics such as age, sex, household size, region, income and education. Healthcare expenditure is recorded in detail, including expenditure not covered by compulsory health insurance, and is valued at the amount paid before reimbursement where applicable. Some healthcare-related expenses, such as transport costs to healthcare providers, are not separately identified. Changes in survey methodology and consumption classification mean that data from before 2012 are not directly comparable with later waves.^{438, 439}

11.2.4 Other data sources

Other sources, including medical records, interviews, ad-hoc surveys, focus groups and direct observation of healthcare activities, may also be relevant for specific evaluations. For example, medical records can provide detailed clinical information but are very hard to acquire, while direct observation can be used to assess practices such as hand hygiene. Interviews, focus groups, direct observations and data collected specifically for an individual evaluation are not discussed further in this chapter.

11.3 Overview of relevant data

The data required for an evaluation depend on the evaluation questions and the type of analysis to be undertaken. Different types of data may be needed to characterise the target population, assess the implementation and delivery of an intervention, estimate resource use and costs, value healthcare services using tariffs or prices, assess quality of care, or estimate the budget impact of an intervention. These purposes are not mutually exclusive, and the same data may be relevant to several types of analysis. For example, demographic and clinical characteristics may be used both to characterise the target population and to stratify quality indicators; healthcare utilisation data may be used for costing, budget impact analysis and assessment of changes in care processes; and data on treatment uptake may be relevant to both process evaluation and budget impact analysis. The sections below therefore distinguish between data required to (1) characterise the target population, (2) estimate costs and tariffs, and (3) assess implementation, care processes and quality, while recognising that some data may be relevant to more than one of these purposes. The overview of relevant data is not exhaustive.

11.3.1 Data for characterising the target population

These data describe who is eligible for, affected by or potentially reached by the intervention.

11.3.1.1 Sociodemographic characteristics

Sociodemographic information describes the characteristics of the population relevant to the evaluation, such as age, sex, household composition, nationality or migration background, and place of residence. These data are primarily used during the planning and design of the evaluation, to characterise the target population, define and assess comparability between study groups, identify relevant subgroups, and examine differences in access, outcomes or effects. They may also be used in the analysis stage to describe the study population and, where appropriate, adjust for potential confounding or assess differences between population groups.

^{kk} The Classification of Individual Consumption According to Purpose (COICOP) is an international nomenclature for classifying household consumption expenditure into categories and subcategories. Healthcare expenditure is included under COICOP 06. Different versions are adapted to specific regions (e.g. Europe), countries or purposes, including household budget surveys.



- **Age, sex, community and the statistical sector^{ll}** of residence are available at individual level in the Population database of IMA – AIM. Statistical sector as such is usually not relevant but it allows to add demographic, socioeconomic and environmental statistics of the immediate neighbourhood where an individual resides.⁴⁴⁰ It can provide otherwise unavailable information, albeit at the cost of precision as the information is at the level of the statistical sector. The dataset with complete information for year Y is made available in November of year Y+2. Data are updated annually.
- **Household size, household members, the relationship between the entitled person and dependants, and whether the household is a single-parent family** are available in the Population database of IMA – AIM at household level.
- **Household composition, including family nucleus^{mm}, household type (LIPRO classification) and household size** are available in [Statistics Belgium \(Demobel\)](#) and the [Datawarehouse Labour Market and Social Protection of the Crossroads Bank for Social Security](#). Some variables are available at the level of a statistical sector. At individual or statistical-sector level, it also provides the **position within the household and civil status**. The latest available data are for 2024 and are updated annually.
- Presence of other household members is available in the Population database of IMA – AIM, which can be used as a proxy for the **availability of informal caregivers**.
- Population characteristics and area of residence can be obtained from [Statistics Belgium \(Statbel\)](#). Relevant information includes **population size, population density and indicators of urbanisation**, such as the Degree of Urbanisation (DEGURBA). These characteristics can be used to describe the geographical context of the population or to identify differences between areas.
- **Nationality and migration background** can be obtained from Statistics Belgium and the Datawarehouse Labour Market and Social Protection, at individual or statistical-sector level. Depending on the evaluation question, indicators may include current nationality, nationality at birth, parents' nationality or country of origin. At statistical-sector level, these data can, for example, be used to identify areas with a higher proportion of residents with a non-EU nationality or non-Belgian country of origin. Nationality and country of residence are also available in the MZG – RHM.
- **Distance-based accessibility:** the 'geographical breakdowns' published by Statistics Belgium provide information on the proximity of the population to hospitals. They identify the proportion of the population living within predefined distances of a hospital, currently 5 km, 10 km and 20 km, and provide results at different geographical levels, including the municipality. The latest data are based on population figures as of 1 January 2025 and hospital locations recorded in June 2025.⁴⁴¹

11.3.1.2 Socioeconomic characteristics

Socioeconomic characteristics can be relevant when differences in health status, healthcare use or access to care are expected across population groups. The available indicators differ in their definition, level of observation and detail; the choice of source should therefore reflect the socioeconomic characteristic required and the level at which it needs to be measured.

^{ll} Statistical sectors are small geographical units within municipalities, used by Statistics Belgium to collect and disseminate detailed demographic, socioeconomic and environmental statistics and to analyse local patterns and disparities. In 2025, Belgium comprised 20 781 statistical sectors.

^{mm} <https://statbel.fgov.be/en/themes/census/households-and-family-nuclei/family-nucleus>

- **Activity status:** the Population database of IMA – AIM and the Datawarehouse Labour Market and Social Protection provide individual-level information on employment, unemployment, incapacity for work, invalidity and other activity categories.
- **Household income:** the Population database of IMA – AIM provides proxies for household financial resources, including entitlement of increased reimbursement and receipt of social assistance. [Statistics Belgium](#) provides household income measures, including taxable income, disposable income and equivalised disposable income, at statistical-sector level.
- **Educational attainment:** the [Census 2021 of Statistics Belgium](#)ⁿⁿ provides information on the highest educational attainment and current participation in education, available at statistical-sector level. The Datawarehouse Labour Market and Social Protection provides individual-level information contributed by the regions.
- **Household work intensity:** the Datawarehouse Labour Market and Social Protection provides information on household work intensity at individual and statistical-sector level, including an indicator identifying households with low work intensity.
- **Work incapacity, invalidity and reintegration:** the Datawarehouse Labour Market and Social Protection provides information on periods of incapacity for work and invalidity, as well as on return-to-work and reintegration trajectories.

11.3.1.3 *Epidemiological data*

Epidemiological data describe the occurrence and distribution of diseases, health conditions and other health-related events in a population, including measures such as incidence, prevalence, mortality and disease progression. Examples of variables relevant to incidence, prevalence and mortality are discussed below. Other examples are not described here, as they have already been covered in the descriptions of the available databases in the previous section.

When **mortality** is an outcome of interest, the research team should determine at an early stage which data source is required and whether data linkage is necessary. In studies involving hospitalised patients, a distinction should be made between in-hospital and post-discharge mortality. In-hospital mortality can be identified from discharge status variables in the MZG – RHM and reflects deaths occurring during the hospital stay. Post-discharge mortality, such as 30-day, 90-day or longer-term mortality, cannot be measured using MZG – RHM alone and requires linkage with a source containing date-of-death information, such as [Statistics Belgium](#) mortality data or the IMA – AIM Population database.

When **incidence or prevalence** is an outcome of interest, particular attention should be paid to the validity of the case definition in relation to the characteristics and limitations of the underlying data. For example, when cases are identified based on reimbursed healthcare use, it should be considered that some individuals with the condition may not have used reimbursed healthcare and vice versa that some of the reimbursed services might include people without the target condition. When cases are defined using diagnostic or procedure classification systems such as ICD or International Classification of Primary Care (ICPC), researchers should verify the coding rules for diagnoses or procedures of the respective data source. For example, the MZG – RHM requires, for each specialty, the registration of one principal diagnosis and all other diagnoses that influence the care provided during the hospital stay, but excludes diagnoses considered irrelevant to the current stay. When pharmaceutical use is used to identify cases, researchers should assess whether the products are sufficiently specific to the condition of interest and are not commonly used for other conditions.

ⁿⁿ The previous census was conducted in 2011.



11.3.2 Data for costing and tariffs data

Cost data should be distinguished from prices or tariffs used to value healthcare resources. Costs reflect the resources required to provide a service, whereas prices or tariffs represent the amounts charged, reimbursed or otherwise paid for that service. A price or tariff may therefore be used as a proxy for the cost of a service when appropriate, but it does not necessarily reflect the resources actually used to provide it.

This distinction is particularly relevant when the amount received by a healthcare provider differs from the cost of providing the service. For example, a hospital may incur costs that are higher or lower than the amount it receives from public health insurance for a particular service. Conversely, an amount paid by a payer may include elements that do not correspond directly to the resources used to provide that service. The appropriate valuation therefore depends on the purpose and perspective of the analysis. Where the objective is to estimate the resources consumed, costs should be estimated as closely as possible from the resources required to provide the service. Where the objective is to estimate expenditure for a payer, the amounts actually paid by that payer may be more appropriate.

The **payer perspective and the provider perspective should therefore not be conflated**. The amount paid by public health insurance, for example, represents expenditure for the payer, but does not necessarily represent the cost incurred by the hospital or other provider. Similarly, a patient's payment represents expenditure from the patient's perspective but does not necessarily correspond to the full cost of providing the service. The valuation approach should consequently be determined by the perspective and purpose of the analysis, taking into account the availability and quality of the data.

11.3.2.1 Cost data

Finhosta is the only national database providing cost data for hospitals; no equivalent national cost database is available for other healthcare providers. For hospitals, the Belgian cost-accounting system records costs by cost centre and cost account. **Cost centres** represent hospital departments, services or activities to which costs are assigned, while **cost accounts** classify costs according to their nature, such as personnel, supplies, accommodation or depreciation. Cost centres are divided into temporary cost centres, where costs are initially recorded or accumulated before further allocation, and final cost centres, such as nursing departments, medical-technical services, consultations and pharmacy, and non-hospital activities. Costs that cannot initially be assigned to a final cost centre are recorded in accounts for costs awaiting allocation and are progressively allocated through a legally defined cascade method, using predefined **allocation keys** such as surface area, staffing or nursing days. This system is intended to improve comparability of costs between hospitals, although some allocation keys may not fully reflect the actual distribution of resources across activities.

11.3.2.2 Prices and reimbursements

We distinguish between healthcare costs covered by insurance and healthcare costs paid directly by patients, with further distinction according to the type of insurance coverage.

HEALTHCARE COSTS COVERED BY INSURANCE

- **Healthcare costs reimbursed by compulsory health insurance:** Information on reimbursed healthcare costs is available from the sickness funds through the IMA – AIM database, subject to the applicable access conditions. Derivatives of sickness-fund data are also available in several other Belgian data sources. These include RIZIV – INAMI datasets such as SHA – AZV / HJA – ADH and Doc. P (profiles of healthcare providers), as well as disease-specific or sample-based datasets such as the Belgian Cancer Registry and the EPS, respectively. The level of detail and the variables available differ between these sources.
- **Co-payments:** Information on the amounts paid by patients as co-payments is collected by the sickness funds and is available through the IMA – AIM database.

- **Supplements:** A distinction should be made between hospital and ambulatory settings. Data on supplements charged in the hospital setting are collected by the sickness funds and are available in the IMA – AIM database. In the ambulatory setting, only partial data are available in the IMA – AIM database. Since 2023, physicians billing electronically have been required to report ambulatory fee supplements to the sickness funds, with the transparency requirements further extended in 2024. Since September 2025, electronic billing in ambulatory care has been mandatory, subject to limited exceptions. These billing data are collected by the sickness funds and are available in the IMA – AIM databases. However, the completeness of the data should be assessed according to the type of supplement and the period considered.⁴⁴²
- Most **transport costs** incurred by patients to access healthcare are not covered by compulsory health insurance and therefore have to be borne by patients themselves. This includes, for example, ordinary travel to a GP, hospital or other healthcare provider. Exceptions apply to certain specific situations, where compulsory health insurance provides a contribution towards transport costs, such as travel for dialysis or cancer treatment and certain rehabilitation programmes, subject to specific eligibility conditions. Transport costs should therefore be considered from the appropriate perspective and, where relevant, included as patient costs.
- **Healthcare costs covered by private health insurance:** Information on healthcare costs reimbursed by private insurers exists at insurer level. However, these data are not routinely available for research through the Belgian health data infrastructure. Consequently, costs reimbursed by private health insurance may not be fully captured in analyses based on compulsory health insurance data.
- **Costs covered by specific solidarity schemes:** information on healthcare costs covered through specific public schemes may be available from the relevant administrative sources. For example, the Special Solidarity Fund provides financial support for certain healthcare expenses that are not reimbursed under the compulsory health insurance scheme.⁴⁴³

HEALTHCARE COSTS NOT COVERED BY INSURANCE

The Household Budget Survey (HBS) provides detailed information on healthcare expenditure, including costs not covered by compulsory health insurance and therefore paid by households.

HBS data are internationally used to calculate indicators of catastrophic and impoverishing out-of-pocket healthcare payments. A limitation for this purpose is the relatively short expenditure-registration period, which may not capture all healthcare expenditure incurred by households over a longer period.⁴³⁸

11.3.3 Administrative data and registries for process evaluation and quality measurement

Process evaluations consider three interrelated dimensions: implementation, which examines whether and how the intervention was delivered as intended; mechanisms of impact, which explore the processes through which the intervention is expected to produce its effects; and context, which considers the factors that may have influenced implementation and outcomes (see Chapter 5). Different types of data may be needed to assess these dimensions, and their suitability varies depending on the evaluation question.

Administrative data and registries are particularly useful for documenting aspects of **implementation** and describing relevant **contextual characteristics**. They can provide objective information on whether, when, where and for whom an intervention was delivered, and on variation in implementation across providers, institutions or population groups. They can also provide information on patient, provider, institutional or geographical characteristics that may help explain differences in implementation or outcomes. Data sources may include routine healthcare administrative data, hospital administrative data and disease-specific registries such as the Belgian Cancer Registry.



For example, in a process evaluation of the implementation of a new primary care reimbursement scheme, IMA – AIM data could be used to assess the uptake of the intervention, the number and characteristics of patients receiving the reimbursed service, the frequency of its use and differences in uptake between providers, regions or population groups. The data could also be used to examine changes in related healthcare utilisation before and after implementation. This is relatively straightforward if interventions are already reimbursed but can be more challenging if a new intervention is not yet reimbursed or only within e.g. a RIZIV – INAMI agreement specifically drafted to reimburse a new intervention including an evaluation of that intervention. It is of vital importance that in such cases extensive concertation between KCE and RIZIV – INAMI is possible during the drafting phase of a RIZIV – INAMI agreement. The objective is to obtain sufficiently detailed data registration, e.g. by using RIZIV – INAMI pseudo-nomenclature codes or other ad hoc data registration to be able to evaluate the actual use of the intervention and potential controls. This should include monitoring of data registration quality by e.g. RIZIV – INAMI or other partners involved in the agreement.

However, administrative data and registries are generally less suited to understanding **mechanisms of impact** within a process evaluation. They rarely capture why an intervention was implemented in a particular way, how it was experienced by patients or professionals, or how implementation processes unfolded. Qualitative methods such as interviews, focus groups and observations, as well as surveys and patient- or provider-reported data, are often needed to understand mechanisms of impact and the experiences underlying observed patterns, and can therefore provide complementary information on these aspects.

The same administrative data and registries can also be used to construct **quality indicators**, which provide standardised measures of whether specified aspects of care meet predefined standards or recommendations. They can be particularly useful for applying such indicators consistently across patients, providers or institutions and, where data are available longitudinally, for monitoring changes over time. For example, cancer registry and administrative data can be used to measure adherence to recommended diagnostic or treatment procedures, the timeliness of care, or compliance with a defined clinical pathway. When quality indicators are compared across providers, relevant case-mix variables should be taken into account, as differences in patient characteristics, such as disease severity, comorbidities or socioeconomic background, may influence the results. Not all relevant case-mix variables are routinely recorded or readily available in administrative databases, which should be considered when interpreting and reporting the indicators (see Chapter 9 for relevant data sources for the development of quality indicators).

11.4 Defining study variables from available data

In most research projects, constructs are defined that need to be measured and analysed to try to answer the research questions. For each of these constructs, an operational definition needs to be developed by selecting and transforming the raw data included in the data sources.⁴⁴⁴

11.4.1 Illustrative applications

Risk adjustment is commonly used when administrative data are used for comparative analyses (see also Chapter 3). Comorbidity is an example of a patient characteristic that may be included for this purpose. In hospital-based studies, several comorbidity algorithms are available, including the Elixhauser and Charlson algorithms, which are based on recorded diagnoses.⁴⁴⁵ The choice between such systems should take into account their respective strengths and limitations, based on the international literature on their validity and reliability. However, other considerations may also be relevant. These include whether an algorithm can be readily applied to the available Belgian data, for example whether an appropriate version of the Elixhauser algorithm is available for ICD-10-CM coding,⁴⁴⁶ and, where the literature does not demonstrate meaningful differences between approaches, consistency with previous KCE studies.

The characteristics of the available variables and coding practices in Belgian administrative data should also be considered. For example, Elixhauser comorbidities can be identified using all recorded secondary diagnoses or restricted to conditions that were already present at admission. The latter approach requires systematic availability of a present-on-admission indicator in addition to secondary diagnosis codes. As coding practices may vary between institutions, consultation with coding experts is recommended to assess the completeness and consistency of the available information and determine the most appropriate approach for the study.

Further methodological choices are required after selecting the comorbidity algorithm. The 38 Elixhauser comorbidity categories, for example, can be included separately in statistical models, preserving detailed information on individual conditions, or combined into a weighted summary index, as in several international studies. A weighted index allows individual comorbidities to contribute differently according to their relative association with the outcome of interest. Alternatively, the number of comorbidities can be counted and categorised into broader groups.

The latter approach was used in KCE Report 412, where the 38 Elixhauser comorbidities were combined into a categorical variable distinguishing patients with no comorbidity, one comorbidity, two comorbidities, and three or more comorbidities.³⁵⁴ This approach was chosen partly to reduce the influence of differences in coding intensity between hospitals. If some hospitals systematically record more secondary diagnoses per admission than others, this may result in artificial differences in measured comorbidity burden. Combining patients with three or more comorbidities into a single category can partly mitigate the influence of such differences while retaining sufficient discrimination for risk adjustment.

Another example is the **need to define the treatment schedule** for each patient in order to assess the typical use of an intervention and variation in its use between patients, prescribers or hospitals. The construct of interest could, for example, be the treatment schedule of bevacizumab among patients with ovarian cancer. The operational definition of this construct could then be specified as follows, assuming that patients with ovarian cancer can be identified using Belgian Cancer Registry data and that reimbursement data on bevacizumab are available:

- Identify per calendar day all reimbursements of bevacizumab defined as the occurrence of all package codes (CNK)) with a non-zero reimbursement that are associated with Anatomical Therapeutic Chemical (ATC) code L01XC07 for data before 1-1-2022 or ATC code L01FG01 for data after 1-1-2022.
- Calculate the dose administered on each day by using the number of daily defined doses per package.
- Calculate the delay in days between consecutive days of bevacizumab use.
- Determine a stopping rule for a single treatment schedule, e.g. one of:
 - no use of bevacizumab for six months;
 - decease of the patient;
 - patient no longer covered in the data (e.g. due to emigration);
 - end of the available data period.

When requesting data for such an analysis, it is important to balance the need for sufficiently detailed patient-level information with the need to minimise the amount and level of detail of the data requested, in accordance with data protection requirements. For example, rather than providing exact dates, the data may contain the number of days since a reference date that is not disclosed to the analyst. Similarly, variables such as RIZIV – INAMI nomenclature codes may be provided at a less detailed level or not provided at all.

Although the operational definition of the treatment schedule may appear specific, its interpretation is subject to limitations in the underlying data. A reimbursed package does not necessarily mean that



the product was administered to the patient, nor that the entire package was used. For example, the dose of bevacizumab depends on body weight, and a vial may therefore not be fully used, resulting in an overestimation of the administered dose when reimbursement data are used as a proxy. Changes in coding systems also need to be considered. The ATC code for bevacizumab officially changed in 2022, but this does not necessarily mean that the new code was adopted in Belgian data at the same time. The previous code may have continued to be used for some time, meaning that both codes may need to be included when identifying bevacizumab use over the relevant study period.

The definition of a stopping rule for an individual treatment schedule also illustrates the issue of right censoring that commonly arises in longitudinal administrative data. In addition, left censoring may occur when patients started treatment before the beginning of the available data period. These issues should be considered when determining the required observation period and, where necessary, addressed in the analysis.

Operational definitions of constructs that are central to the research questions should be validated both internally and externally, for example by subject-matter experts and coding or data experts.

11.4.2 Good-practice principles

Several principles should guide the selection and use of data sources.

- **The data requirements should be defined on the basis of the evaluation questions and the analytical approach.** Researchers should first determine which variables, populations, time periods and level of detail are needed to answer the evaluation questions and perform the planned analyses. Data should not be selected simply because they are readily available.
- **The origin, purpose, population covered, definitions, reference periods and methods of data collection should be assessed to determine whether a source is appropriate for the intended use.** Researchers should check why the data were originally collected, who is responsible for their collection, which population and period they cover, how variables are defined and how the information is recorded. It is, for instance, important to understand the classification and coding systems that are used. For example, when patient or service selection is based on nomenclature codes, searching [Nomenclsoft](#) by keywords alone may be insufficient, as this will only retrieve codes whose labels contain the search terms. The nomenclature is organised into chapters and articles, with codes grouped according to medical speciality. It may therefore be necessary to review all codes within a relevant article to ensure that the selection is complete. Another example is that of a reimbursement database that may provide reliable information on reimbursed healthcare use but may not capture care that is not reimbursed or information on the clinical reason for treatment. Similarly, a survey may provide information on health status or patient experiences that is not available in administrative data but may rely on self-report and recall.
- **Data quality should be evaluated, including completeness, consistency, accuracy and comparability over time.** Researchers should assess whether the required information is available for the relevant population and period, how much information is missing and whether values are plausible and internally consistent. They should also consider whether the recorded information accurately reflects the concept of interest. A healthcare service may be systematically recorded when it generates a reimbursement claim but be absent when it is not reimbursed. The extent and pattern of missing data should be examined, rather than assuming that missing observations are random. For example, cancer staging may be missing for a substantial proportion of cancer patients for certain hospitals or age groups. How such missing data are handled, and the potential impact of missing staging information on the validity of the study results, should therefore be considered.
- **Researchers should consider whether the population represented in the data corresponds to the target population of the evaluation and whether selection or coverage issues may affect the results.** They should check whether particular groups are excluded or under-

represented, for example because a database covers only people insured under a particular scheme, patients receiving a particular type of care, participating healthcare providers or people living in a particular region. They should also consider whether inclusion in the data depends on characteristics that are related to the evaluation, as this may result in selection bias. For example, participation in a survey may differ according to age, health status or socioeconomic characteristics, while healthcare providers participating in a registry may differ systematically from those that do not participate. Similarly, when complex pancreatic cancer surgery is identified using claims data, researchers should verify whether the reimbursed procedure used for case selection, such as a Whipple procedure, is performed exclusively for cancer patients or also for patients with other conditions. If the procedure is not specific to pancreatic cancer, linking claims data with the Belgian Cancer Registry may be necessary to identify the target population more accurately. Differences between the population represented in the data and the target population may limit the generalisability of the findings.

- **Any changes in definitions, coding, registration practices or data collection methods over time should be identified and taken into account.** Researchers should check whether variables have been redefined, coding systems have changed, new registration requirements have been introduced or the way information is recorded has changed. Apparent changes in healthcare use or outcomes may otherwise reflect changes in data collection rather than actual changes in practice. Where possible, researchers should use consistent definitions over time or document and account for breaks in comparability.
- **Data access and governance requirements should also be considered from the outset** (see section 11.5). Researchers should determine early whether the required data can legally and practically be accessed and whether their use requires specific approvals, agreements, authorisations or data-protection measures. The level of identifiability and the need to link data sources should be considered when planning the study. Data requirements should also be sufficiently well defined to allow the necessary access and governance procedures to be initiated in time.
- **Assumptions, limitations and important uncertainties related to the data should be documented and transparently reported.** Researchers should record important decisions about variable definitions, missing data, exclusions, data linkage and the interpretation of measures, as well as limitations that may affect the results. Where the data do not directly measure the concept of interest, the assumptions used to derive or approximate it should be made explicit. These issues should be considered when interpreting the findings and reported sufficiently clearly to allow readers to assess the robustness and applicability of the results.

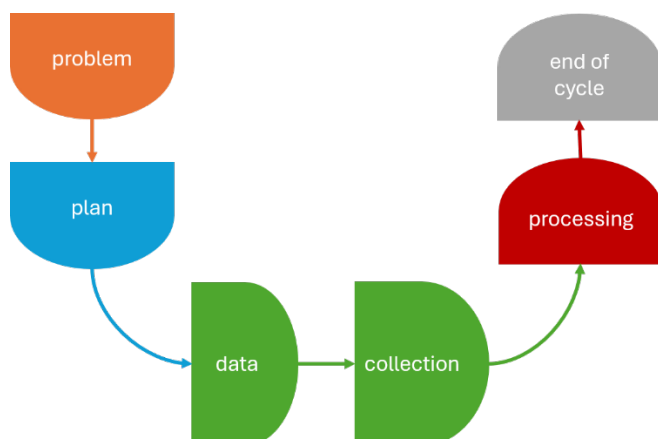
These principles help ensure that data are not only available, but also appropriate and sufficiently reliable for the intended evaluation.

11.5 Procedures for data access

KCE has established procedures covering the entire data life cycle (see Figure 41). This section only touches on aspects related to the acquisition of data relevant to the research questions agreed upon in the problem definition and scoping phase.



Figure 41 – Data processing life cycle



A **data analysis plan** as part of the study protocol should be developed before requesting the data, specifying the data sets and variables required, planned analyses, as well as the link to the research questions. The data analysis plan should be used to determine whether the requested data are proportionate to the study objectives.

When defining the required data, a distinction must be made between the need for anonymous data or personal data. Data requests follow different procedures depending on this distinction as well as on the source of the data, the purpose of the analysis and the proposed data processing.

For **anonymous data**, researchers can contact the KCE data manager for advice on the availability and characteristics of the required data, the appropriate data holder and access conditions. For some data sources, data holders can propose a ‘terms of use’ contract that stipulates the modalities of use including what parts of the data are to be considered confidential. KCE legal should also be involved when such a contract is required.

For **personal data**, researchers should always contact the KCE data manager at the start of the process, before approaching a data provider or submitting applications for approval. Here too, the data manager can advise on the availability and characteristics of the required data, the appropriate data holder, access conditions, and the procedures and approvals that apply to the specific request.

For personal data directly available at KCE, an a priori assessment of the requested data by the KCE supervising physician is mandatory. The KCE data manager will then collect the data and provide access in a dedicated analysis environment. After analysis and publishing of the research (end of life cycle), the data will be removed from the analysis environment, after consultation with the research team. Personal data acquired for a specific research project cannot be reused in a different research project.

For personal data that needs to be acquired from data holders or data subjects, additional steps are usually necessary. If new data are collected prospectively, either from data subjects or data holders, approval by an ethics committee is required for personal data related to health. New data in this context are data that have not yet been collected for another purpose. In some studies, the data collection or processing of personal data is subcontracted to a third party. As KCE remains the responsible controller, a data processing agreement (DPA) stipulating the type and means of processing and the respective responsibilities needs to be concluded. KCE legal should also be involved in the process of drafting and negotiating the DPA.

If the required data concern reuse of already collected personal data, an ‘a priori assessment’ by the Information Security Committee (*Informatieveiligheidscomité / Comité de sécurité de l’information*; IVC – CSI) will be necessary. A positive assessment by IVC – CSI results in a publicly available deliberation. In addition, most data holders require a data transfer agreement (DTA) that stipulates the conditions of use and transfer modalities among other things. KCE legal should also be involved in

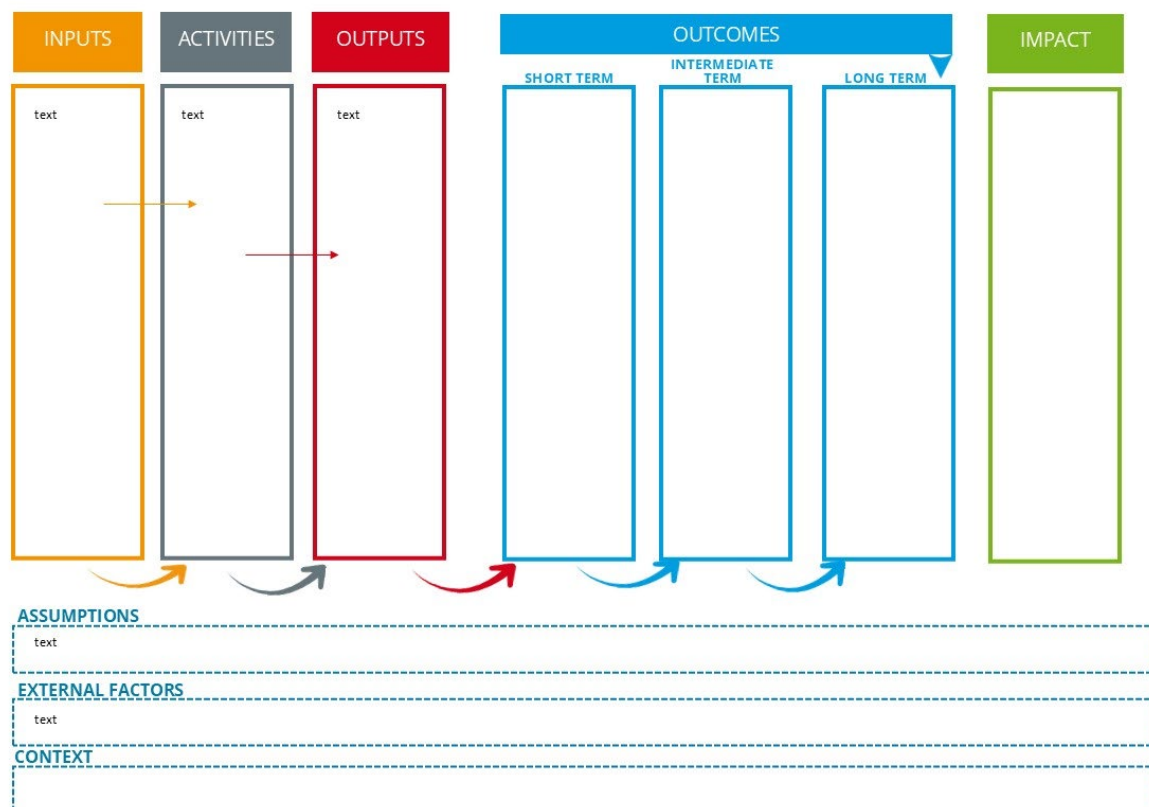


the process of concluding a DTA. The KCE data manager, in concertation with the research team and external parties, provides follow-up and performs the necessary steps during the execution of the dataflow to acquire the requested data. Once the data are acquired, the data follow a similar life cycle as described earlier.



■ APPENDICES

APPENDIX 1. TO CHAPTER 1



APPENDIX 2. TO CHAPTER 4

Appendix 2.1. Generalised pairwise comparisons

In some studies, **selecting a single primary endpoint can be difficult**. This may occur for several reasons. In certain areas, there is no widely accepted primary endpoint within the scientific community. In other cases, multiple outcomes may be considered equally important. One approach to address this issue is the **use of composite endpoints**, which combine several outcomes into a single binary measure.

A more recent approach is to account for the relative importance of different outcomes by introducing a predefined hierarchy among them. One of the methods based on this principle is Generalized Pairwise Comparisons (GPC), which can be viewed as an **extension of the classical non-parametric Mann-Whitney U test**. **The key idea of GPC is to prioritise outcomes.** For example, in a study evaluating a heart failure management program, the outcomes might be ordered as follows:

1. All-cause mortality at 12 months
2. Hospitalization for heart failure
3. Quality of life score.

This ordering must be pre-defined in the protocol. In the analysis, each participant in the experimental arm is compared with every participant in the control arm. If there are n_E participants in the experimental arm and n_C in the control arm, this results in $n_E \times n_C$ pairwise comparisons. For example, if both arms contain 10 participants, this yields $10 \times 10 = 100$ comparisons (also referred to as pairs).

For each pair:

1. The comparison starts with the highest-priority endpoint (e.g. mortality)
2. If one participant has a better outcome, this determines the result for that pair (“win” or “loss” for experimental vs control), and the comparison stops.
3. If the outcomes are equal, the comparison proceeds to the next endpoint in the hierarchy
4. This continues until a difference is observed or all endpoints have been considered.

Each pair can result in:

- A win for the experimental intervention
- A win for the control intervention (i.e. a loss for experimental), or
- A tie (no difference across all endpoints).

After evaluating all pairs, the overall result is summarised using the net benefit (NB):

$$NB = \frac{\#wins_E}{n_E \times n_C} - \frac{\#losses_E}{n_E \times n_C}$$

The null hypothesis is: $H_0: NB = 0$. Statistical testing is typically performed using nonparametric methods, such as randomisation tests. The net benefit can be interpreted as the difference between the probability that a randomly selected experimental arm participant has a better outcome than a randomly selected control arm participant, and the reverse probability.

Sample size calculation for GPC is complex and requires simulation, as it depends on the joint distribution of the prioritised outcomes. This means that evidence from prior studies or expert elicitation is needed, not only to specify the effects on each outcome, but also to characterize the correlations between them.



Appendix 2.2. Sample size calculation for cluster randomised trial without repeated measures

INTRA-CLASS CORRELATION

In cluster randomised trials (CRTs), entire groups (or clusters) of individual participants are randomised to interventions rather than individual participants. Examples of clusters include general practices, schools, hospitals, or families. Compared to individually randomised trials, **CRTs are generally less statistically efficient**. Note that we will only discuss the case of continuous outcomes in this process note.

The main reason of lower statistical efficiency is that **observations within the same cluster tend to be correlated**, with as a consequence that the **effective sample size is smaller than the total number of participants**. Such correlations may arise due to factors like shared healthcare providers (therapist effects), mutual influence between group members, or shared environmental or genetic characteristics. The degree of similarity of observations within a cluster is quantified by the **intraclass correlation coefficient (ICC), usually denoted by ρ** .

The ICC takes values between 0 and 1:

- $\rho=0$: observations within clusters are independent, and clustering has no impact on sample size.
- $\rho=1$: all observations within a cluster are identical, meaning that each cluster effectively provides only one piece of information.

The ICC tends to be inversely related to the cluster size. In larger clusters, individuals are typically more heterogeneous and mutual influences between members are relatively weaker. Conversely, smaller and more homogeneous clusters – such as families – tend to exhibit stronger within-cluster similarity. For example, general practices with 100s or 1000s of patients usually have very small ICCs, whereas family-based clusters may have higher ICCs, although these are still rarely close to 1.^{107, 108, 203}

IMPACT ON SAMPLE SIZE: DESIGN EFFECT

Because individual participants within clusters are correlated, the effective sample size is reduced compared to an individually randomised trial. To maintain the desired statistical power, the total sample size must therefore be increased, i.e. $N_{CRT} = N_{RCT} \times D$. **This inflation is captured by the design effect D (also called the variance inflation factor):**

$D = 1 + (n_{csize} - 1)\rho$ where n_{csize} is the cluster size (number of individuals per cluster) and ρ is the ICC.

This means that unless $\rho = 0$, the total CRT sample size will always be higher than the total RCT sample size, for the same design parameters.

The sample size formula for the difference of two normally distributed means, based on a RCT design without clustering, equals²⁰⁹

$$n_e = n_c = 2 \left[\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{\delta}{\sigma}} \right]^2, \text{ with}$$

- $n_e = n_c$ the sample size in the experimental and control arm, respectively
- $z_{1-\frac{\alpha}{2}}$ the z-score (quantile of standard normal distribution) corresponding with a $1 - \frac{\alpha}{2}$ probability
- δ the expected effect size under H_a
- σ the standard deviation in the population, assuming equal variance in the experimental and control arm of the study

Complete formulas for n_{nc} (the number of clusters) and n_{csize} (the cluster size) for the difference of two normal means are as follows:

$$n_{nc} = 4 \frac{1+(n_{csize}-1)\rho}{n_{csize}} \left[\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\delta} \right]^2$$

$$n_{csize} = \frac{4(1-\rho)}{\left[\frac{\delta}{z_{1-\frac{\alpha}{2}} + z_{1-\beta}} \right]^2 n_{nc} + 4\rho}$$

These formulas assume that the number of clusters is at least 32. Note further that a design with a fixed cluster size is more efficient. In case of varying cluster size, this is replaced with the average cluster size in the formulas.

Appendix 2.3. Sample size calculation for longitudinal CRTs, including stepped-wedge design

In Appendix 2.2, we described that sample size calculation for a CRT requires an estimate of the intra-cluster correlation coefficient (ICC). In longitudinal CRTs, where outcomes are measured repeatedly within clusters over time, additional parameters are required to describe the more complex correlation structure.

As mentioned before, three main types of longitudinal CRT designs can be distinguished. In a closed cohort design, the same individuals are followed and measured repeatedly over time, resulting in longitudinal (or repeated measures) data. In contrast, a repeated cross-sectional study assesses different individuals at each time point. Finally, in an open cohort design, individuals may enter and leave the study over time.

In longitudinal CRTs, outcomes are correlated at multiple levels: within clusters, across time within individuals (in closed cohort designs) and across time within clusters. To capture these dependencies, three types of correlations are commonly distinguished: the **intra-cluster correlation (ICC)** measures the similarity of individuals within the same cluster at a given time point. The **cluster autocorrelation (CAC)** reflects how similar cluster-level outcomes are across different timepoints. Finally, the **individual autocorrelation (IAC)** measures the similarity of repeated measurements within the same individual in closed cohort designs. IAC does not apply to cross-sectional longitudinal designs.

The above correlations can be expressed formally through a variance-covariance structure, which describes how outcomes are related, and which links to the appropriate statistical mixed-effects model. Three common correlation structures for longitudinal CRTs are exchangeable, nested (or block) exchangeable, and exponential decay.

Obtaining appropriate estimates of these three types of correlation is not straightforward. The optimal scenario is the analysis of historical data obtained in similar settings.

The simplest model, requiring the smallest number of parameters is the exchangeable correlation structure. However, in most cases it leads to an under-estimation of the sample size.

Sample size considerations become **even more complex in stepped-wedge designs**, where clusters switch from control to intervention over time. In these designs, the **intervention effect is partially confounded with time**. In addition, **assumptions about how the intervention effect evolves over time are required. The most common assumption is an immediate and sustained effect, but alternative patterns, such as delayed or gradually increasing effects, may also be considered.**^{108, 447, 448}

Several tools are available to support sample size calculation for longitudinal CRTs and stepped-wedge designs, including:

<https://clusterrcts.shinyapps.io/rshinyapp/>



<https://researchmethodsresources.nih.gov/swgrt-calculator?page=4>

<https://douyang.shinyapps.io/swcrtcalculator/>

Appendix 2.4. Drop-out and missing data in sample size calculation

In the planning of a study, it is important to account for participants who drop out or are otherwise unevaluable. This will also depend on the pre-specified analysis population for the primary analysis (see also section on estimands).^{203, 211}

Often, the primary analysis is an 'intention-to-treat' analysis. This means that all participants should be included in the primary statistical analysis, and that they should be analysed in the intervention arm, as dictated by the randomisation process. A common conservative assumption in this context is that there is no treatment difference among unevaluable participants of the two different arms. Under this assumption, their contribution to the overall treatment effect is zero. As a result, the observed treatment effect is diluted compared to the true effect among evaluable patients. Formally, if the true treatment effect is denoted by δ , the observed effect δ' becomes: $\delta' = (1 - x) \times \delta + x \times 0 = (1 - x) \times \delta$, with x the proportion of unevaluable patients (e.g. 0.1, 0.15 or even 0.5). This reflects the fact that only a proportion $(1 - x)100\%$ of participants contributes to the treatment difference. This reduction in effect size has direct implications for sample size. To maintain the same statistical power, the required sample size must increase.

The sample size formula for the difference of two normally distributed means is provided in Appendix 2.2.

The adjusted sample size, taking into account dropout, for the control arm becomes:

$$n'_c = 2 \left[\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{(1-x)\delta}{\sigma}} \right]^2 = 2 \left[\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{\delta}{\sigma}} \right]^2 \frac{1}{(1-x)^2} = n_c \frac{1}{(1-x)^2},$$
 with n_c the sample size for the control intervention without expected drop-out.

For small values of x , this corresponds approximately to an increase of about $2 \times x \times 100\%$ in sample size. For example, if 10% of patients are expected to drop out: $\frac{1}{(1-x)^2} = 1.23$, corresponding with an increase in the sample size of about 20%.

If, on the other hand, only evaluable participants are included in the final analysis, such as in a per protocol analysis, this means that a portion of the total sample size is not evaluable, namely xn' . So, in order to keep an evaluable sample size of n , the updated sample size n' needs to be $n' = \frac{n}{(1-x)}$ as the evaluable sample size will then be $n' - xn' = n'(1 - x) = n$.

Finally, it is important to distinguish between dropout (missing data) and intercurrent events, as emphasised in the estimand framework. These concepts have different implications for analysis and interpretation and should be carefully addressed during study design.

Appendix 2.5. Change from baseline and sample size calculation

In a randomised trial, baseline result values are expected to be approximately balanced between intervention arms. Consequently, in principle, interventions can be compared using only post-baseline measurements without introducing bias. In some settings, baseline values are not available, for example when comparing outcomes such as during surgery.

However, **when baseline measurements are available, it is generally preferable to incorporate them into the analysis. Doing so improves statistical efficiency by reducing variability and thereby increasing power.**

There are several approaches to incorporating baseline measurements in the statistical analysis. One option is to analyse change from baseline, defined as the difference between one single follow-up timepoint and the baseline measurement for each participant. This usually makes sense if there is only one follow-up measurement at the end of the study. Alternatively, baseline can be included as

part of a longitudinal analysis, treating it as an additional time point. A third approach is to include the baseline value as a covariate in a linear model, i.e. an ANCOVA analysis, which makes most sense in case there are repeated measures. These approaches are closely related and are, in general, equivalent.⁴⁴⁹ In practice, sample size calculations are often based on the change from baseline formulation because of its simplicity.

Assuming that baseline means are equal across intervention arms, the expected effect size is the same as when analysing post-baseline values alone. However, the relevant variance differs. For change from baseline, the sample size calculation is driven by the within-participant variability rather than the between-participant variability.

Let Y_F be the single follow-up timepoint measurement and Y_B the baseline measurement. The variance of the change from baseline is then given by:

$$\text{Var}(Y_F - Y_B) = 2\sigma_t^2(1 - \rho) = 2\sigma_w^2, \text{ where}$$

- σ_t^2 is the total variance of the measurement, which is simply the variance of the outcome, with usually equal variances assumed across intervention arms
- ρ the intra-class-correlation coefficient, representing the proportion of total variance attributable to between-participant differences
- σ_w^2 the within-participant variance

If historical data with repeated measurements are available, the within-participant variance can be estimated directly. Otherwise, the variance of the outcome as such may be obtained from prior studies, and an additional assumption is required for the intra-class correlation.

Appendix 2.6. Adaptive designs

The sample size calculations we have considered are based on a *fixed sample design*, in which no substantive changes are planned during the conduct of the study. This approach does not take advantage of data that accumulate while the trial is ongoing. Unless an IDMC makes unplanned changes – for example stopping the study due to unacceptable toxicity – the study proceeds to completion without modifications.

In contrast, *adaptive designs* are planned from the outset to allow **pre-specified modifications based on interim data**, with both the types of adaptations and the rules governing them clearly defined in the study protocol. A wide range of study design aspects may be adapted, but these can be broadly divided into two categories:²¹⁶

1. Early stopping of the study for success (**superior efficacy** of the experimental intervention) or failure (**futility**)
2. More **structural design modifications**, such as
 - a. Changing the sample size
 - b. Modifying interventions, such as dropping an unsuccessful arm
 - c. Changing eligibility criteria of participants

Note that this type of interim analysis is not usually done in studies with complex interventions.⁷⁰

GROUP SEQUENTIAL DESIGNS⁴⁵⁰

In clinical trials with a fixed sample design, the total sample size is determined in advance and a single final analysis is conducted once all data have been collected. A more flexible approach involves conducting interim analysis during the study, **allowing the study to stop early for efficacy or futility**, based on the final results of a part of the study population. Note that this is only applicable if enrolment happens in a staggered way. If all participants would be randomised at once, then this type of design would not be applicable.



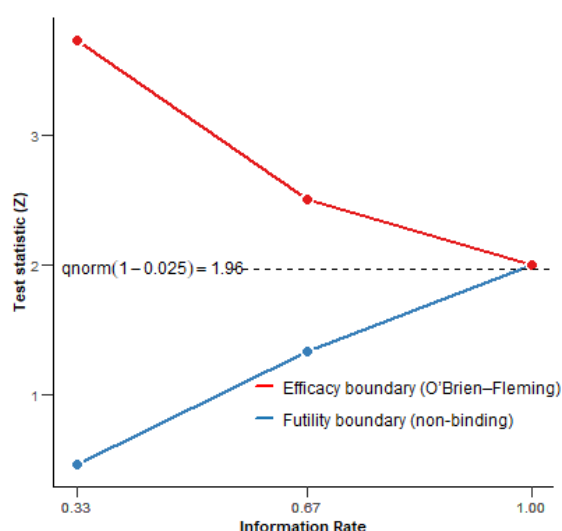
In theory, one could perform a statistical analysis after each new observation, i.e. the consecutive incoming final results of each participant. However, this would require continuous monitoring, which is rarely feasible.⁴⁵¹ In practice, interim analyses are **usually conducted at a small number of pre-planned occasions**. These are usually scheduled by the **information fraction, not the calendar time**, e.g. 1 interim when $\frac{1}{2}$ of participants have reached the final endpoint, or 2 interims when $\frac{1}{3}$ and $\frac{2}{3}$ of the participants have reached it. This is the defining feature of group sequential designs, in which accumulating data are analysed at specific intervals, **based on a proportion of the participants who have reached the final study endpoint**.

As explained before, performing multiple analyses at the conventional significance level α inflates the type I error rate. Conversely, stopping early for futility reduces study power. The first group sequential designs were proposed by **Haybittle and Peto, Pocock and by O'Brien and Fleming**, and they required interim analyses at fixed time points defined by the proportion of participants observed. Later developments introduced **alpha-spending (for efficacy) and beta-spending (for futility) functions**, which allow more flexibility in the timing of interim analyses. Commonly used approaches include the Lan-DeMets alpha spending methods. Despite this flexibility, it is still strongly recommended to pre-specify the timing of interim analyses.

Group sequential methods usually use more conservative thresholds than would be the case for one final analysis. This means that, in general, the earlier the interim analysis, the better the observed interim results need to be to allow stopping for efficacy and the worse to allow stopping for futility.

Figure A. 1 gives an illustration of efficacy (in red) and futility (in blue) stopping rules at $\frac{1}{3}$ and $\frac{2}{3}$ information time. While a conventional significance level $\alpha = 0.025$ corresponds with a z-score of 1.96, it can be seen in the figure that at the final analysis, this value is a bit higher, corresponding with a slightly lower α at the last analysis. This is an example of rather conservative stopping rules, where the threshold needs to be rather extreme in order to stop the study early for efficacy or futility. More aggressive stopping rules, such as Pocock's allow to stop the trial in case of less extreme measures. But these require a higher sample size, in order to preserve pre-specified type I and type II errors.

Figure A. 1 – Group-sequential design with Lan-DeMets O'Brien-Fleming type alpha and beta spending functions



ADAPTIVE TRIAL DESIGN

Group sequential designs are adaptive only with respect to the decision to stop the study early, while all other design features remain unchanged. In contrast, general adaptive study designs allow for



prospectively specified modifications to one or more structural elements of the study design, based on interim data. Such adaptations result in (usually only two) distinct stages: participants whose final outcomes are available before the design change, and participants whose outcomes are observed after the change. Separate statistics are calculated for each stage and then combined in a pre-defined manner to preserve the overall type I error rate. Implementing this approach requires specialized statistical methods. Two important approaches are the ***combination-of-tests method*** and the ***conditional Type I error approach***.

Adaptive trials may raise concerns mainly related to credibility. As a result, their planning requires **extensive simulation studies to evaluate operating characteristics, including power and type I error control**.⁴⁵²



APPENDIX 3. TO CHAPTER 5: EXAMPLES OF TMFs USED IN PROCESS EVALUATIONS

Appendix 3.1. Focus on the mechanisms of impact and outcome: the 'outcome and system-oriented' TMFs

Table A. 1 – ENCOMPASS evaluation framework

ENCOMPASS	
Origin of the TMF	Developed by A. Luna Pinzon and colleagues (2022)
Objectives	▪ Guide evaluation of complex public health programmes, operating within complex adaptive systems ▪ Support the design of systems evaluations ▪ Capture interactions between programme activities, system actors and context ▪ Support adaptive, theory-informed and systems-oriented evaluation approaches
Main components	Five iterative stages for conducting a systems evaluation: - Stage 1: Adopt a system dynamics perspective on the overall evaluation design: 1a. Specify the logic model 1b. Define the aims and scope of the evaluation 1c. Formulate evaluation questions 1d. Form a multidisciplinary evaluation team - Stage 2: Define the system boundaries: what is inside and outside of the system under study - Stage 3: Develop an understanding of the pre-existing system to inform system change: 3a. Map the pre-existing system 3b. Identify leverage points - Stage 4: Monitor dynamic programme output at different system levels and establish real-time monitoring mechanisms, track programme outputs and system responses, and refine the theory of change through the identification of action functions and feedback processes - Stage 5a: measure programme outcome, include intended and unintended outcomes - Stage 5b: measure programme impact in terms of system changes, examine the extent to which changes in the system resulted from the programme implemented → Steps are interconnected and revisited iteratively rather than followed linearly
When to use?	It can be applied to the evaluation of public health programmes in complex adaptive systems, from understanding the system, to developing actions to change the system, to measuring system changes
Comments	Explicitly designed as an iterative and adaptive evaluation process, rather than a linear sequence of steps Strong emphasis on complexity, systems thinking, and co-evolution of intervention and context Encourages ongoing refinement of programme theory and methods during the evaluation. Value of ENCOMPASS lies in integration of key characteristics from existing systems evaluation studies and in its practical, applied focus

Source: Luna Pinzon et al.⁵⁵

Appendix 3.2. Focus on the context: 'context and system complexity' TMFs

Table A. 2 – NASSS: Non-adoption, Abandonment, Scale-up, Spread and Sustainability framework

NASSS	
Origin of the TMF	Developed by Trisha Greenhalgh and colleagues (2017)
Objectives	▪ Explain why complex health and care technologies succeed or fail ▪ Understand adoption, non-adoption, abandonment, and sustainability ▪ Identify factors influencing scale-up and spread ▪ Support planning for long-term implementation success
Main components	Six core domains plus a seventh domain addressing adaptation and sustainability over time: 1. Condition (nature of the health condition) 1A. What is the nature of the condition or illness? 1B. What are the relevant sociocultural factors and comorbidities? 2. Technology (characteristics and usability) 2A. What are the key features of the technology? 2B. What kind of knowledge does the technology bring into play? 2C. What knowledge and/or support is required to use the technology? 2D. What is the technology supply model? 3. Value proposition (benefits and costs) 3A. What is the developer's business case for the technology (supply-side value)? 3B. What is its desirability, efficacy, safety, and cost effectiveness (demand-side value)? 4. Adopter system (staff, patient, and lay caregiver[s]), the health or care organisation(s) (including attention to the work of implementation and adaptation) 4A. What changes in staff roles, practices, and identities are implied? 4B. What is expected of the patient (and/or immediate caregiver) and is this achievable by, and acceptable to, them? 4C. What is assumed about the extended network of lay caregivers? 5. Organisation(s) and social context 5A. What is the organisation's capacity to innovate? 5B. How ready is the organisation for this technology-supported change? 5C. How easy will the adoption and funding decision be? 5D. What changes will be needed in team interactions and routines? 5E. What work is involved in implementation and who will do it? 6. Wider system (policy, regulation, infrastructure) 6A. What is the political, economic, regulatory, professional (e.g., medicolegal), and sociocultural context for programme rollout? A seventh domain considers embedding and adaptations <u>over time</u> (complexity dynamics) 7A. How much scope is there for adapting and coevolving the technology and the service over time? 7B. How resilient is the organisation to handling critical events and adapting to unforeseen eventualities? → Domains are interdependent and dynamic rather than linear
When to use?	To assess and anticipate challenges in scaling, sustaining, or discontinuing complex interventions or technologies Could be used: 1. at an early stage in a technology development to inform technology and service design, 2. in strategic planning to identify technological innovations that (perhaps despite policy enthusiasm) have limited chance of achieving large-scale, sustained adoption 3. in technology implementation projects to address the micro-level challenges of individual adoption, the meso-level challenges of organisational assimilation, and the macro-level challenges of the policy and regulatory environment 4. to inform and support scale-up and rollout of technology programs; and 5. retrospectively to explain programme failures



Comments	The framework is intended to be used reflexively to guide conversations and help generate ideas, not as a checklist Useful for digital health and service innovations
----------	---

Source: Greenhalgh et al.²⁴⁹

Table A. 3 – BANANA: Basel Approach for coNtextual ANalysis

BANANA	
Origin of the TMF	Developed by a group of researchers at the university of Basel (2022) (based on CICI framework)
Objectives	▪ Provide methodological guidance for planning, conducting and reporting contextual analyses in implementation science ▪ Support the identification of contextual barriers and facilitators ▪ Inform intervention adaptation and implementation strategy selection ▪ Improve consideration of context throughout implementation project
Main components	BANANA comprises six interacting components that guide contextual analysis from framework selection to reporting and application of findings: Component 1: Choose a TMF to guide contextual analysis → Considerations when selecting a TMF for contextual analysis - TMF acknowledges the multidimensional, multilevel, and dynamic nature of context - TMF fits the intervention and/or setting in which the intervention will be implemented <i>Consider combination of a context and setting-specific TMF</i> Component 2: Use empirical evidence to identify relevant contextual and setting factors → Identification of empirical evidence on relevant contextual and setting factors for implementation using various sources of evidence - Local data and information - Professional knowledge/clinical experience - Patient experiences and preferences - Research Component 3: Involve stakeholder - Identification and listing of relevant stakeholders for contextual analysis (target group, implementers, decision-makers, others) from different levels (micro, meso, macro) - Mapping of stakeholders in a stakeholder matrix specifying their characteristics (e.g., influence, role, activity, product) - Visualising stakeholder characteristics in an influence-interest-capacity matrix - Verifying stakeholder availability and commitment - Developing a stakeholder strategy specifying stakeholder tasks, timepoints, and methods for involvement - Evaluation of stakeholder involvement and adaption if needed Component 4: Develop a study design for contextual analysis → Data collection is guided by theory, empirical evidence, and stakeholder input. Choice of appropriate methods to answer the research questions such as - Quantitative methods (e.g., survey, routine data) - Qualitative methods (e.g., interview, focus group, observation) - (Rapid) ethnography <i>Consider changes of context over time</i> <i>Plan (if possible) several timepoints for data collection (e.g., prior, during, and at the end of the project)</i> Component 5: Determine the relevance of context for intervention co-design, choice of implementation strategies, and interpretation of outcomes → Findings from the contextual analysis can be used for:

	- Development/adaption of the intervention - Choice/adaption of implementation strategies - Interpretation of implementation and effectiveness outcomes - Choice of sustainability strategies <i>Consider the development of a programme theory to describe/visualise causal pathways between intervention components, implementation strategies, and contextual factors</i> Component 6: Report on contextual analysis → Reporting contextual analysis as part of the implementation intervention study (detailed findings can be reported in a separate paper) Suggestions for reporting based on BANANA: - Definition of context and operationalisations of contextual and setting factors studied - TMF applied for contextual and setting analysis and description of how it was used - Overview of empirical evidence identified and used - Stakeholder involvement (i.e., stakeholder strategy) - Reporting methods applied for data collection and analysis (e.g., study design, measures used, contextual and setting factors assessed) - Use of findings from the contextual analysis for subsequent project phases (cf. component 5)
When to use?	BANANA is useful for earlier stage work (e.g., preparatory work) as well as implementation studies that include the development/adaptation, implementation, and evaluation of an intervention and its implementation strategies
Comments	BANANA provides methodological guidance on planning and conducting contextual analyses BANANA describes how contextual information can be reported and used to inform further project phases (e.g., intervention development) The first three components (TMF selection, evidence gathering and stakeholder involvement) are often conducted iteratively and partly in parallel Deliberately not a “checklist” approach

Source: Mielke et al.²⁵⁰

Appendix 3.3. Focus on implementation and context: ‘determinant’ TMFs

Table A. 4 – TDF: Theoretical Domains Framework

TDF	
Origin of the TMF	Developed by a multidisciplinary group of behaviour change and implementation researchers (2005); refined and validated by James Cane, Denise O'Connor and Susan Michie (2012)
Objectives	▪ Understand determinants of behaviour change among stakeholders ▪ Identify barriers and facilitators to implementation at the individual level ▪ Inform the design of behaviour change interventions and implementation strategies ▪ Link theoretical constructs to practical implementation strategies
Main components	14 domains, 84 constructs (number in brackets) 1. Knowledge (awareness, procedural knowledge) (3) 2. Skills (competencies, abilities) (7) 3. Social/professional role and identity (9) 4. Beliefs about capabilities (self-efficacy) (8) 5. Optimism (4) 6. Beliefs about consequences (outcome expectations) (5) 7. Reinforcement (7) 8. Intentions (3) 9. Goals (goal setting, priorities, goal conflict) (6)



	10. Memory, attention and decision processes (5) 11. Environmental context and resources (6) 12. Social influences (norms, support, networks) (11) 13. Emotions (affect, stress) (7) 14. Behavioural regulation (monitoring, action planning) (3) → Not all are required in a single study
When to use?	To explore behavioural determinants, particularly through qualitative methods (e.g. interviews, focus groups), and to guide intervention design and evaluation
Comments	Primarily focused on individual-level mechanisms but can incorporate social and environmental influences

Source: Cane et al.²⁵²

Table A. 5 – i-PARIHS: Integrated-Promoting Action on Research Implementation in Health Services

i-PARIHS	
Origin of the TMF	Developed by Gill Harvey and colleagues (1998, updated in 2016), building on the PARIHS framework
Objectives	▪ To help both explain and predict why the implementation of evidence into practice is or is not successful ▪ Emphasise the role of facilitation in implementation processes ▪ Account for interactions between innovation, recipients, and context
Main components	Core constructs organised around four interacting elements: - Innovation (characteristics of the intervention): ○ Clarity ○ Evidence base ○ Usability ○ Fit with practice - Recipients (individuals and teams involved): ○ Motivation ○ Values ○ Skills ○ Knowledge ○ Time and resources - Context: ○ <u>Inner context</u> (local setting): ▪ Culture ▪ Leadership ▪ Evaluation and feedback ○ <u>Outer context</u> (wider system): ▪ Policy ▪ Regulatory environment ▪ Socio-political factors - Facilitation: ○ Roles ○ Skills ○ Strategies ○ Processes supporting change
When to use?	To guide implementation planning and evaluation, particularly when facilitation is a central component, and to understand dynamic interactions between key elements
Comments	i-PARIHS makes explicit the multi-dimensional and complex nature of implementation as well as highlighting the central importance of context Strong emphasis on facilitation as a mechanism for change

Source: Harvey et al.²⁵³

Appendix 3.4. Focus on implementation: ‘process’ TMFs

Table A. 6 – EPIS: Exploration, Preparation, Implementation, Sustainment

EPIS	
Origin of the TMF	Developed by Gregory Aarons and colleagues (2011)
Objectives	- Understand and guide implementation across its different phases (and consider challenges and opportunities in implementation) - Identify multilevel factors influencing implementation over time - Support the adoption, scaling, and sustainment of interventions - Analyse interactions between context and implementation processes
Main components	Four phases with multilevel, interacting determinants Exploration (identification of needs, fit, and potential interventions) Preparation (planning, adaptation, capacity building) Implementation (active delivery, monitoring, problem-solving) Sustainment (integration into routine practice, long-term maintenance) Cross-cutting factors: - Outer context (policy, funding, system-level influences) - Inner context (organisational characteristics, leadership, culture) - Bridging factors (linkages between outer and inner contexts) - Innovation factors (fit and characteristics of the intervention)
When to use?	To understand how an “intervention” is adopted, implemented and sustained → It analyses implementation as a dynamic, phased process, particularly in longitudinal or system-level studies
Comments	Main idea of the EPIS framework: the implementation of an evidence-based practice depends not only on the quality of the intervention itself, but also on organisational context, policy environments, funding structures, leadership, the stakeholders involved, and the interactions between external and internal systems Strong temporal perspective on implementation Explicit consideration of multilevel and cross-phase interactions Particularly useful for complex, system-wide interventions

Source: Aarons et al.²⁵⁴

Table A. 7 – Proctor’s Implementation Outcomes Framework

Proctor’s implementation outcomes framework	
Origin of the TMF	Developed by Enola Proctor and colleagues (2011)
Objectives	- Define and standardise key ‘implementation outcomes’ (as indicators of implementation success and proximal indicators of implementation process) - Distinguish implementation outcomes from service and patient outcomes - Support rigorous evaluation of implementation processes and strategies - Improve comparability and measurement across implementation studies
Main components	Heuristic, working “taxonomy” of eight implementation outcomes: Acceptability: perception among stakeholders that an intervention is agreeable or satisfactory Adoption: intention, decision or action to use an intervention Appropriateness: perceived fit, relevance or compatibility Feasibility: extent to which an intervention can be successfully used or carried out Fidelity: degree to which an intervention is implemented as intended Implementation cost: costs associated with implementation efforts



	- Penetration: integration of a practice within a service setting and its subsystems - Sustainability: extent to which the intervention is maintained or institutionalised over time → Selection depends on study objectives
When to use?	To structure the assessment of implementation success, particularly in implementation research and process evaluations
Comments	Focuses specifically on implementation outcomes rather than intervention effectiveness, and help to structure the implementation results (acceptability, feasibility, fidelity, etc.) May be used alongside other frameworks (e.g. RE-AIM or logic models) Applicable prospectively and retrospectively

Sources: Proctor et al.²⁵⁵, the website of the University of Colorado Denver (Dissemination & Implementation models in health)⁴⁵³, and the website Implementation Outcome Repository⁴⁵⁴

Appendix 3.5. Focus on adaptation and fidelity in implementation: ‘adaptation and fidelity’ TMFs

Table A. 8 – FRAME: Framework for Reporting Adaptations and Modifications-Enhanced

FRAME	
Origin of the TMF	Developed by Laura Stirman and colleagues (2013; updated as the Framework for Reporting Adaptations and Modifications Expanded (FRAME) in 2019)
Objectives	▪ Systematically document adaptations and modifications to interventions ▪ Improve transparency in reporting changes made during implementation ▪ Distinguish between planned and unplanned adaptations ▪ Support interpretation of intervention effectiveness in real-world settings
Main components	1. Timing of the adaptation - Before implementation - During implementation - Sustainment/maintenance phase (if applicable) 2. Content of the modification (what is modified) - Core elements of the intervention - Contextual elements - Format of delivery - Frequency/dose/intensity - Sequence of intervention components - Personnel delivering the intervention 3. Level at which adaptation occurs - Individual level (patient/client) - Provider level - Team/service level - Organisational level - System/policy level 4. Type of modification - Addition of components - Removal of components - Substitution of components - Reordering/sequencing changes - Shortening or extending components - Tailoring/cultural adaptation - Mode of delivery change (e.g., in-person vs digital) 5. Who decided the adaptation - Researchers/intervention developers - Clinicians/practitioners - Organisation/service managers - Policymakers/external stakeholders - End-users (patients/clients) 6. Rationale for adaptation (why)



	- Improve fit with context or population - Increase feasibility - Improve acceptability - Address resource constraints - Improve reach or equity - Enhance fidelity or adherence - Respond to external policy or organisational requirements 7. Planned vs unplanned adaptation - Planned (proactive, pre-specified) - Unplanned (emergent, reactive during implementation) 8. Fidelity to core intervention components - Maintained core components - Partially modified core components - Compromised/removed core components - Whether fidelity-consistent modifications were maintained → Structured framework for documenting and reporting adaptations. Not all adaptation dimensions need to be assessed in every study
When to use?	The FRAME provides a structured framework for systematically documenting when adaptations occur, what is modified, at what level, by whom, for what reason, whether adaptations are planned or unplanned, their extent, and their implications for fidelity and outcomes → To report and analyse adaptations made to interventions during design, implementation, or evaluation phases Can be used as data collection form or reporting tool
Comments	Helps distinguish between fidelity and adaptation Useful in pragmatic and real-world implementation studies May be used alongside implementation frameworks such as CFIR

Sources: Stirman *et al.*^{256, 257}

Table A. 9 – FRAME-IS: Framework for Reporting Adaptations and Modifications to Evidence-based Implementation Strategies

FRAME-IS	
Origin of the TMF	Developed by Christopher Miller and colleagues (2021)
Objectives	▪ Extend FRAME to implementation strategies rather than interventions ▪ Systematically capture modifications to implementation processes ▪ Improve reporting of how implementation strategies evolve over time ▪ Enhance reproducibility and interpretation of implementation outcomes
Main components	Core components often called “modules” Description of the evidence-based practice, implementation strategy, and modification(s) - What is changed? - Intervention, implementation strategy, both - What is the nature of the content, evaluation, or training modification? - Content, evaluation/feedback, training, delivery/format, intensity/modality - What is the rationale for the modification? (why is the modification implemented?) - Fits with the context, feasibility, resources, acceptability, performance/effectiveness When the modification is initiated, and whether it is planned - Before implementation, during implementation, after start - Planned (proactive) or unplanned (reactive) Who participates in the decision to modify? - Researchers, clinicians, managers/leadership, stakeholders, users - How widespread is the modification (scope of the modification)? - One-site vs multisite, individual vs organisational, local vs systemic, spread of the modification within the system → Flexible framework. Modules may be used selectively depending on context.



When to use?	To document and analyse adaptations made to implementation strategies (rather than interventions themselves)
Comments	The FRAME-IS is designed to document changes made to a specific implementation strategy that has been defined in advance → To properly track these changes, it is important to clearly define what the original strategy includes Complements FRAME by focusing specifically on implementation strategies, may be used alongside assessment of implementation outcomes (e.g., using RE-AIM or similar evaluation-oriented implementation frameworks) Can support understanding of dynamic and iterative implementation processes. May be useful in complex and adaptive implementation contexts

Source: Miller et al.²⁵⁸

■ REFERENCES

1. Federaal Kenniscentrum voor de Gezondheidszorg (KCE). Studie 2023-53 (HSR) Evaluatie van het proefproject voor de implementatie van het New Deal-model [Web page].2026 [cited 04/06/2026]. Available from: <https://www.kce.fgov.be/nl/studie-2023-53-hsr-evaluatie-van-het-proefproject-voor-de-implementatie-van-het-new-deal-model>
2. Van den Heede K, Camberlin C, Detollenaere J, Kohn L, Leroy R, Van Deynse H, et al. Pilot study on a new payment system for home nursing. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2025. KCE Reports (405C)
3. Federaal Kenniscentrum voor de Gezondheidszorg (KCE). Process Book [Web page].2026 [cited 04/06/2026]. Available from: <https://processbook.kce.be/>
4. Sermeus W. Complex Intervention Research: What, Why, and How. In: Kaehne A, Feather J, editors. Handbook of Health Services Evaluation: Theories, Methods and Innovative Practices. Cham, Switzerland: Springer; 2025.
5. Curran GM, Landes SJ, McBain SA, Pyne JM, Smith JD, Fernandez ME, et al. Reflections on 10 years of effectiveness-implementation hybrid studies. Front Health Serv. 2022;2:1053496.
6. Kaehne A, Feather J. Handbook of Health Services Evaluation: Theories, Methods and Innovative Practices. Amsterdam: Springer; 2025.
7. Richards DA. Complex Interventions in Health: An Overview of Research Methods. 2nd ed.: Taylor & Francis; 2026.
8. Richards DA, Hallberg IR. Complex Interventions in Health: An Overview of Research Methods. Richards DA, Hallberg IR, editor. Oxon, United Kingdom: Routledge; 2015.
9. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061.
10. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Int J Nurs Stud. 2024;154:104705.
11. Craig P, Di Ruggiero E, Frohlich KL, Mykhalovskiy E, White M, on behalf of the Canadian Institutes of Health Research (CIHR)–National Institute for Health Research (NIHR) Context Guidance Authors Group. Taking account of context in population health intervention research: guidance for producers, users and funders of research. Southampton, UK: 2018.
12. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med. 2018;16(1):95.
13. Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised controlled trial be? BMJ. 2004;328(7455):1561-3.
14. Hawe P, Shiell A, Riley T. Theorising interventions as events in systems. Am J Community Psychol. 2009;43(3-4):267-76.
15. Rutter H, Savona N, Glonti K, Bibby J, Cummins S, Finegood DT, et al. The need for a complex systems model of evidence for public health. Lancet. 2017;390(10112):2602-4.
16. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. Int J Nurs Stud. 2013;50(5):587-92.
17. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ. 2008;337:a1655.
18. Lawless A, Baum F, Delany-Crowe T, MacDougall C, Williams C, McDermott D, et al. Developing a Framework for a Program Theory-Based Approach to Evaluating Policy



- Processes and Outcomes: Health in All Policies in South Australia. *Int J Health Policy Manag.* 2018;7(6):510-21.
19. O'Cathain A, Croot L, Duncan E, Rousseau N, Sworn K, Turner KM, et al. Guidance on how to develop complex interventions to improve health and healthcare. *BMJ Open.* 2019;9(8):e029954.
 20. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ.* 2000;321(7262):694-6.
 21. Harris P. Researching Healthy Public Policy: Navigating the 'Black Box' Means Thinking More About Power Comment on "Developing a Framework for a Program Theory-Based Approach to Evaluating Policy Processes and Outcomes: Health in All Policies in South Australia". *Int J Health Policy Manag.* 2018;7(9):874-6.
 22. Clarke GM, Conti S, Wolters AT, Steventon A. Evaluating the impact of healthcare interventions using routine data. *BMJ.* 2019;365:l2239.
 23. Moore G, Audrey S, Barker M, Bond L, Bonell C, Cooper C, et al. Process evaluation in complex public health intervention studies: the need for guidance. *J Epidemiol Community Health.* 2014;68(2):101-2.
 24. Moore G, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions. UK Medical Research Council (MRC) guidance. 2015. Available from: <https://mrc.ukri.org/documents/pdf/mrc-phsrn-process-evaluation-guidance-final/>
 25. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. *BMJ.* 2015;350:h1258.
 26. Renmans D, Castellano Pleguezuelo V. Methods in realist evaluation: A mapping review. *Eval Program Plann.* 2023;97:102209.
 27. Palm R, Hochmuth A. What works, for whom and under what circumstances? Using realist methodology to evaluate complex interventions in nursing: A scoping review. *Int J Nurs Stud.* 2020;109:103601.
 28. Sørensen OH, Garn SD, Nielsen SB. And then a miracle occurs—a review of theory of change models for societal impact of research. *Research Evaluation.* 2025;34.
 29. Ryom K, Mygind A, Terkildsen Maindal H. The use of programme theory in intervention development. In: David A. Richards IRH, Carole A. Estabrooks, Sascha Köpke, Gabriele Meyer, Lars Wallin, Yvonne Wengström, and Joanne Woodford, editor. *Complex interventions in health: an overview of research methods* (second edition). London: Routledge; 2026.
 30. Kneale D, Bangpan M, Thomas J, HS W. Chapter 4.10 Using logic models in research and evaluation of Health EDRM interventions. In: WHO guidance on research methods for health emergency and disaster risk management; 2022. Available from: <https://wkc.who.int/our-work/health-emergencies/research-methods/sections-and-chapters/section-4/chapter-4-10-logic-models>
 31. BetterEvaluation. Using logic models and theories of change better in evaluation [Web page].2017 [cited 23/02/2026]. Available from: <https://www.betterevaluation.org/blog/using-logic-models-theories-change-better-evaluation>
 32. Early Intervention Foundation. What is a theory of change? [Web page].2021 [cited 23/02/2026]. Available from: <https://evaluationhub.eif.org.uk/theory-of-change/>
 33. Breuer E, Lee L, De Silva M, Lund C. Using theory of change to design and evaluate public health interventions: a systematic review. *Implement Sci.* 2016;11:63.
 34. Funnell S, Rogers P. *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models.* 2011.
 35. Brousselle A, Champagne F. Program theory evaluation: Logic analysis. *Evaluation and Program Planning.* 2011;34(1):69-78.

36. Van den Heede K, Camberlin C, Detollenaere J, Kohn L, Leroy R, Van Deynse H, et al. Evaluatie van een nieuwe financiering voor thuisverpleging. Health Services Research (HSR). Brussel: Federaal Kenniscentrum voor de Gezondheidszorg (KCE); 2025. KCE Reports (405A)
37. Sharp S. Logic Models vs Theories of Change [Web page].2021 [cited 23/02/2026]. Available from: <https://cere.olemiss.edu/logic-models-vs-theories-of-change/>
38. Orozco C. Logic Models vs Theories of Change: What is the difference? [Web page]. UNM Evaluation Lab Summer Institute 2022 [cited 23/02/2026]. Available from: <https://evallab.unm.edu/lmt.pdf>
39. Voss S, Bauer J, Coenen M, Jung-Sievers C, Moore G, Rehfuss E. Logic models for the evaluation of complex interventions in public health: lessons learnt from a staged development process. BMC Public Health. 2025;25(1):1923.
40. W.K. Kellogg Foundation. Logic Model Development Guide. Battle Creek, MI: 2004.
41. Tapplin. Theory of Change: TECHNICAL PAPERS. 2013.
42. McGill S, Borghi J. Theory of Change: a tool to guide the evaluation of complex health interventions [Web page].London;2015. Available from: <https://resyst.lshtm.ac.uk/resources/theory-of-change-a-tool-to-guide-the-evaluation-of-complex-health-interventions>
43. Centre for Epidemiology and Evidence. Developing and Using Program Logic: A Guide. Sydney: NSW Ministry of Health; 2023.
44. McLaughlin JA, Jordan GB. Logic models: a tool for telling your programs performance story. Evaluation and program planning. 1999;22(1):65-72.
45. Manchester University. Creating your own Theory of Change [Web page].n.d. Available from: <https://assets.bmh.manchester.ac.uk/evaluation/guides/about-theory-of-change/content/index.html#/lessons/WwyhG6EjKWme51GAX1bEANjpgdgHQzxv>
46. Fridrich A, Jenny GJ, Bauer GF. The Context, Process, and Outcome Evaluation Model for Organisational Health Interventions. Biomed Res Int. 2015;2015:414832.
47. De Silva MJ, Breuer E, Lee L, Asher L, Chowdhary N, Lund C, et al. Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. Trials. 2014;15:267.
48. BetterEvaluation. Theory of change software [Web page].2023 [cited 23/02/2026]. Available from: <https://www.betterevaluation.org/tools-resources/theory-change-software>
49. Tsantila F, Coppens E, De Witte H, Abdulla K, Amann BL, Arensman E, et al. Developing a framework for evaluation: a Theory of Change for complex workplace mental health interventions. BMC Public Health. 2023;23(1):1171.
50. Sewell J, Rodger A, Sherr L, Currie R, Sparrowhawk A, Lampe FC, et al. Theory of change to develop a health and well-being complex intervention for people with HIV. AIDS Care. 2026;38(2):263-74.
51. Nunez LPC, Martinez GM, Gallo EG. Analysis of a theory of change to evaluate the Health Care-Associated Infection Prevention Program (HAI) in Colombia. BMC Health Serv Res. 2025;25(1):442.
52. Maini R, Mounier-Jack S, Borghi J. How to and how not to develop a theory of change to evaluate a complex intervention: reflections on an experience in the Democratic Republic of Congo. BMJ Glob Health. 2018;3(1):e000617.
53. Remawi BN, Preston N, Gadoud A. Development of a complex palliative care intervention for patients with heart failure and their family carers: a theory of change approach. BMC Palliat Care. 2025;24(1):129.
54. Mugotitsa B, Andeso P, Owoko H, Momanyi R, Mailosi D, Ssebunnya J, et al. Enhancing mental health outcomes through a theory of change: supporting open-source data science



- platforms to promote FAIR data and evidence-based decisions. *BMC Health Serv Res.* 2025;25(1):1239.
55. Luna Pinzon A, Stronks K, Dijkstra C, Renders C, Altenburg T, den Hertog K, et al. The ENCOMPASS framework: a practical guide for the evaluation of public health programmes in complex adaptive systems. *Int J Behav Nutr Phys Act.* 2022;19(1):33.
 56. Hudon C, Chouinard MC, Brousselle A, Bisson M, Danish A. Evaluating complex interventions in real context: Logic analysis of a case management program for frequent users of healthcare services. *Eval Program Plann.* 2020;79:101753.
 57. Hawkes RE, Miles LM, French DP. The theoretical basis of a nationally implemented type 2 diabetes prevention programme: how is the programme expected to produce changes in behaviour? *Int J Behav Nutr Phys Act.* 2021;18(1):64.
 58. Bugge SJ, Hangaard J, Damkier P, Jensen PB, Johansson SL, Nagarajah S, et al. Network of doctors for multimorbidity and diabetes: Development of the NOMAD intervention - a multidisciplinary team for patients with diabetes and multimorbidity. *J Multimorb Comorb.* 2025;15.
 59. Esmail LC, Barasky R, Mittman BS, Hickam DH. Improving Comparative Effectiveness Research of Complex Health Interventions: Standards from the Patient-Centered Outcomes Research Institute (PCORI). *J Gen Intern Med.* 2020;35(Suppl 2):875-81.
 60. Bonell C, Jamal F, Melendez-Torres GJ, Cummins S. 'Dark logic': theorising the harmful consequences of public health interventions. *J Epidemiol Community Health.* 2015;69(1):95-8.
 61. Sermeus W. Modelling process and outcomes in complex interventions. In: Richards D, Rahm Hallberg I, editors. *Complex interventions in health: an overview of research methods.* London: Routledge; 2015.
 62. Shakman K, Rodriguez S. Logic models for program design, implementation, and evaluation: Workshop toolkit. U.S. Department of Education - Institute of Educational Sciences; 2015 Available from: https://ies.ed.gov/ncee/rel/regions/northeast/pdf/rel_2015057.pdf
 63. Centre for Epidemiology and Evidence. *Study Design for Evaluating Population Health and Health Service Interventions: A Guide.* 2023
 64. Berzins K, Lowther-Payne H, Christian D. An introduction to logic models [Web page]. University of Central Lancashire & NIHR; n.d. [cited 24/02/2026]. Available from: <https://arc-nwc.nihr.ac.uk/wp-content/uploads/2023/03/Logic-Models-presentation.pdf>
 65. Onyura B, Mullins H, Hamza DM. Five ways to get a grip on the shortcomings of logic models in program evaluation. *Can Med Educ J.* 2021;12(6):96-9.
 66. World Health Organization. Clinical trials [Web page]. 2020 [cited 18 May 2026]. Available from: <https://www.who.int/news-room/questions-and-answers/item/clinical-trials>
 67. World Health Organization. *The World Health Report 2000: Health Systems: Improving Performance.* Geneva: World Health Organization; 2000. Available from: https://cdn.who.int/media/docs/default-source/health-financing/whr-2000.pdf?sfvrsn=95d8b803_1&download=true
 68. Curtis L, Hernandez AF, Weinfurt KP. Choosing and Specifying Endpoints and Outcomes: Introduction [Web page]. [cited 26 May 2026]. Available from: <https://rethinkingclinicaltrials.org/chapters/design/choosing-specifying-end-points-outcomes/choosing-and-specifying-endpoints-and-outcomes-introduction/>
 69. European Commission Directorate-General for Health and Food Safety. *Guidance on Validity of Clinical Studies* [Web page]. 2024 [updated 22 January 2024; cited 11 March 2026]. Available from: https://health.ec.europa.eu/publications/guidance-validity-clinical-studies-joint-clinical-assessments_en

70. Leroy JL, Frongillo EA, Kase BE, Alonso S, Chen M, Dohoo I, et al. Strengthening causal inference from randomised controlled trials of complex interventions. *BMJ Glob Health*. 2022;7(6).
71. Roland M, Torgerson DJ. What are pragmatic trials? *BMJ*. 1998;316(7127):285.
72. Roland M, Torgerson D. Understanding controlled trials: what outcomes should be measured? *BMJ*. 1998;317(7165):1075.
73. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Clin Epidemiol*. 2009;62(5):499-505.
74. Sedgwick P. Explanatory trials versus pragmatic trials. *BMJ*. 2014;349:g6694.
75. Turner KM, Huntley A, Yardley T, Dawson S, Dawson S. Defining usual care comparators when designing pragmatic trials of complex health interventions: a methodology review. *Trials*. 2024;25(1):117.
76. Kalkman S, van Thiel G, Grobbee DE, van Delden JJM. Pragmatic clinical trials: ethical imperatives and opportunities. *Drug Discov Today*. 2018;23(12):1919-21.
77. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med*. 1987;317(3):141-5.
78. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*. 2015;350:h2147.
79. Centre for Evidence-Based Medicine (CEBM) University of Oxford. Catalogue of Bias [Web page]. [cited 11 March 2026]. Available from: <https://catalogofbias.org/>
80. Creanor S. Evaluating efficacy and effectiveness in the context of complex interventions. In: Richards DA, Hallberg IR, editors. *Complex Interventions in Health*. Second edition ed. 4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN and 605 Third Avenue, New York, NY 10158: Routledge; 2026. p. 177-96.
81. Liu Y, Zhou K, Sims K. Tipping Point Analysis: Assessing the Potential Impact of Missing Data. *JAMA*. 2025;334(3):265-6.
82. The SPIRIT-CONSORT Group. The SPIRIT-CONSORT guidelines [Web page]. 2025 [cited 11 March 2026]. Available from: <https://www.consort-spirit.org/>
83. Lachin JM. Statistical properties of randomization in clinical trials. *Control Clin Trials*. 1988;9(4):289-311.
84. Buyse M. Centralized Treatment Allocation in Comparative Clinical Trials. *Applied Clinical Trials*. 2000;9(6):32-7.
85. Cook TD, DeMets DL. *Introduction to Statistical Methods for Clinical Trials*. CHAPMAN & HALL/CRC; 2008.
86. Kahan BC, Jairath V, Dore CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15:139.
87. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol*. 1999;52(1):19-26.
88. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH E9 statistical principles for clinical trials - Scientific guideline [Web page]. 1998 [cited 11 March 2026]. Available from: <https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline>
89. Coart E, Bamps P, Quinaux E, Sturbois G, Saad ED, Burzykowski T, et al. Minimization in randomized clinical trials. *Stat Med*. 2023;42(28):5285-311.
90. Browner WS. *Designing clinical research*. Fifth Edition ed.: Lippincott Williams & Wilkins; 2023.
91. Pocock SJ. *Clinical Trials: A Practical Approach*. The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons Ltd; 1983.



92. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Soc Sci Med*. 2012;75(12):2299-306.
93. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. *BMC Med Res Methodol*. 2003;3:26.
94. Mdege ND, Brabyn S, Hewitt C, Richardson R, Torgerson DJ. The 2 x 2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: a systematic review. *J Clin Epidemiol*. 2014;67(10):1083-92.
95. Montgomery AA, Astin MP, Peters TJ. Reporting of factorial trials of complex interventions in community settings: a systematic review. *Trials*. 2011;12:179.
96. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Stat Med*. 1989;8(12):1421-32.
97. Sibbald B, Roberts C. Understanding controlled trials. Crossover trials. *BMJ*. 1998;316(7146):1719.
98. Hainaut P. The Gambia hepatitis intervention study (GHIS). *SUPLEMENTO IATREIA*. 2007;20:S21-2.
99. Leupold F, Karimzadeh A, Breitzkreuz T, Draht F, Klidis K, Grobe T, et al. Digital redesign of hypertension management with practice and patient apps for blood pressure control (PIA study): A cluster-randomised controlled trial in general practices. *EClinicalMedicine*. 2023;55:101712.
100. Peden CJ, Stephens T, Martin G, Kahan BC, Thomson A, Rivett K, et al. Effectiveness of a national quality improvement programme to improve survival after emergency abdominal surgery (EPOCH): a stepped-wedge cluster-randomised trial. *Lancet*. 2019;393(10187):2213-21.
101. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*. 2015;350:h391.
102. Leontjevas R, Gerritsen DL, Smalbrugge M, Teerenstra S, Vernooij-Dassen MJ, Koopmans RT. A structural multidisciplinary approach to depression management in nursing-home residents: a multicentre, stepped-wedge cluster-randomised trial. *Lancet*. 2013;381(9885):2255-64.
103. Coldiron M, Grais RF. Cluster randomized trials (Chapter 4.3) [Web page]. World Health Organization (WHO);2022. Available from: <https://wkc.who.int/our-work/health-emergencies/research-methods/sections-and-chapters/section-4/chapter-4-3-cluster-randomized-trials>
104. Dron L, Taljaard M, Cheung YB, Grais R, Ford N, Thorlund K, et al. The role and challenges of cluster randomised trials for global health. *Lancet Glob Health*. 2021;9(5):e701-e10.
105. Weijer C, Grimshaw JM, Eccles MP, McRae AD, White A, Brehaut JC, et al. The Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials. *PLoS Med*. 2012;9(11):e1001346.
106. Goldstein CE, du Toit J, Murphy NB, Nicholls SG, Shaw JF, Althabe F, et al. Gaps in the Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials: a citation analysis reveals a need for updated ethics guidelines. *Res Integr Peer Rev*. 2025;10(1):10.
107. Moerbeek M, Teerenstra S. *Power Analysis of Trials with Multilevel Data*. Boca Raton: CRC Press; 2016.
108. Crespi CM. *Power and Sample Size in R*. Boca Raton: CRC Press; 2025.
109. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. *J Clin Epidemiol*. 2012;65(12):1249-52.



110. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Researchers should convince policy makers to perform a classic cluster randomized controlled trial instead of a stepped wedge design when an intervention is rolled out. *J Clin Epidemiol*. 2012;65(12):1255-6.
111. Piantadosi S, Meinert CL. Principles and Practice of Clinical Trials. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer Nature Switzerland AG; 2022.
112. Kessels R, May AM, Koopman M, Roes KCB. The Trial within Cohorts (TwICs) study design in oncology: experience and methodological reflections. *BMC Med Res Methodol*. 2023;23(1):117.
113. Krishnan P. A review of the non-equivalent control group post-test-only design. *Nurse Researcher*. 2018.
114. Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhu J, Peterson DE, et al. The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American medical informatics association*. 2006;13(1):16-23.
115. White H, Sabarwal S. Quasi-Experimental Design and Methods. Florence: UNICEF Office of Research; 2014. Methodological Briefs: Impact Evaluation 8
116. American Psychological Association. quasi-experimental design [Web page].2018. Available from: <https://dictionary.apa.org/quasi-experimental-design>
117. Gertler PJ, Martinez S, Premand P, Rawlings LB, Vermeersch CMJ. Impact Evaluation in Practice. Second Edition. Worldbank Group; 2016.
118. Craig P, Katikireddi SV, Leyland A, Popham F. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annual review of public health*. 2017;38:39-56.
119. McKenna C, Morrison A. Guide 3: Natural experiments Available from: <https://www.gov.scot/binaries/content/documents/govscot/publications/advice-and-guidance/2009/12/social-research-methods-guides/documents/natural-experiments/natural-experiments/govscot%3Adocument/Natural%2BExperiments.pdf>
120. Leatherdale ST. Natural experiment methodology for research: a review of how different methods can support real-world research. *International Journal of Social Research Methodology*. 2019;22(1):19-35.
121. Fu EL. Target trial emulation to improve causal inference from observational data: what, why, and how? *Journal of the American Society of Nephrology*. 2023;34(8):1305-14.
122. Tran V-T, Porcher R, Perrodeau E, Yaacoub S, Boutron I, Ravaud P. Practical elements to consider when emulating a target trial. *Journal of Clinical Epidemiology*. 2026;194:112205.
123. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183(8):758-64.
124. Hernán MA, Wang W, Leaf DE. Target trial emulation: a framework for causal inference from observational data. *Jama*. 2022;328(24):2446-7.
125. Ben-Michael E, Feller A, Stuart EA. A Trial Emulation Approach for Policy Evaluations with Group-level Longitudinal Data. *Epidemiology*. 2021;32(4):533-40.
126. Seewald NJ, McGinty EE, Stuart EA. Target trial emulation for evaluating health policy. *Annals of internal medicine*. 2024;177(11):1530-8.
127. Andrew NE, Ung D, Olaiya MT, Dalli LL, Kim J, Churilov L, et al. The population effect of a national policy to incentivize chronic disease management in primary care in stroke: a population-based cohort study using an emulated target trial approach. *The Lancet Regional Health – Western Pacific*. 2023;34.
128. University of Oxford. Catalogue of Bias [Web page].Oxford. Available from: <https://catalogofbias.org>
129. Siedlecki L, Siedlecki SL. Using Research to Advance Nursing Practice. 2019.



130. Benjamin-Chung J, Arnold BF, Berger D, Luby SP, Miguel E, Colford Jr JM, et al. Spillover effects in epidemiology: parameters, study designs and methodological considerations. *International journal of epidemiology*. 2018;47(1):332-47.
131. Polit D, Beck C. *Nursing Research: Generating and Assessing Evidence for Nursing Practice*. 10th Edition ed. Philadelphia: Wolters Kluwer; 2017.
132. Fahmy JN, Cichocki MN, Chung KC. Quasi-experimental design for health policy research: a methodology overview. *Plastic and Reconstructive Surgery*. 2023;151(3):667-75.
133. Ye K, Bilinski A, Lee Y. Difference-in-differences analysis with repeated cross-sectional survey data. *Health Services and Outcomes Research Methodology*. 2025:1-19.
134. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama*. 2014;312(22):2401-2.
135. Bilinski A, Feng S, Hatfield LA. Estimating Intervention Effects With Difference-in-Differences. *JAMA*. 2026;336(6):508-9.
136. Bosco BP, Maranzano P. Vis Inertiae and Statistical Inference: A Review of Difference-in-Differences Methods Employed in Economics and Other Subjects. *Econometrics*. 2025;13(4):38.
137. Dahabreh IJ, Lewis RJ. Causal Language for Studies Using Difference-in-Differences Analyses. *JAMA*. 2026;336(6):462-3.
138. Columbia Mailman School of Public Health. Difference-in-Difference Estimation [Web page]. Available from: <https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation>
139. Sant'Anna PH, Xu Q. Difference-in-differences with compositional changes. *Journal of Econometrics*. 2026;253:106147.
140. Maciejewski ML, Basu A. Regression discontinuity design. *Jama*. 2020;324(4):381-2.
141. The Academy of Medical Sciences. *Identifying the environmental causes of disease: how should we decide what to believe and when to take action?* London: The Academy of Medical Sciences; 2007.
142. Feir D, Lemieux T, Marmer V. Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*. 2016;34(2):185-96.
143. Hernán M, Robins J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2025.
144. Gail MH. Frequency Matching. In: *Wiley StatsRef: Statistics Reference Online*. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05163>
145. Iwagami M, Shinozaki T. Introduction to matching in case-control and cohort studies. *Annals of Clinical Epidemiology*. 2022;4(2):33-40.
146. Columbia Mailman School of Public Health. Propensity Score Analysis [Web page]. Available from: <https://www.publichealth.columbia.edu/research/population-health-methods/propensity-score-analysis>
147. Imbens G, Rubin D. *Causal inference for statistics, social, and biomedical sciences: an introduction*. New York: Cambridge University Press; 2015.
148. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*. 2010;25(1):1.
149. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*. 2011;46(3):399-424.
150. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*. 2005;34(1):215-20.



151. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology*. 2017;46(1):348-55.
152. Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The Multiple Baseline Design for Evaluating Population-Based Research. *American Journal of Preventive Medicine*. 2007;33(2):162-8.
153. Kennedy CH. The nonconcurrent multiple-baseline design: It is what it is and not something else. *Perspectives on Behavior Science*. 2022;45(3):647-50.
154. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA, US: Houghton, Mifflin and Company; 2002.
155. Carter KN, Imlach-Gunasekara F, McKenzie SK, Blakely T. Differential loss of participants does not necessarily cause selection bias. *Australian and New Zealand journal of public health*. 2012;36(3):218-22.
156. Sax Institute. *Translational Research Framework: Source Book*. Sydney: 2016. Available from: https://www.saxinstitute.org.au/wp-content/uploads/Translation-Research-Framework_Source-Book.pdf
157. Krass I. Quasi experimental designs in pharmacist intervention research. *International journal of clinical pharmacy*. 2016;38(3):647-54.
158. Sanson-Fisher RW, D'Este CA, Carey ML, Noble N, Paul CL. Evaluation of systems-oriented public health interventions: alternative research designs. *Annual review of public health*. 2014;35:9-27.
159. Penfold RB, Zhang F. Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements. *Academic Pediatrics*. 2013;13(6, Supplement):S38-S44.
160. Christ TJ. Experimental control and threats to internal validity of concurrent and nonconcurrent multiple baseline designs. *Psychology in the Schools*. 2007;44(5):451-9.
161. Rhoda DA, Murray DM, Andridge RR, Pennell ML, Hade EM. Studies with staggered starts: multiple baseline designs and group-randomized trials. *American Journal of Public Health*. 2011;101(11):2164-9.
162. Kreif N, Grieve R, Hangartner D, Turner AJ, Nikolova S, Sutton M. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics*. 2016;25(12):1514-28.
163. Bouttell J, Craig P, Lewsey J, Robinson M, Popham F. Synthetic control methodology as a tool for evaluating population-level health interventions. *J Epidemiol Community Health*. 2018;72(8):673-8.
164. Klößner S, Kaul A, Pfeifer G, Schieler M. Comparative politics and the synthetic control method revisited: a note on Abadie et al. (2015). *Swiss Journal of Economics and Statistics*. 2018;154(1):11.
165. O'Neill S, Kreif N, Grieve R, Sutton M, Sekhon JS. Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*. 2016;16(1):1-21.
166. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*. 2010;105(490):493-505.
167. Hill AB. *The environment and disease: association or causation?* Sage Publications; 1965. (0035-9157)
168. Pearl J, Mackenzie D. *The book of why*. 2018.
169. Rogers P. *Overview: Strategies for Causal Attribution*. Florence: UNICEF Office of Research; 2014. *Methodological Briefs: Impact Evaluation* 6



170. Tappen R. *Advanced Nursing Research: From Theory to Practice*. 2nd Edition ed. Burlington: Jones and Bartlett Learning; 2015.
171. Chatton A, Rohrer JM. The causal cookbook: Recipes for propensity scores, g-computation, and doubly robust standardization. *Advances in Methods and Practices in Psychological Science*. 2024;7(1):25152459241236149.
172. Cunningham S. *Causal inference: The mixtape*. Yale university press; 2021.
173. Lawes M, West SG, Eid M. A guide to causal inference in life-event studies. *Advances in Methods and Practices in Psychological Science*. 2025;8(1):25152459241302015.
174. Sourial N, Longo C, Vedel I, Schuster T. Daring to draw causal claims from non-randomized studies of primary care interventions. *Family Practice*. 2018;35(5):639-43.
175. Kainz K, Greifer N, Givens A, Swietek K, Lombardi BM, Zietz S, et al. Improving causal inference: recommendations for covariate selection and balance in propensity score methods. *Journal of the Society for Social Work and Research*. 2017;8(2):279-303.
176. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American journal of epidemiology*. 2006;163(12):1149-56.
177. da Costa BR, Gahl B, Jueni P. Tools & techniques—statistics: propensity score techniques. *EuroIntervention*. 2014;10(6):761-7.
178. Connelly BS, Sackett PR, Waters SD. Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology*. 2013;66(2):407-42.
179. Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, et al. An introduction to inverse probability of treatment weighting in observational research. *Clinical kidney journal*. 2022;15(1):14-20.
180. Kühne F, Schomaker M, Stojkov I, Jahn B, Conrads-Frank A, Siebert S, et al. Causal evidence in health decision making: methodological approaches of causal inference and health decision science. *GMS German Medical Science*. 2022;20:Doc12.
181. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *American journal of epidemiology*. 2011;173(7):761-7.
182. Raad H, Cornelius V, Chan S, Williamson E, Cro S. An evaluation of inverse probability weighting using the propensity score for baseline covariate adjustment in smaller population randomised controlled trials with a continuous outcome. *BMC Medical Research Methodology*. 2020;20(1):70.
183. John E, Abrams K, Brightling C, Sheehan N. Assessing causal treatment effect estimation when using large observational datasets. *BMC medical research methodology*. 2019;19(1):207.
184. Wang A, Nianogo RA, Arah OA. G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology*. 2017;17(1):3.
185. Pirracchio R, Petersen ML, Van Der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology*. 2015;181(2):108-19.
186. Le Borgne F, Chatton A, Léger M, Lenain R, Foucher Y. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. *Scientific reports*. 2021;11(1):1435.
187. Li X, Shen C. Doubly robust estimation of causal effect: upping the odds of getting the right answers. *Circulation: Cardiovascular Quality and Outcomes*. 2020;13(1):e006065.
188. Tan Z. Comment: Understanding or, PS and DR. *Statistical Science*. 2007;22(4):560-8.
189. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. 2004;23(19):2937-60.



190. Widding-Havneraas T, Zachrisson HD. A Gentle Introduction to Instrumental Variables. *Journal of Clinical Epidemiology*. 2022;149:203-5.
191. Aso S, Yasunaga H. Introduction to instrumental variable analysis. *Annals of Clinical Epidemiology*. 2020;2(3):69-74.
192. Feeney T, Hartwig FP, Davies NM. How to use directed acyclic graphs: guide for clinical researchers. *BMJ*. 2025;388:e078226.
193. Kutcher SA, Brophy JM, Banack HR, Kaufman JS, Samuel M. Emulating a randomised controlled trial with observational data: an introduction to the target trial framework. *Canadian Journal of Cardiology*. 2021;37(9):1365-77.
194. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med*. 2010;29(30):3245-57.
195. Agresti A. *Categorical Data Analysis*. Second Edition ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2002.
196. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):1080.
197. Molenberghs G, Buyse M, Burzykowski T. 1. Introduction. In: Burzykowski T, Molenberghs G, Buyse M, editors. *The Evaluation of Surrogate Endpoints*. 233 Spring Street, New York, NY 10013, USA: Springer Science+Business Media, Inc.; 2005.
198. Buyse M. I.1. Introduction. In: Alonso A, Bigirimurame T, Burzykowski T, Buyse M, Molenberghs G, Muchene L, et al., editors. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742: CRC Press Taylor & Francis Group; 2017.
199. Moore DM, G, Craig B. *Introduction to the Practice of Statistics*. 9th edition ed. New York, USA: W.H. Freeman and Company; 2017.
200. Akobeng AK. Understanding type I and type II errors, statistical power and sample size. *Acta Paediatrica*. 2016;105(6):605-9.
201. Bishara S. Adapting power calculations to include a superiority margin: what are the implications? *Biochem Med (Zagreb)*. 2024;34(1):010101.
202. Krousel-Wood MA, Chambers RB, Muntner P. Clinicians' guide to statistics for medical practice and research: Part II. *Ochsner Journal*. 2007;7(1):3-7.
203. Chow S-C, Shao J, Wang H, Lohknygina Y, Chow S-C, Shao J, et al. *Sample Size Calculations in Clinical Research: Third Edition*. 2017.
204. Gelman A, Carlin J. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on psychological science*. 2014;9(6):641-51.
205. Lakens D. Sample Size Justification. *Collabra: Psychology*. 2022;8(1).
206. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Controlled clinical trials*. 1989;10(4):407-15.
207. Wang Y, Devji T, Carrasco-Labra A, King MT, Terluin B, Terwee CB, et al. A step-by-step approach for selecting an optimal minimal important difference. *Bmj*. 2023;381:e073822.
208. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat Med*. 1990;9(1-2):65-71; discussion -2.
209. Cytel. *EAST 6 User Manual*. 2018.
210. Fitzmaurice GML, Nan M., Ware JH. *Applied Longitudinal Analysis*. 2nd edition ed. Balding DJ, Cressie NAC, Fitzmaurice GM, Goldstein H, Johnstone IM, Molenberghs G, et al., editor. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2011.
211. Piantadosi S. *Clinical Trials. A Methodologic Perspective*. Third Edition ed. 11 RiverStreet, Hoboken, NJ07030, USA: JohnWiley&Sons, Inc.; 2017.



212. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ*. 2001;323(7321):1123-4.
213. Bretz F, Hothorn T, Westfall P. Multiple Comparisons Using R. 1st Edition ed. New York: Chapman and Hall/CRC; 2011.
214. Dmitrienko A, D'Agostino R, Sr. Traditional multiplicity adjustment methods in clinical trials. *Stat Med*. 2013;32(29):5172-218.
215. U.S. Food and Drug Administration. Multiple Endpoints in Clinical Trials. Guidance for Industry. U.S. Department of Health and Human Services, Food and Drug Administration; 2022 October 2022. (FDA-2016-D-4460) Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials>
216. Cook JA. An Introduction to Clinical Trials. Great Clarendon Street, Oxford, OX2 6DP, United Kingdom: Oxford University Press; 2023.
217. Federal Drug Administration. Adaptive designs for clinical trials of drugs and biologics. Guidance for industry. 2019.
218. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. Fundamentals of Clinical Trials. Fifth Edition ed. Cham, Switzerland: Springer International Publishing; 2015.
219. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH E6(R3) Guideline for Good Clinical Practice (GCP)-Step 5. Amsterdam: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use; 2025. (EMA/CHMP/ICH/135/1995) Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e6-r3-guideline-good-clinical-practice-gcp-step-5_en.pdf
220. European Medicines Agency. Guideline on data monitoring committees. London: European Medicines Agency; 2005 27 July 2005. (EMA/CHMP/EWP/5872/03 Corr) Available from: <https://www.ema.europa.eu/en/data-monitoring-committees-scientific-guideline>
221. Herson J. Data and Safety Monitoring Committees in Clinical Trials. Second Edition ed. 6000 Broken Sound Parkway NW, Suite 300. Boca Raton, FL 33487-2742: CRC Press. Taylor & Francis Group; 2017.
222. Molero-Calafell J, Burón A, Castells X, Porta M. Intention to treat and per protocol analyses: differences and similarities. *Journal of Clinical Epidemiology*. 2024;173:111457.
223. Heinrich M, Zagorscak P, Bohn J, Knaevelsrud C, Schulze L. Using the ICH estimand framework to improve the interpretation of treatment effects in internet interventions. *npj Digital Medicine*. 2025;8(1):535.
224. Ahn E, Kang H. Intention-to-treat versus as-treated versus per-protocol approaches to analysis. *Korean Journal of Anesthesiology*. 2023;76(6):531-9.
225. Smith VA, Coffman CJ, Hudgens MG. Interpreting the results of intention-to-treat, per-protocol, and as-treated analyses of clinical trials. *Jama*. 2021;326(5):433-4.
226. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH E9(R1): Addendum on estimands and sensitivity analysis in clinical trials. London: European Medicines Agency; 2020. (ICH E9(R1)) Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf
227. Hopewell S, Chan AW, Collins GS, Hrobjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *BMJ*. 2025;389:e081123.
228. McGill E, Er V, Penney T, Egan M, White M, Meier P, et al. Evaluation of public health interventions from a complex systems perspective: A research methods review. *Soc Sci Med*. 2021;272:113697.



229. Minary L, Trompette J, Kivits J, Cambon L, Tarquinio C, Alla F. Which design to evaluate complex interventions? Toward a methodological framework through a systematic review. *BMC Med Res Methodol.* 2019;19(1):92.
230. French C, Dowrick A, Fudge N, Pinnock H, Taylor SJC. What do we want to get out of this? a critical interpretive synthesis of the value of process evaluations, with a practical planning framework. *BMC Med Res Methodol.* 2022;22(1):302.
231. Stame N. What Doesn't Work? Three Failures, Many Answers. *Evaluation.* 2010;16(4):371-87.
232. Nilsen P, Bernhardsson S. Context matters in implementation science: a scoping review of determinant frameworks that describe contextual determinants for implementation outcomes. *BMC Health Serv Res.* 2019;19(1):189.
233. Kumpunen S, Edwards N, Georgiou T, Hughes G. Evaluating integrated care - Why are evaluations not producing the results we expect? [Web page]. 2019 [updated 1/06/2026]. Available from: <https://www.nuffieldtrust.org.uk/sites/default/files/2019-11/the-challenges-of-evaluating-integrated-care-briefing-3.pdf>
234. Durlak JA, DuPre EP. Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation. *Am J Community Psychol.* 2008;41:327-50.
235. Rathod L, Heine M, Boateng D, Martens M, van Olmen J, Ku GM, et al. Process Evaluations for the Scale-Up of Complex Interventions - a Scoping Review. *Int J Integr Care.* 2024;24(4):6.
236. Moore G, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: UK Medical Research Council (MRC) guidance. Medical Research Council (MRC), UK Research and Innovation (UKRI); 2026. Available from: <https://www.ukri.org/wp-content/uploads/2026/07/MRC-210726-ProcessEvaluationComplexInterventionsFullGuidance.pdf>
237. McGill E, Moore G. Process evaluation of complex interventions. In: *Complex Interventions in Health*; 2026. p. 197-207.
238. Wensing M, Geng EH. Making process evaluations more explanatory - key opportunities to advance implementation research. *Implement Sci.* 2026;21(1).
239. Lohmann J. Evaluation of public health interventions. Mechanisms of Change. PowerPoint slides. London: Department of Global Health and Development, London School of Hygiene and Tropical Medicine; 2022.
240. Nilsen P. Making sense of implementation theories, models and frameworks. *Implement Sci.* 2015;10:53.
241. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. *Implement Sci.* 2022;17(1):75.
242. Arizmendi-Barrera KA, Arrieta J, Nevarez Tinoco T, Chahal P, Barker P. Mapping Theories, Models and Frameworks from implementation science for evaluating quality improvement initiatives: a scoping review. *BMJ Qual Saf.* 2026.
243. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health.* 1999;89(9):1322-7.
244. Jung ME, Bourne JE, Gainforth HL. Evaluation of a community-based, family focused healthy weights initiative using the RE-AIM framework. *Int J Behav Nutr Phys Act.* 2018;15(1):13.
245. RE-AIM. Welcome to RE-AIM and PRISM: Implementation in Context [Web page]. [updated 1/06/2026]. Available from: <https://re-aim.org/>
246. Glasgow RE, Trinkley KE, Ford B, Rabin BA. The Application and Evolution of the Practical, Robust Implementation and Sustainability Model (PRISM): History and Innovations. *Glob Implement Res Appl.* 2024;4(4):404-20.



247. RE-AIM. Frequently Asked Questions about PRISM [Web page]. [updated 1/06/2026]. Available from: <https://re-aim.org/learn/prism/>
248. Pfadenhauer LM, Gerhardus A, Mozygemba K, Lysdahl KB, Booth A, Hofmann B, et al. Making sense of complexity in context and implementation: the Context and Implementation of Complex Interventions (CICI) framework. *Implement Sci.* 2017;12(1):21.
249. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *J Med Internet Res.* 2017;19(11):e367.
250. Mielke J, Leppla L, Valenta S, Zullig LL, Zuniga F, Staudacher S, et al. Unraveling implementation context: the Basel Approach for coNtextual ANALysis (BANANA) in implementation science and its application in the SMILe project. *Implement Sci Commun.* 2022;3(1):102.
251. CFIR. What is CFIR? [Web page]. [updated 1/06/2026]. Available from: <https://cfirguide.org/what-is-cfir>
252. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. *Implement Sci.* 2012;7:37.
253. Harvey G, Kitson A. PARIHS revisited: from heuristic to integrated framework for the successful implementation of knowledge into practice. *Implement Sci.* 2016;11:33.
254. Aarons GA, Hurlburt M, Horwitz SM. Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Adm Policy Ment Health.* 2011;38(1):4-23.
255. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm Policy Ment Health.* 2011;38(2):65-76.
256. Wiltsey Stirman S, Baumann AA, Miller CJ. The FRAME: an expanded framework for reporting adaptations and modifications to evidence-based interventions. *Implement Sci.* 2019;14(1):58.
257. Stirman SW, Miller Cj Fau - Toder K, Toder K Fau - Calloway A, Calloway A. Development of a framework and coding system for modifications and adaptations of evidence-based interventions. 2013(1748-5908 (Electronic)).
258. Miller CJ, Barnett ML, Baumann AA, Gutner CA, Wiltsey-Stirman S. The FRAME-IS: a framework for documenting modifications to implementation strategies in healthcare. 2021(1748-5908 (Electronic)).
259. de Souza DE. Elaborating the Context-Mechanism-Outcome configuration (CMOc) in realist evaluation: A critical realist perspective. *Evaluation.* 2013;19(2):141-54.
260. Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. *BMC Med.* 2016;14(1):96.
261. Zullig LL, Drake C, Check DK, Brunkert T, Deschodt M, Olson MS, et al. Embedding implementation science in the research pipeline. *Transl Behav Med.* 2024;14(2):73-9.
262. Salter KL, Kothari A. Using realist evaluation to open the black box of knowledge translation: a state-of-the-art review. *Implementation Science.* 2014;9(1):115.
263. Cornelis J, Van Grootven B, Alvarez Irusta L, de Meester C, Christiaens W, Van Durme T, et al. Telemonitoring bij patiënten met COVID-19. Health Services Research (HSR). Brussel: Federaal Kenniscentrum voor de Gezondheidszorg (KCE); 2022. KCE Reports 354
264. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review - a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy.* 2005;10(1_suppl):21-34.
265. Pawson R, Tilley N. Realistic evaluation. 1997.
266. Westhorp G. Realist impact evaluation: an introduction. 2014



267. Lee M. Introduction to Realist Evaluation. [Web page]. PAN;2017. Available from: <https://paninbc.ca/2017/04/03/introduction-to-realist-evaluation/>
268. Evaluation Task Force. Supplementary Guide: Realist evaluation. 2026 Available from: <https://www.gov.uk/government/publications/the-magenta-book/supplementary-guide-realist-evaluation-html>
269. Graham AC, McAleer S. An overview of realist evaluation for simulation-based education. *Adv Simul (Lond)*. 2018;3:13.
270. Mukumbang FC, Wong G. Mechanism-Based Middle-Range Theories: Using Realist Syntheses to Reconcile Specificity to Context and Generalizability. *Qual Health Res*. 2025.
271. De Weger E, Van Vooren NJE, Wong G, Dalkin S, Marchal B, Drewes HW, et al. What's in a Realist Configuration? Deciding Which Causal Configurations to Use, How, and Why. *International Journal of Qualitative Methods*. 2020;19.
272. Greenhalgh J, Manzano A. Understanding 'context' in realist evaluation and synthesis. *International Journal of Social Research Methodology*. 2022;25(5):583-95.
273. Lacouture A, Breton E, Guichard A, Ridde V. The concept of mechanism from a realist approach: a scoping review to facilitate its operationalization in public health program evaluation. *Implementation Science*. 2015;10(1):153.
274. Dalkin SM, Greenhalgh J, Jones D, Cunningham B, Lhussier M. What's in a mechanism? Development of a key concept in realist evaluation. *Implementation Science*. 2015;10(1):49.
275. Williams L, Rycroft-Malone, J. and Burton, C.R. Bringing critical realism to nursing practice: Roy Bhaskar's contribution. *Nursing Philosophy*. 2017;18(2):e12130.
276. Haynes A, Gilchrist H, Oliveira JS, Tiedemann A. Using Realist Evaluation to Understand Process Outcomes in a COVID-19-Impacted Yoga Intervention Trial: A Worked Example. *Int J Environ Res Public Health*. 2021;18(17).
277. Fletcher A, Jamal F, Moore G, Evans RE, Murphy S, Bonell C. Realist complex intervention science: Applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions. *Evaluation (Lond)*. 2016;22(3):286-303.
278. Gilmore B, McAuliffe E, Power J, Vallières F. Data Analysis and Synthesis Within a Realist Evaluation: Toward More Transparent Methodological Approaches. *International Journal of Qualitative Methods*. 2019;18.
279. Rycroft-Malone J, McCormack B, Hutchinson AM, DeCorby K, Bucknall TK, Kent B, et al. Realist synthesis: illustrating the method for implementation research. *Implementation Science*. 2012;7(1):33.
280. INTRAC. Qualitative comparative analysis (QCA) [Web page]. Intrac;2017. Available from: <https://www.intrac.org/app/uploads/2017/01/Qualitative-comparative-analysis.pdf>
281. Pawson R, Tilley N. Realist evaluation. 2004.
282. Marchal B, van Belle S, van Olmen J, Hoérée T, Kegels G. Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research. *Evaluation*. 2012;18(2):192-212.
283. Pawson R. *The Science of Evaluation: A Realist Manifesto*. London: SAGE Publications Ltd; 2013.
284. Belle SV, Westhorp G, Marchal B, Stevens K, Rogers P, Punton M. Realist evaluation [Web page]. 2024. Available from: <https://www.betterevaluation.org/methods-approaches/realist-evaluation>
285. May C. A rational model for assessing and evaluating complex interventions in health care. *BMC Health Serv Res*. 2006;6:86.
286. May C, Finch T. Implementing, Embedding, and Integrating Practices: An Outline of Normalization Process Theory. *Sociology*. 2009;43(3):535-54.



287. May CR, Mair F, Finch T, MacFarlane A, Dowrick C, Treweek S, et al. Development of a theory of implementation and integration: Normalization Process Theory. *Implementation Science*. 2009;4(1):29.
288. May C. Towards a general theory of implementation. *Implement Sci*. 2013;8:18.
289. May CR, Cummings A, Girling M, Bracher M, Mair FS, May CM, et al. Using Normalization Process Theory in feasibility studies and process evaluations of complex healthcare interventions: a systematic review. *Implement Sci*. 2018;13(1):80.
290. Dalkin SM, Hardwick RJL, Haighton CA, Finch TL. Combining Realist approaches and Normalization Process Theory to understand implementation: a systematic review. *Implement Sci Commun*. 2021;2(1):68.
291. May CR, Albers B, Bracher M, Finch TL, Gilbert A, Girling M, et al. Translational framework for implementation evaluation and research: a normalisation process theory coding manual for qualitative research and instrument development. *Implement Sci*. 2022;17(1):19.
292. Jagosh J, Stott H, Halls S, Thomas R, Liddiard C, Cupples M, et al. Benefits of realist evaluation for rapidly changing health service delivery. *BMJ Open*. 2022;12(7):e060347.
293. Emmel N. *Sampling and Choosing Cases in Qualitative Research: A Realist Approach*. London: SAGE Publications Ltd; 2013.
294. Moore G. Process evaluation of complex interventions: a summary of Medical Research Council guidance. [Preprint]. 2021]Available from: <https://www.ukri.org/wp-content/uploads/2021/11/MRC-291121-PHSRN-ProcessEvaluationSummaryGuidance.pdf>
295. Manzano A. The craft of interviewing in realist evaluation. *Evaluation*. 2016;22(3):342-60.
296. Pawson R. Theorizing the interview. *British Journal of Sociology*. 1996:295-314.
297. Mukumbang FC, Marchal B, Van Belle S, van Wyk B. Using the realist interview approach to maintain theoretical awareness in realist studies. *Qualitative Research*. 2020;20(4):485-515.
298. Bos C, de Weger E, Wildeman I, Pannebakker N, Kemper PF. Implement social prescribing successfully towards embedding: what works, for whom and in which context? A rapid realist review. *BMC Public Health*. 2024;24(1):1836.
299. Neyt M, Thiry N, Cleemput I. Belgian guidelines for economic evaluations and budget impact analyses: third edition. Methods. Brussels: Belgian Health Care Knowledge Centre (KCE); 2025 08/05/2025. KCE Reports 400 Available from: <https://doi.org/10.57598/R400C>
300. Baghbanian A, Carter D, Merlin T. Methods for the health technology assessment of complex interventions: A scoping review. *PLoS One*. 2025;20(3):e0315381.
301. Sullivan SD, Mauskopf JA, Augustovski F, Jaime Caro J, Lee KM, Minchin M, et al. Budget impact analysis-principles of good practice: report of the ISPOR 2012 Budget Impact Analysis Good Practice II Task Force. *Value Health*. 2014;17(1):5-14.
302. Camberlin C, Obyn C, Neyt M. Transversal Budget Impact Analysis. Method. Brussels: Belgian Health Care Knowledge Centre (KCE); 2023 Method. KCE Reports 369 (369C) Available from: <https://doi.org/10.57598/R369C>
303. Turner HC, Rivillas-Garcia JC, Prinja S, Hung TM, Dabak SV, Asare BA, et al. An Introduction to Costing and the Types of Costs Used within Health Economic Studies. *Pharmacoecon Open*. 2025;9(6):849-68.
304. Spacirova Z, Epstein D, Garcia-Mochon L, Rovira J, Olry de Labry Lima A, Espin J. A general framework for classifying costing methods for economic evaluation of health care. *Eur J Health Econ*. 2020;21(4):529-42.
305. Tan SS, Rutten FF, van Ineveld BM, Redekop WK, Hakkaart-van Roijen L. Comparing methodologies for the cost estimation of hospital services. *Eur J Health Econ*. 2009;10(1):39-45.
306. Van de Voorde C, Lefèvre M, Devriese S, Obyn C, Swartenbroekx N. Indirect hospital costs as a mark-up percentage on direct costs in Belgium. *Health Services Research (HSR)*.



- Brussels: Belgian Health Care Knowledge Centre (KCE); 2024 04/2024. KCE Reports 380 (380C) Available from: <https://doi.org/10.57598/R380C>
307. Guerre P, Hayes N, Bertaux AC. [Hospital costs estimation by micro and gross-costing approaches]. *Rev Epidemiol Sante Publique*. 2018;66 Suppl 2:S65-S72.
 308. Shakya S, Bohingamu Mudiyansele S, Robinson S, Randall S, Gao L. Time-Driven Activity-Based Costing and Its Use in Health Economic Analysis: A Systematic Literature Review. *Appl Health Econ Health Policy*. 2025;23(6):989-1004.
 309. Wensing M, Sales A, Wilson P, Armstrong R, Kislov R, Rankin NM, et al. Implementation Science and Implementation Science Communications: a refreshed description of the journals' scope and expectations. *Implement Sci*. 2021;16(1):103.
 310. Wensing M, Grol R. Implementatie: effectieve verbetering van de patiëntenzorg (8ste, herziene druk). Amsterdam: Bohn Stafleu Van Loghum; 2023.
 311. Grimshaw JM, Patey AM, Kirkham KR, Hall A, Dowling SK, Rodondi N, et al. De-implementing wisely: developing the evidence base to reduce low-value care. *BMJ Qual Saf*. 2020;29(5):409-17.
 312. Wensing M. Challenges facing implementation science researchers. In: Richards DP, Hallberg IR, editors. *Complex interventions in health: an overview of research methods* (second edition). London: Routledge (Taylor & Francis Group); 2026.
 313. Eccles MP, Mittman BS. Welcome to implementation science. *Implementation Science*. 2006;1(1):1-3.
 314. Grol R. Personal paper. Beliefs and evidence in changing clinical practice. *BMJ*. 1997;315(7105):418-21.
 315. Wensing M, Grol R. Knowledge translation in health: how implementation science could contribute more. *BMC Med*. 2019;17(1):88.
 316. Boaz A, Baeza J, Fraser A, Persson E. 'It depends': what 86 systematic reviews tell us about what strategies to use to support the use of research in clinical practice. *Implement Sci*. 2024;19(1):15.
 317. Wensing M, Grol R. Methoden om invoeringsproblemen op te sporen. In: Wensing M, Grol R, editors. *Implementatie: effectieve verbetering van patiëntenzorg* (8ste herziene druk). Amsterdam: Bohn Stafleu Van Loghum; 2023.
 318. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. *Lancet*. 2003;362(9391):1225-30.
 319. Holtrop JS, Dorsey-Holliman B, Hamilton AB. Navigating qualitative methods choices in dissemination and implementation research. *Implement Sci*. 2025;21(1):11.
 320. Reardon CM, Damschroder LJ, Ashcraft LE, Kerins C, Bachrach RL, Nevedal AL, et al. The Consolidated Framework for Implementation Research (CFIR) User Guide: a five-step guide for conducting implementation research using the framework. *Implement Sci*. 2025;20(1):39.
 321. Jager C, Steinhäuser J, Freund T, Baker R, Agarwal S, Godycki-Cwirko M, et al. Process evaluation of five tailored programs to improve the implementation of evidence-based recommendations for chronic conditions in primary care. *Implement Sci*. 2016;11(1):123.
 322. Lewis CC, Powell BJ, Brewer SK, Nguyen AM, Schriger SH, Vejnosa SF, et al. Advancing mechanisms of implementation to accelerate sustainable evidence-based practice integration: protocol for generating a research agenda. *BMJ Open*. 2021;11(10):e053474.
 323. Kok G, Gottlieb NH, Peters GJ, Mullen PD, Parcel GS, Rutter RA, et al. A taxonomy of behaviour change methods: an Intervention Mapping approach. *Health Psychol Rev*. 2016;10(3):297-312.
 324. Grol R, Wensing M. Ontwikkeling en selectie van implementatiestrategieën. In: Wensing M, Grol R, editors. *Implementatie: effectieve verbetering van de patiëntenzorg* (8ste herziene druk). Amsterdam: Bohn Stafleu Van Loghum; 2023.



325. Vis C, van Bodegom-Vos L, Hipple-Walters B, Powell BJ, Ista E, van Nassau F. Applied methods for matching implementation strategies to determinants: a scoping review of scientific and grey literature, and qualitative exploration of practice experiences. *Implement Sci.* 2025;21(1):14.
326. Ulrich W, Reynolds M. Critical Systems Heuristics: The Idea and Practice of Boundary Critique. In: Reynolds M, Holwell S, editors. *Systems Approaches to Making Change: A Practical Guide*. London: Springer London; 2020. p. 255-306. Available from: https://doi.org/10.1007/978-1-4471-7472-1_6
327. Effective Practice and Organisation of Care (EPOC). EPOC Taxonom [Web page].2015. Available from: <https://epoc.cochrane.org/epoc-taxonomy>
328. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. *Implement Sci.* 2015;10:21.
329. Waltz TJ, Powell BJ, Fernandez ME, Abadie B, Damschroder LJ. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. *Implement Sci.* 2019;14(1):42.
330. Meng M, Wang Z, Yang D, Jiang H, Lu J, Chen S, et al. Application status of the CFIR-ERIC matching tool in healthcare context: a scoping review. *Implement Sci.* 2026;21(1).
331. McHugh S, Riordan F, O'Mahony A, McCarthy LJ, Contreras Navarro A, Kerins C, et al. Tailored interventions to address determinants of practice. *Cochrane Database Syst Rev.* 2026;2:CD005470.
332. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med Care.* 2012;50(3):217-26.
333. Williams NJW, Smith JD, Landes SJ, Curran GM. Effectiveness-implementation hybrid studies to speed the translation of complex health interventions into practice. In: Richards DP, Rahm Hallberg I, editors. *Complex interventions in Health: an overview of research methods (second edition)*; 2026.
334. Becker M, Breuing J, Nothacker M, Deckert S, Brombach M, Schmitt J, et al. Guideline-based quality indicators-a systematic comparison of German and international clinical practice guidelines. *Implement Sci.* 2019;14(1):71.
335. Braspenning J, Hermens R, Calsbeek H, Campbell S, Wees P, Grol R. Indicators for Quality and Safety of Care. In: Wensing M, Grol R, Grimshaw J, editors. *Improving Patient Care: The Implementation of Change in Health Care*. Chichester, UK: John Wiley & Sons, Ltd; 2020. p. 131-54.
336. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *Jama.* 2006;296(9):1094-102.
337. Leroy R, De Gendt C, Stordeur S, Silversmit G, Verleye L, Schillemans V, et al. Quality indicators for the management of head and neck squamous cell carcinoma. *Health Services Research (HSR)*. Brussels: Belgian Health Care Knowledge Centre (KCE); 2019 01/2019. KCE Reports 305 (305) Available from: https://www.kce.fgov.be/sites/default/files/2021-11/KCE_305_Quality_indicators_Head_and_neck_Report_5.pdf
338. Quentin W, Partanen V-M, Bronwood I, Klazinga N. Measuring healthcare quality. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. *Improving healthcare quality in Europe. Characteristics, effectiveness and implementation of different strategies: World Health Organization (acting as the host organization for, and secretariat of, the European Observatory on Health Systems and Policies) and OECD*; 2019. p. 31-62. Available from: <https://apps.who.int/iris/bitstream/handle/10665/327356/9789289051750-eng.pdf?sequence=1&isAllowed=y>

339. Busse R, Panteli D, Quentin W. An introduction to healthcare quality: defining and explaining its role in health systems. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving healthcare quality in Europe. Characteristics, effectiveness and implementation of different strategies: World Health Organization (acting as the host organization for, and secretariat of, the European Observatory on Health Systems and Policies) and OECD 2019; 2019. p. 3-18. Available from: <https://apps.who.int/iris/bitstream/handle/10665/327356/9789289051750-eng.pdf?sequence=1&isAllowed=y>
340. Institute of Medicine. In: Crossing the Quality Chasm: A New Health System for the 21st Century. Washington (DC): National Academies Press (US); 2001.
341. Vlayen J, Verstreken M, Mertens C, Van Eycken E, Penninckx F. Quality insurance for rectal cancer - phase 2: development and testing of a set of quality indicators. Good Clinical Practice (GCP). Brussels: Belgian Health Care Knowledge Centre (KCE); 2008 03/07/2008. KCE Reports 81 Available from: <https://kce.fgov.be/publication/report/quality-insurance-for-rectal-cancer-phase-2-development-and-testing-of-a-set-of-q>
342. Vlayen J, Vrijens F, Beirens K, Stordeur S, Devriese S, Van Eycken E. Quality indicators in oncology: testis cancer. Good Clinical Practice (GCP). Brussels: Belgian Health Care Knowledge Centre (KCE); 2010. KCE Reports 149 Available from: <https://kce.fgov.be/publication/report/quality-indicators-in-oncology-testis-cancer>
343. Stordeur S, Vrijens F, Beirens K, Vlayen J, Devriese S, Van Eycken E. Quality indicators in oncology: breast cancer. Good Clinical Practice (GCP). Brussels: Belgian Health Care Knowledge Centre (KCE); 2010. KCE Reports 150C (D/2010/10.273/101) Available from: <https://doi.org/10.57598/R150C>
344. Vlayen J, De Gendt C, Stordeur S, Schillemans V, Camberlin C, Vrijens F, et al. Quality indicators for the management of upper gastrointestinal cancer. Good Clinical Practice (GCP). Brussels: Belgian Health Care Knowledge Centre (KCE); 2013. KCE Reports 200 (D/2013/10.273/15) Available from: <https://doi.org/10.57598/R200C>
345. Vrijens F, Verleye L, De Gendt C, Schillemans V, Robays J, Camberlin C, et al. Quality indicators for the management of lung cancer. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2016 22042016. KCE Reports 266 Available from: <https://doi.org/10.57598/R266C>
346. Leroy R, De Gendt C, Bourgeois J, Verbeeck J, Savoye I, Silversmit G, et al. Survival and Quality of care offered in Belgian hospitals with and without recognition for breast cancer. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2023 03/2023. KCE Reports 365 (365C) Available from: <https://doi.org/10.57598/R365C>
347. Kötter T, Blozik E, Scherer M. Methods for the guideline-based development of quality indicators--a systematic review. Implement Sci. 2012;7:21.
348. Lawrence M, Frede O. Indicators of quality in health care. Eur J Gen Pract. 1997;3:103-8.
349. Braithwaite J, Hibbert P, Blakely B, Plumb J, Hannaford N, Long JC, et al. Health system frameworks and performance indicators in eight countries: A comparative international analysis. SAGE Open Med. 2017;5.
350. Parmelli E, Langendam M, Piggott T, Adolfsson J, Akl EA, Armstrong D, et al. Guideline-based quality assurance: a conceptual framework for the definition of key elements. BMC Health Serv Res. 2021;21(1):173.
351. Panteli D, Quentin W, Busse R. Understanding healthcare quality strategies: a five-lens framework. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving healthcare quality in Europe. Characteristics, effectiveness and implementation of different strategies: World Health Organization (acting as the host organization for, and secretariat of, the European Observatory on Health Systems and Policies) and OECD; 2019. p. 19-30. Available from: <https://apps.who.int/iris/bitstream/handle/10665/327356/9789289051750-eng.pdf?sequence=1&isAllowed=y>
352. Endometriosis UK. Analysis of Endometriosis Service Provision in Scotland. Endometriosis UK report for the Scottish Government. Edinburgh 2022. Available from:



- https://www.endometriosis-uk.org/sites/default/files/files/Analysis%20of%20Endometriosis%20Service%20Provision%20in%20Scotland%20-%20Final_0.pdf
353. Gameiro S, Sousa-Leite M, Vermeulen N. Dissemination, implementation and impact of the ESHRE evidence-based guidelines. *Hum Reprod Open*. 2019;2019(3):hoz011.
 354. Bouckaert N, Savoye I, Zeevaert R, Devos C, San Miguel L, Van den Heede K. Shifting care from inpatient towards day-care and ambulatory care: impact on readmission rates & post-discharge mortality. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2026 08/01/2026. KCE Reports 412 Available from: <https://doi.org/10.57598/R412C>
 355. Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q*. 1966;44(3):Suppl:166-206.
 356. Palmer RH, Reilly MC. Individual and institutional variables which may serve as indicators of quality of medical care. *Med Care*. 1979;17(7):693-717.
 357. Becker CM, Bokor A, Heikinheimo O, Horne A, Jansen F, Kiesel L, et al. ESHRE guideline: endometriosis. *Hum Reprod Open*. 2022;2022(2):hoac009.
 358. Weissman NW, Allison JJ, Kiefe CI, Farmer RM, Weaver MT, Williams OD, et al. Achievable benchmarks of care: the ABCs of benchmarking. *J Eval Clin Pract*. 1999;5(3):269-81.
 359. Mainz J. Defining and classifying clinical indicators for quality improvement. *Int J Qual Health Care*. 2003;15(6):523-30.
 360. Donabedian A. The quality of care. How can it be assessed? *JAMA*. 1988;260(12):1743-8.
 361. Rubin HR, Pronovost P, Diette GB. From a process of care to a measure: the development and testing of a quality indicator. *Int J Qual Health Care*. 2001;13(6):489-96.
 362. Terris DD, Aron DC. Attribution and causality in health-care performance measurement. In: Smith PC, Mossialos E, Papanicolas I, Leatherman S, editors. *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge: Cambridge University Press; 2010. p. 311-38.
 363. Iezzoni LI. Risk adjustment for measuring health care outcomes. 3rd ed. Chicago, Illinois: Health Administration Press; 2003.
 364. Wollersheim H, Hermens R, Hulscher M, Braspenning J, Ouwens M, Schouten J, et al. Clinical indicators: development and applications. *Neth J Med*. 2007;65(1):15-22.
 365. Cebam. CEQUID: Cebam Quality Indicator Development. Methodologisch document. Leuven: 2026. Available from: <https://cdn.nimbu.io/s/q6nkby/assets/1788248413616/cequid-cebam-quality-indicator-development-methodologisch-document-2026.pdf>
 366. Leroy R, De Gendt C, Bourgeois J, Verbeeck J, Savoye I, Van Damme N, et al. The Challenge of Calculating EUSOMA Quality Indicators with Administrative Data: A Pilot Study in Belgium. *Breast Care (Basel)*. 2025;20(2):88-94.
 367. Jespers V, Verleye L, Costa E, Vlayen J, Struyf T. KCE process note: Research Methods for guidelines – update 2025. Methods. Brussels: Belgium Health Care Knowledge Centre (KCE); 2025 6/01/2026. KCE Reports 411 Available from: <https://doi.org/10.57598/R411C>
 368. Fitch K, Bernstein SJ, Aguilar MD, Burnand B, J.R. L, Lazaro P, et al. The RAND/UCLA Appropriateness Method. User's Manual. Santa Monica, CA: 2001. Available from: https://www.rand.org/content/dam/rand/pubs/monograph_reports/2011/MR1269.pdf
 369. Campbell SM, Cantrill JA, Roberts D. Prescribing indicators for UK general practice: Delphi consultation study. *Bmj*. 2000;321(7258):425-8.
 370. Hermens RP, Ouwens MM, Vonk-Okhuijsen SY, van der Wel Y, Tjan-Heijnen VC, van den Broek LD, et al. Development of quality indicators for diagnosis and treatment of patients with non-small cell lung cancer: a first step toward implementing a multidisciplinary, evidence-based guideline. *Lung Cancer*. 2006;54(1):117-24.



371. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-updagger. *Ann Oncol*. 2019;30(8):1194-220.
372. Rubio IT, Marotti L, Biganzoli L, Aristei C, Athanasiou A, Campbell C, et al. EUSOMA quality indicators for non-metastatic breast cancer: An update. *Eur J Cancer*. 2024;198:113500.
373. Gurol-Urganci I, Campbell F, Black N. Quality assessment. In: Gurol-Urganci I, Campbell F, Black N, editors. Second edition ed. London (England): Open University Press, McGraw-Hill Education; 2017. p. 208-25.
374. Tolonen H, Reinikainen J, Koponen P, Elonheimo H, Palmieri L, Tijhuis MJ, et al. Cross-national comparisons of health indicators require standardized definitions and common data sources. *Arch Public Health*. 2021;79(1):208.
375. Vermeulen N, Abrao MS, Einarsson JI, Horne AW, Johnson NP, Lee TTM, et al. Endometriosis classification, staging and reporting systems: a review on the road to a universally accepted endometriosis classification. *Facts Views Vis Obgyn*. 2021;13(4):305-30.
376. Leroy R, Camberlin C, Lefèvre M, Van den Heede K, Van de Voorde C, Beguin C. Variability in elective day-surgery rates between Belgian hospitals - Analyses of administrative data explained by surgical experts. *Int J Surg*. 2017;45:118-24.
377. Leroy R, Desimpel F, Ombelet S, De Jaeger M, Benahmed N, Camberlin C, et al. How can we improve endometriosis care in Belgium? Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2024 04/2024. KCE Reports 378 (378C) Available from: <https://doi.org/10.57598/R378C>
378. Kannan V, Fish JS, Mutz JM, Carrington AR, Lai K, Davis LS, et al. Rapid Development of Specialty Population Registries and Quality Measures from Electronic Health Record Data*. An Agile Framework. *Methods Inf Med*. 2017;56(99):e74-e83.
379. Van Brabandt H, Camberlin C, Neyt M, De Laet C, Stroobandt S, Devriese S, et al. Cardiac resynchronisation therapy. A Health technology Assessment. Health Technology Assessment (HTA). Brussels: Belgian Health Care Knowledge Centre (KCE); 2010 2011-02-15 (2nd print; 1st print: 2011-01-28). KCE Reports 145C (D/2010/10.273/84) Available from: <https://doi.org/10.57598/R145C>
380. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J-L, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. *BMC Public Health*. 2011;11:450.
381. Paridaens R, Van den Bulck S, De Jonghe M, Fauquert B, Meel L, Raat W, et al. Development of Quality Indicators for the Correct Use of Electronic Medical Records in Primary Care: Modified Delphi Study. *JMIR Med Inform*. 2026;14:e80057.
382. Piggott T, Langendam MW, Parmelli E, Adolfsson J, Akl EA, Armstrong D, et al. The GIN-McMaster guideline tool extension for the integration of quality improvement and quality assurance in guidelines: a description of the methods for its development. *J Clin Epidemiol*. 2023;154:197-203.
383. Kelley E, Hurst J. Health Care Quality Indicators Project: Conceptual framework paper. Paris: 2006. OECD Health Working Papers 23 (JT03205355) Available from: https://www.oecd.org/content/dam/oecd/en/publications/reports/2006/03/health-care-quality-indicators-project_g17a17b5/440134737301.pdf
384. Reiter A, Fischer B, Kötting J, Geraedts M, Jäckel WH, Döbler K. [QUALIFY--a tool for assessing quality indicators]. *Z Arztl Fortbild Qualitatssich*. 2007;101(10):683-8.
385. de Koning JS, Kallewaard M, Klazinga NS. Prestatie-indicatoren langs de meetlat – het AIRE instrument. *TVGW*. 2007;85:261–4.
386. Center for Health Policy/Center for Primary Care and Outcomes Research & Battelle Memorial Institute. Quality Indicator Measure Development, Implementation, Maintenance, and Retirement (Prepared by Battelle, under Contract No. 290-04-0020). Rockville, MD: Agency



- for Healthcare Research and Quality; 2011. Available from: https://qualityindicators.ahrq.gov/downloads/resources/publications/2011/qi_measure_development_implementation_maintenance_retirement_full_5-3-11.pdf
387. Nothacker M, Stokes T, Shaw B, Lindsay P, Sipilä R, Follmann M, et al. Reporting standards for guideline-based performance measures. *Implement Sci.* 2016;11:6.
 388. International Society for Quality in Health Care External Evaluation Association (ISQua EEA). Guidelines and Principles for the Development of Health and Social Care Standards Genève, Switzerland: 2022. Available from: <https://ieea.ch/wp-content/uploads/2024/09/Guidelines-and-Principles-for-the-Development-of-Health-and-Social-Care-Standards-5th-Edition-v1.1.pdf>
 389. Copnell B, Hagger V, Wilson SG, Evans SM, Sprivilis PC, Cameron PA. Measuring the quality of hospital care: an inventory of indicators. *Intern Med J.* 2009;39(6):352-60.
 390. Schang L, Blotenberg I, Boywitt D. What makes a good quality indicator set? A systematic review of criteria. *Int J Qual Health Care.* 2021;33(3).
 391. Shwartz M, Restuccia JD, Rosen AK. Composite Measures of Health Care Provider Performance: A Description of Approaches. *Milbank Q.* 2015;93(4):788-825.
 392. Reeves D, Doran T, Valderas JM, Kontopantelis E, Trueman P, Sutton M, et al. How to identify when a performance indicator has run its course. *Bmj.* 2010;340:c1717.
 393. Piggott T, Langendam M, Parmelli E, Adolfsson J, Akl EA, Armstrong D, et al. Bringing two worlds closer together: a critical analysis of an integrated approach to guideline development and quality assurance schemes. *BMC Health Serv Res.* 2021;21(1):172.
 394. Piggott T, Langendam MW, Parmelli E, Adolfsson J, Akl EA, Armstrong D, et al. Integrating Quality Assurance and Quality Improvement With Guidelines: Systematic Stakeholder-Driven Development of an Extension of the Guidelines International Network-McMaster Guideline Development Checklist. *Ann Intern Med.* 2022;175(5):735-9.
 395. Meadows DH. Thinking in systems. A primer. Wright D, editor. Vermont: Chelsea Green Publishing; 2008.
 396. Patton MQ. Qualitative research & evaluation methods: Integrating theory and practice. Sage publications; 2014.
 397. Walton M. Setting the context for using complexity theory in evaluation: boundaries, governance and utilisation. *Evidence & Policy.* 2016;12(1):73-89.
 398. Walton M. Applying complexity theory: a review to inform evaluation design. *Eval Program Plann.* 2014;45:119-26.
 399. Peeters J, Abts F, Albert N, Bloemen H, Coates A, Beukelaer E, et al. Veerkracht en burgerschap. Sociaal werk in transitie [Resilience and citizenship. Social work in transition]. 2015.
 400. Nicolis G, Prigogine I. Exploring Complexity: an introduction. New York: W.H. Freeman and company; 1989.
 401. Johnson S. Emergence: The Connected Lives of Ants, Brains, Cities, and Software. New York: Scribner; 2002.
 402. Senge P, Kleiner A. The Fifth Discipline Fieldbook: Strategies and Tools for Building a Learning Organization. Nicholas Brealey Publishing; 2011.
 403. Checkland P, Poulter J. Learning For Action: A Short Definitive Account of Soft Systems Methodology, and its use for Practitioners, Teachers and Students. Wiley; 2007.
 404. Goodman M. Systems Thinking: What, Why, When, Where, and How? The Systems Thinker. 1997;8(2):2.
 405. Reynolds M, Holwell S. Introducing Systems Approaches. In: Reynolds M, Holwell S, editors. Systems Approaches to Making Change: A Practical Guide. London: Springer London; 2020. p. 1-24. Available from: https://doi.org/10.1007/978-1-4471-7472-1_1



406. Horst W, Rittel J, Webber MM. Dilemmas in a general theory of planning. *Policy Science*. 1973;4(2):155-69.
407. Carmichael T, Hadžikadić M. The fundamentals of complex adaptive systems. In: *Complex adaptive systems: Views from the physical, natural, and social sciences*: Springer; 2019. p. 1-16.
408. Renger R. *System Evaluation Theory: A Blueprint for Practitioners Evaluating Complex Interventions Operating and Functioning as Systems*. Emerald Publishing Limited; 2022.
409. Brant S. If Russ Ackoff had given a TED talk (video): YouTube; 1994.
410. Reason P, Bradbury H, Ison R. *Systems Thinking and Practice for Action Research*. 2nd ed.
411. Morecroft J. System Dynamics. In: Reynolds M, Holwell S, editors. *Systems Approaches to Making Change: A Practical Guide*. London: Springer London; 2020. p. 25-88. Available from: https://doi.org/10.1007/978-1-4471-7472-1_2
412. Gates EF. Making sense of the emerging conversation in evaluation about systems thinking and complexity science. *Evaluation and Program Planning*. 2016;59:62-73.
413. Arnold RD, Wade JP. A definition of systems thinking: A systems approach. *Procedia computer science*. 2015;44:669-78.
414. Williams B, Imam I. *Systems concepts in evaluation: An expert anthology*. EdgePress of Inverness Point Reyes, CA; 2007.
415. Westhorp G. Using complexity-consistent theory for evaluating complex systems. *Evaluation*. 2012;18(4):405-20.
416. McGill E, Marks D, Er V, Penney T, Petticrew M, Egan M. Qualitative process evaluation from a complex systems perspective: a systematic review and framework for public health evaluators. *PLoS medicine*. 2020;17(11):e1003368.
417. Bustamante M, Vidueira P, Baker L. Systems thinking and complexity science–informed evaluation frameworks: assessment of the economics of ecosystems and biodiversity for agriculture and food. *New directions for evaluation*. 2021;2021(170):81-100.
418. Association SiETIGotAE. Principles for Effective Use of Systems Thinking in Evaluation Practice. 2018. Available from: <https://www.aes.asn.au/images/stories/regions/QLD/SETIG-Principles-FINAL-DRAFT-2018-9-9.pdf>
419. Jackson MC. Critical systems thinking and practice. *European Journal of Operational Research*. 2001;128(2):233-44.
420. Rogers PJ. Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation*. 2008;14(1):29-48.
421. Wholey JS. Using policy evaluation to support decision making and improve program performance. *Japanese Journal of Evaluation Studies*. 2003;3(2):6-21.
422. Chen HT. Interfacing theories of program with theories of evaluation for advancing evaluation practice: Reductionism, systems thinking, and pragmatic synthesis. *Evaluation and program planning*. 2016;59:109-18.
423. Patton MQ. *Utilization-focused evaluation*. Sage publications; 2008.
424. Pinzon-Salcedo LA. Critical systems thinking: an evolving systemic approach. *Journal of Systems Thinking*. 2023;3(1):1-14.
425. Vandenbroeck P. Systems thinking and four forms of complexity. ShiftN; 2015. Available from: https://shiftn.com/uploads/pdf/shiftN_WP_Systems-Thinking_Web-kopie.pdf
426. Martinez-Moyano IJ. System Dynamics in Action. *Journal of Systems Thinking*. 2023;3(1):1-16.
427. Kim DH. *Systems Archetypes I*. Pegasus Communications; 1992.



428. Cassidy R, Singh NS, Schiratti PR, Semwanga A, Binyaruka P, Sachingongu N, et al. Mathematical modelling for health systems research: a systematic review of system dynamics and agent-based models. BMC Health Serv Res. 2019;19(1):845.
429. Hoverstadt P. The Viable System Model. In: Reynolds M, Holwell S, editors. Systems Approaches to Making Change: A Practical Guide. London: Springer London; 2020. p. 89-138. Available from: https://doi.org/10.1007/978-1-4471-7472-1_3
430. Vandenbroeck P, Christiaens W, Dauvrin M, Wickert R, Becher K, Hendrickx E, et al. An evaluation protocol for NIHDI conventions. Health Services Research (HSR). Brussels: Health Care Knowledge Centre (KCE); 2018. D/2018/10.273/22
431. Simonsen J. Soft Systems Methodology: an introduction. Roskilde: Roskilde University; 1994. Available from: <https://www.jespersimonsen.dk/Downloads/SSM-IntroductionJS.pdf>
432. Boulanger V, MacLaurin A, Quach C. Barriers and facilitators for using administrative data for surveillance purpose: a narrative overview. J Hosp Infect. 2025;155:25-36.
433. Bouckaert N, Lefèvre M, Van den Heede K, Detollenaere J, Castanares-Zapatero D, Van de Voorde C. Organisation and financing of intensive care services in Belgium. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Center (KCE); 2025 04/07/2025. KCE Reports 404 Available from: <https://doi.org/10.57598/R404C>
434. Belgian Cancer Registry. BCR's Role in Cancer Control. Celebrating 20 years. 2005-2025. 2026. Available from: https://kankerregister.org/sites/default/files/2026/BCRs_Role_in_Cancer_Control_Celebrating_20_years_Belgium_2005_2025_3.pdf
435. Zayed AM, Janssens A, Caspers M, Mamouris P, Beerten SG, De Burghgraeve T, et al. Data resource profile: the Intego-II primary care database. Int J Epidemiol. 2025;54(6).
436. Eurostat. Income and living conditions. Modules [Web page]. Available from: <https://ec.europa.eu/eurostat/web/income-and-living-conditions/database/modules>
437. Statbel. Risico op armoede of sociale uitsluiting [Web page]. Available from: <https://statbel.fgov.be/nl/themas/huishoudens/armoede-en-levensomstandigheden/risico-op-armoede-sociale-uitsluiting#documents>
438. Bouckaert N, Maertens de Noordhout C, Van de Voorde C. Health System Performance Assessment: how equitable is the Belgian health system? Health Services Research (HSR). Brussel: Belgian Health Care Knowledge Centre (KCE); 2020 12/2020. KCE Reports 334 Available from: <https://doi.org/10.57598/R334C>
439. Statbel. Household budget survey (HBS) [Web page]. Available from: <https://statbel.fgov.be/en/themes/households/household-budget-survey-hbs#documents>
440. Statbel. REDEGEO: het moderniseren van de statistische sectoren van België in 2025 [Web page]. Available from: <https://statbel.fgov.be/nl/nieuws/redegeo-het-moderniseren-van-de-statistische-sectoren-van-belgie-2025>
441. Statbel. 58% of the Belgian population lives in a 5-kilometre radius of a hospital [Web page]. Available from: <https://statbel.fgov.be/en/news/58-belgian-population-lives-5-kilometre-radius-hospital>
442. RIZIV-INAMI. Artsen: transparant info bezorgen aan de ziekenfondsen [Web page]. Available from: <https://www.inami.fgov.be/nl/professionals/individuele-zorgverleners/artsen/uitoefening-van-het-beroep/artsen-elektronische-facturatie-en-transparant-info-bezorgen-aan-de-ziekenfondsen>
443. RIZIV-INAMI. Het Bijzonder solidariteitsfonds (BSF): uitzonderlijke vergoeding van medische verstrekkingen [Web page]. Available from: <https://www.riziv.fgov.be/nl/thema-s/verzorging-kosten-en-terugbetaling/financiele-toegankelijkheid/bijzonder-solidariteitsfonds>
444. Frederick J Gravetter, Lori-Ann B. Forzano, Rakow T. Research Methods For The Behavioural Sciences, 1st Edition. 2021.



445. Moore BJ, White S, Washington R, Coenen N, Elixhauser A. Identifying Increased Risk of Readmission and In-hospital Mortality Using Hospital Administrative Data: The AHRQ Elixhauser Comorbidity Index. *Med Care*. 2017;55(7):698-705.
446. AHRQ. Elixhauser Comorbidity Software Refined for ICD-10-CM [Web page]. Available from: https://hcup-us.ahrq.gov/toolssoftware/comorbidityicd10/comorbidity_icd10.jsp
447. Ouyang Y, Li F, Preisser JS, Taljaard M. Sample size calculators for planning stepped-wedge cluster randomized trials: a review and comparison. *Int J Epidemiol*. 2022;51(6):2000-13.
448. Ouyang Y, Hemming K, Li F, Taljaard M. Estimating intra-cluster correlation coefficients for planning longitudinal cluster randomized trials: a tutorial. *Int J Epidemiol*. 2023;52(5):1634-47.
449. Brown KS. Analysis of Covariance. In: Armitage P, Colton T, editors. *Encyclopedia of Biostatistics*. Chichester, UK: John Wiley & Sons, Ltd; 2005.
450. Ellenberg SS, Fleming TR, DeMets DL. *Data Monitoring Committees in Clinical Trials. A Practical Perspective*. The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons Ltd; 2002.
451. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. New York: J. Wiley & Sons; 1984.
452. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Stat Med*. 2009;28(8):1181-217.
453. University of Colorado Denver. Proctor's Implementation Outcomes [Web page]. [updated 1/06/2026]. Available from: <https://dissemination-implementation.org/models/proctors-implementation-outcomes/>
454. Centre for Implementation Science, King's College London. Implementation Outcome Repository [Web page]. [updated 1/06/2026]. Available from: <https://implementationoutcomerepository.org/implementation-outcomes>