Quality Insurance of rectal cancer – phase 3: statistical methods to benchmark centers on a set of quality indicators

KCE reports 161C

Belgian Health Care Knowledge Centre
Federaal Kenniscentrum voor de Gezondheidszorg
Centre fédéral d'expertise des soins de santé
2011

The Belgian Health Care Knowledge Centre

Introduction: The Belgian Health Care Knowledge Centre (KCE) is an organization of public interest, created on the 24th of December 2002 under the supervision of the Minister of Public Health and Social Affairs.
KCE is in charge of conducting studies that support the political decision making on health care and health insurance.

Executive Board

Actual Members: Pierre Gillet (President), Dirk Cuypers (Vice-president), Jo De Cock (Vice-president), Frank Van Massenhove (Vice-president), Maggie De Block, Jean-Pierre Baeyens, Ri de Ridder, Olivier De Stexhe, Johan Pauwels, Daniel Devos, Jean-Noël Godin, Xavier De Cuyper, Palstermans Paul, Xavier Brenez, Rita Thys, Marc Moens, Marco Schetgen, Patrick Verertbruggen, Michel Foulon, Myriam Hubinon, Michael Callens, Bernard Lange, Jean-Claude Praet.

Substitute Members: Rita Cuypers, Christiaan De Coster, Benoît Collin, Lambert Stamatakis, Karel Vermeyen, Katrien Kesteloot, Bart Ooghe, Frederic Lernoux, Anne Vanderstappen, Greet Musch, Geert Messiaen, Anne Remacle, Roland Lemeye, Annick Poncé, Pierre Smiets, Jan Bertels, Celien Van Moerkerke, Yolande Husden, Ludo Meyers, Olivier Thonon, François Perl.

Government commissioner: Yves Roger

Management

Chief Executive Officer: Raf Mertens

Assistant Chief Executive Officer: Jean-Pierre Closon

Information

Federaal Kenniscentrum voor de gezondheidszorg - Centre fédéral d'expertise des soins de santé –
Belgian Health Care Knowlegde Centre.
Centre Administratif Botanique, Doorbuilding (10th floor)
Boulevard du Jardin Botanique 55
B-1000 Brussels
Belgium
Tel: +32 [0]2 287 33 88
Fax: +32 [0]2 287 33 85
Email : info@kce.fgov.be
Web : http://www.kce.fgov.be

Quality Assurance of rectal cancer diagnosis and treatment – phase 3: statistical methods to benchmark centres on a set of quality indicators

KCE reports 161C

ELS GOETGHEBEUR, RONAN VAN ROSSEM, KATRIEN BAERT, KURT VANHOUTTE, TOM BOTERBERG, PIETER DEMETTER, MARK DE RIDDER, DAVID HARRINGTON, MARC PEETERS, GUY STORME, JOHANNA VERHULST, VLAYEN JOAN, VRIJENS FRANCE, STIJN VANSTEELANDT, WIM CEELEN.

Belgian Health Care Knowledge Centre
Federaal Kenniscentrum voor de Gezondheidszorg
Centre fédéral d'expertise des soins de santé
2011

KCE reports 161C

Title : Quality Insurance of rectal cancer – phase 3: statistical methods to benchmark centers on a set of quality indicators

Authors : Els Goetghebeur (UGent), Ronan Van Rossem (UGent), Katrien Baert (UGent), Kurt Vanhoutte (UGent), Tom Boterberg (UZ Gent), Pieter Demetter (Erasme), Mark De Ridder (UZ Brussel), David Harrington (Harvard), Marc Peeters (UZ Antwerp), Guy Storme (UZ Brussel), Johanna Verhulst (UZGent), Vlayen Joan (KCE), Vrijens France (KCE), Stijn Vansteedlandt (Ugent), Wim Ceelen (UZgent)

Reviewers : none

External experts : PROCARE members: Jouret-Mourin Anne (UCL), Kartheuser Alex (UCL), Laurent Stephanie (UZGent), Molle Gaëtan (Hôpital Jolimont La Louvière), Penninckx Freddy (UZ Leuven, president steering group), Van Laethem Jean-Luc (ULB), Vindevoghel Koen (OLV Lourdes Waregem)

De Bethune Xavier (ANMC), Legrand Catherine (UCL), Michiels Stefan (Institut Bordet), Rommel Ward (Vlaamse Liga tegen Kanker)

Acknowledgements : The authors thank the PROCARE steering group, their volunteer contributors and patients for making this continued effort for quality improvement. The authors are also grateful to Alain Visscher (Ugent), Geert Silversmit (Ugent) and Carine Staessens (Ugent) for technical assistance in the making of this report. Finally the authors thank Elisabeth Van Eyck (BCR) and Koen Beirens (BCR) for their assistance on the PROCARE database.

External validators : Pr Johan Hellings (ICURO), Pr Pierre Honoré (CHU Liège), Pr Hans C. van Houwelingen (Leiden University).

Conflict of interest : Any other direct or indirect relationship with a producer, distributor or healthcare institution that could be interpreted as a conflict of interests: Vindevoghel Koen

Disclaimer :

- The external experts were consulted about a (preliminary) version of the scientific report. Their comments were discussed during meetings. They did not co-author the scientific report and did not necessarily agree with its content.
- Subsequently, a (final) version was submitted to the validators. The validation of the report results from a consensus or a voting process between the validators. The validators did not co-author the scientific report and did not necessarily all three agree with its content.
- Finally, this report has been approved by common assent by the Executive Board.
- Only the KCE is responsible for errors or omissions that could persist. The policy recommendations are also under the full responsibility of the KCE.

Layout: Ine Verhulst

Brussels, July 12th 2011

Study nr 2010-04

Domain: Good Clinical Practice (GCP)

MeSH: Rectal neoplasms ; Quality of health care ; Quality indicators, health care ; Benchmarking ; Regression Analysis

NLM classification: WI 610

Language: English

Format: Adobe® PDF™ (A4)

Legal depot: D/2011/10.273/40

This document is available on the website of the Belgian Health Care Knowledge Centre

KCE reports are published under a "by/nc/nd" Creative Commons Licence (http://creativecommons.org/licenses/by-nc-nd/2.0/be/deed.en).

How to refer to this document?

Goetghebeur E, Van Rossem R, Baert K, Vanhoutte K, Boterberg T, Demetter P, De Ridder M, Harrington D, Peeters M, Storme G, Verhulst J, Vlayen J, Vrijens F, Vansteedlandt S, Ceelen W. Quality Insurance of rectal cancer – phase 3: statistical methods to benchmark centers on a set of quality indicators. Good Clinical Practice (GCP). Brussels: Belgian Health Care Knowledge Centre (KCE). 2011. KCE Report 161C. D/2011/10.273/40



Executive summary

INTRODUCTION

In 2004, the Belgian Section for Colorectal Surgery, a section of the Royal Belgian Society for Surgery, decided to start PROCARE (PROject on CAncer of the REctum), a multidisciplinary, profession-driven and decentralized project with as main objectives the reduction of diagnostic and therapeutic variability and improvement of outcome in patients with rectal cancer. All medical specialties involved in the care of rectal cancer established a multidisciplinary steering group in 2005. They agreed to approach the stated goal by means of treatment standardization through guidelines, implementation of these guidelines and quality assurance through registration and feedback.

In 2007, the PROCARE guidelines were updated (Procare Phase I, KCE report 69). In 2008, a set of 40 process and outcome quality of care indicators (QCI) was developed and organized into 8 domains of care: general, diagnosis/staging, neoadjuvant treatment, surgery, adjuvant treatment, palliative treatment, follow-up and histopathologic examination. These QCIs were tested on the prospective PROCARE database and on an administrative (claims) database (Procare Phase II, KCE report 81). Afterwards, 4 QCIs were added by the PROCARE group.

Centres have been receiving feedback from the PROCARE registry on these QCIs with a description of the distribution of the unadjusted centre-averaged observed measures and the centre's position therein. To optimize this feedback, centres should ideally be informed of their risk-adjusted outcomes and be given some benchmarks. The PROCARE Phase III study is devoted to developing a methodology to achieve this feedback.

OBJECTIVE AND RESEARCH QUESTIONS

The current study aims to develop a methodology to evaluate the hospital quality of care for rectal cancer patients on the basis of the available set of QCIs. Notwithstanding its specific application**,** the methodology developed ideally is generic in nature and applicable to other cancers and other (chronic) diseases.

This project specifically addresses the following research questions:

1. Which patient characteristics (prognostic factors) should be taken into account to obtain a fair evaluation of the quality of care for rectal cancer patients, based on the available set of QCIs?
2. What is the most suitable statistical method - taking into account feasibility - to identify low- and high-performing centres?
3. How to combine process and outcome indicators in the 8 domains of care? How to define a composite score based on these 8 domains of care? Can some indicators be omitted, or in other words, which indicators are essential to evaluate the performance of centres?
4. How can results be presented into a feedback to individual teams?

FACTORS AFFECTING OUTCOME IN RECTAL CANCER

METHODS

A literature search was performed on the indexed literature. Prognostic factors were searched for overall survival and local recurrence.

RESULTS

The main prognostic factor for overall survival is stage at presentation. Patients with bowel obstruction, perforation, serosal invasion or peritoneal metastasis fare worse, after adjustment for other prognostic factors. Contrary to age, gender is not a prognostic factor. Socioeconomic deprivation was also prognostic of colorectal cancer survival in several studies. In addition, a wide array of pathological prognostic variables, macroscopic as well as microscopic and molecular, were identified. Prognostic factors for local recurrence can be found in appendix 2.

CONCEPTUAL FRAMEWORK TO BENCHMARK CENTRES ON QUALITY INDEXES

The methods proposed here build on the previously developed set of 44 QCIs.

Methods to perform provider profiling recognize three steps: (1) a risk-adjustment step, (2) the choice of a QCI (which can be a composite score), and (3) the benchmarking itself. Each of these steps has been studied based on methodological literature and results from simulations studies:

STEP 1: RISK ADJUSTMENT

The purpose of risk adjustment is to correct the centre's effect on patients' treatment outcome for patient characteristics. Four types of statistical methods are available to perform risk adjustment:

1. Standard outcome regression methods, adjusting for available confounders (called the fixed effects model);
2. Outcome regression methods incorporating random centre effects (called the random effects or hierarchical model);
3. Methods using the propensity score, i.e. the estimated probability that a patient with a given set of risk factors was treated in each of the considered hospitals;
4. Instrumental variables methods, where the instrumental variable, a variable which predicts the choice of the hospital and which not further predicts the outcome, is used as a vehicle to estimate the true hospital effect. The patient's home residence-centre distance is usually used as an instrumental variable.

The latter method was not pursued in this project due to the unavailability of the patient's home residence-centre distance.

The methods were evaluated on factors such as assumption of the model, precision and accuracy of the estimation, interpretation of results, feasibility, influence on regression-to-the-mean bias (the tendency for institutions that have been identified as 'extreme' to become less extreme when monitored in the future), shrinkage of centre effects estimates, ability of the model to detect outliers, ability of the model to handle different sample sizes and handling of the multiple testing issue. Results from an extensive simulation exercise showed that none of the three former methods uniformly performs best.

STEP 2: DIMENSIONS OF CARE AND THEIR AGGREGATION

Reducing a set of 44 indicators to a more manageable size is the scope of this chapter. There is substantial methodological literature on the construction of composite scores for benchmarking of centers in cardiac surgery in the US, and this step builds on these methods.

The selection of QCIs to be retained in the construction of the composite quality index is guided by the following main principles: involve indicators of clinical relevance/importance, involve indicators which allow for meaningful discrimination due to substantial variation between centers; involve indicators for which available data are reliable; and involve indicators that empirically show a strong association with the different dimensions covered by the QCIs.

Once the selection of individual indicators is performed, different methods exist to combine them into a composite score (quality index): simple or weighted averaging, composite opportunity model, scaled combinations or all-or-none scoring. Arguments described in the scientific report lead to the choice of all-or-none scoring.

Two strategies have been proposed to construct quality indexes:

1. The clinical-statistical approach, which starts from QCIs deemed clinically most relevant and discriminating, and then performs a factor analysis to evaluate whether selected individual QCIs capture the full variation of good-quality QCIs;
2. The statistical-clinical approach, which starts from a factor analysis of the good-quality QCI, and then selects the most clinically important and discriminating individual QCIs.

Feedback and consensus of clinicians are essential in both approaches. The second approach (the statistical-clinical approach) was selected and implemented in this project.

STEP 3: REVIEW OF STATISTICAL METHODS FOR PROVIDER PROFILING

Once a composite score has been defined, the next step is to compare centres. This step builds on previous work performed on cardiac surgery in the US and on audits of colorectal cancer surgery in The Netherlands. Visualisation of variation in quality indexes between centres is studied caterpillar plots (either on the logit scale, on the natural scale or on excess probabilities).

DESCRIPTION OF THE PROCARE DATABASE

METHODS

The PROCARE database available for this project covered the period between 2006 and August 2010. The national coverage of the database was assessed by comparing it to the statistics of the Belgian Cancer Registry (available until 2008). Descriptive statistics of the baseline variables in the PROCARE database were produced.

RESULTS

Representativeness of the PROCARE database

The national coverage of the database is around 30% (34.6% in 2006, 31.2% in 2007, 26.4% in 2008). Large variations in coverage are observed by clinical stage: while cStages I, II and III achieve coverage rates above 35% (48% for cStage III), they drop to 25% for patients with cStage IV.

Description of the study cohort

A total of 3328 patients treated in 79 centres were available for analysis (median number of patients included per centre = 33). Mean age was 67 years (SD=12 years), and 61% of patients were male. To be noted was the high number of missing data in the database: age at diagnosis (6%), cStage (15%), pStage (14%), tumor level (10%), Body Mass Index (37%), ASA score (21%), mode of surgery (14%), location of tumour (25%), and circumferential margin (36%). Furthermore, follow-up was poor, as only 56% of patients had at least one follow-up form recorded (first follow-up scheduled 6 months after incidence date). Vital status, however, was available for almost all patients because of a direct link with the Belgian national registry.

ANALYSIS OF INDIVIDUAL QUALITY OF CARE INDICATORS

METHODS

All 44 QCIs were analysed individually, and a separate prognostic model was constructed for each of them, using stepwise regression methods to select prognostic factors. Separate categories were introduced or multiple imputation methods were used to handle missing data. Then, three methods (fixed effect, random effect and doubly robust propensity score) were tested to estimate centre effects for outcome QCIs, using the same baseline factors as those identified in previous steps. For process indicators, no adjustment was performed. Finally, centre effects were presented as an excess (relative to the average centre) probability or outcome value.

To protect confidentiality of centres, no centre level descriptive statistics are presented.

RESULTS

For each of the 44 QCIs, the working definition, number of eligible patients, national event rate, prognostic factors and detailed analysis per centre are available in appendix 6 of the scientific report.

The observed results for each individual indicator were high in most cases: mean 90% (SD 8%) for the 16 outcome indicators and mean 85% (SD 11%) for 25 process indicators. However, 3 process QCIs achieved very low rates: the percentage of patients with cStage I-III that have a reported circumferential resection margin (cCRM, 26%), the accuracy of cM0 staging (6%) and the percentage of patients with stage I that received neoadjuvant radio(chemo) therapy treatment (17%).

AGGREGATED QUALITY INDEXES

METHODS

QCIs were aggregated into combined quality indices as a summary measure of quality (1) for all outcome QCIs, (2) for all process QCIs, (3) per domain of care, and (4) eventually into one global quality index.

The following steps were performed to select QCIs to be involved in the aggregated index:

1. Assessment of quality and reliability of individual QCIs: this was performed through a survey among clinical members of the PROCARE group;
2. Evaluation of the number of dimensions needed through principal component analysis and selection of representative QCIs through factor analysis;
3. Assessment of clinical importance through consensus among the PROCARE group;
4. Assessment of discriminating ability (significant p-value for predictive ability with regard to overall survival).

Two methods of aggregating scores were compared: the *all-or-none method* and the *simple averaging method*. The *all-or-none method* evaluates whether a patient reaches patient-level benchmarks for all QCIs included in the index. In the *simple averaging*, excess probabilities are averaged. When follow-up was not sufficiently long (for instance 3-year survival), multiple imputations techniques were used to impute survival of these patients for the all-or-none score calculation.

Aggregated scores involving outcome indicators were adjusted for patient baseline characteristics, using a stepwise regression procedure. The following variables were initially entered in the model: age, gender, BMI, cStage, ASA comorbidity score, level of tumour, mode of surgery, ventral tumour, cCRM, cT4, preoperative incontinence and surgical technique. Centres were labelled as significantly performing below or above percentile 25 (P25) or percentile 75 (P75) if the limits of their 95% confidence interval (95% CI) of excess probability were below the P25 or above the P75.

RESULTS

Aggregated outcome quality index

Of the 16 outcome QCIs, 6 met the requirements for data quality. Three of these were retained to construct the outcome quality index: (1) overall survival, (2) proportion of R0 resections and (3) absence of postoperative major surgical morbidity with reintervention under narcosis after radical surgical resection.

The all-or-none approach revealed to be more discriminating between centres than the simple averaging method. The average quality index observed across all centres was 68%, meaning that in the average centre the three indicators mentioned above were all positive for 68% of the patients. After adjustment of patients characteristics, one centre was statistically significantly below the P25 (excess probability compared to the average centre: -9%) and 3 centres were above the P75 (excess probability compared to the average centre: 8%).

Aggregated process quality index

Of the 28 process QCIs, 12 met the requirements for data quality. The final process quality index was based on 4 QCIs: (1) time between first histopathological diagnosis and first treatment less than 30 days, (2) percentage of patients without APR-Hartmann procedures or total excision of colon and rectum with definitive ileostomy, (3) distal tumor-free margin mentioned in pathology report and (4) at least 10 lymph nodes examined.

Again, the all-or-none approach was preferred to the simple averaging method, because it allowed a better discrimination of centres. All centres considered, the average quality index was 40%. Four and 5 centres significantly fell below the P25 (-11%) and above the P75 (+17%) excess probability respectively.

Aggregated quality indexes per domain

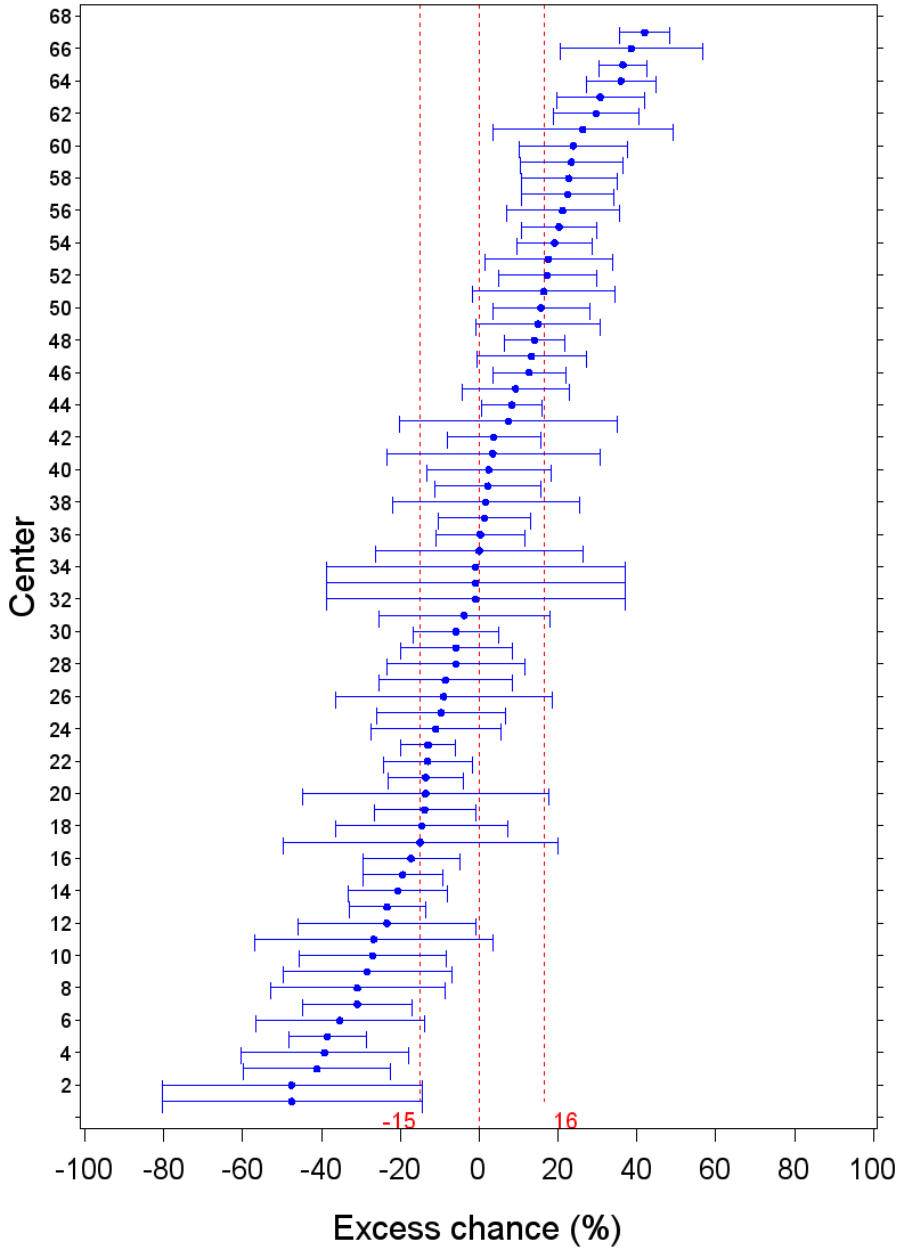
A composite index was created for 5 of the 8 domains. There were not enough eligible patients for domain 5 (adjuvant treatment), 6 (palliative treatment) and 7 (follow-up). For the general domain, only one QCI was selected (overall survival). Results of the selection process are presented in Table 1.

Table 1: Results of the construction of Quality Index per domain of care, using the all-or-none method

Domain of care	Initial number of outcome QCIs	Initial number of process QCIs	QCIs retained in Quality index	Observed Score (mean across centres)	Variability across centres (Interquartile range of excess probabilities)
1 General quality indicators	5	0	Overall Survival	72 %	14%
2 Diagnosis and staging	0	11	Proportion of patients with a documented distance from the anal verge Time between first histopathologic diagnosis and first treatment	60 %	31%
3 Neoadjuvant treatment	1	6	Proportion of cStage II-III patients that received neoadjuvant pelvic radiotherapy	71 %	27%
4 Surgery	9	1	Proportion of R0 resections Major leakage after total mesorectal excision + sphincter-sparing surgery + reconstruction (global) Proportion of abdomino-perineal resections - Hartmann's procedures or total excisions of colon and rectum with definitive ileostomy	70 %	12%
5 Adjuvant treatment	1	4	Not enough eligible patients		
6 Palliative treatment	0	1	Not enough eligible patients		
7 Follow-up	1	1	Not enough eligible patients		
8Histopathologic examination	0	7	Distal margin involvement mentioned after sphincter-sparing surgery or Hartmann's procedure Number of lymph nodes examined	62%	25%

Domain 2, 3 and 8 showed the highest variability between centres. The efforts to harmonize surgical treatment (domain 4) appear to have paid off in this cohort, since less variation was seen in this domain. For more outlying centres, the caterpillar plots warrant an investigation of the centre-specific covariates that contribute to their under/out-performance. Figure 1 presents the caterpillar plot of domain 2.

Figure 1: Caterpillar plot of the adjusted center-specific 'excess' probability for the "all-or-none" domain 2 (Diagnosis and staging) score



Overall composite quality index

Starting from 28 process QCIs and 16 outcome QCIs, and after the application of stringent data quality criteria, 6 outcome and 12 process indicators were retained to construct a composite index. Following the statistical-clinical aggregation procedure, only 2 outcome and 4 process QCIs could eventually be included.

DISCUSSION

The present report documents results of the development and application of methods for the evaluation of quality of care among rectal cancer centres entering information on quality indicators for some of their patients in the voluntary PROCARE registry. The methodology developed is generic in nature and the approach applicable for the monitoring of care in a whole range of cancers and other areas of disease.

The monitoring and evaluation of care envisaged here primarily serves an educational purpose, realized through confidential feedback to individual care centres on their performance on quality indicators - adjusted for patient mix when necessary. This is accompanied by a flagging of lower performance. While this effort is expected to lead to improved care, no specific monitoring of response to this feedback is currently conducted. Whether and how feedback (best) leads to the desired response resulting in an optimization of care is yet to be addressed explicitly and is work for further study.

Feedback takes the form of domain-specific (or more global) presentations of centre performance expressed as an excess percentage: the expected percentage of patients in the centre, beyond that in the average centre, for which all relevant quality indicators in the composite index are met. Centres are plotted in order from the highest to the lowest estimated excess probability and presented with the 95% CI on this excess probability. Confidence intervals overlapping between centres emphasize that the plotting order should not be seen as an indication of the intrinsic rank order of the centres. Substantial uncertainty on the relative quality index performance remains, particularly in centres which contributed few patients to the database so far.

Ideally, confidence intervals on the quality indexes should be compared to absolute benchmarks of good quality care. In practice, such objective benchmarks are not yet available and hence the intervals are simply related to observed percentiles (quartiles) of performance across the centres. Relative benchmarks have the obvious drawbacks that 1) stricter demands are made in domains which generally perform better (i.e. with less variability over the centres) and 2) they are likely to change over time and cannot be used for international comparison.

Domain-specific aggregate scores allow the comparison of specific services across centres. Their construction is designed to involve fewer QCIs than the original set of QCIs in the database without losing much information. The obtained reduction in QCIs for this evaluation thus suggests more economic ways of gathering data.

The application of the proposed methods to the PROCARE database demonstrates the feasibility of the approach in this setting, but also indicates the limitations caused by incompleteness and lack of data quality of the database. Results should not uncritically be used to objectively judge the quality of rectal cancer care in Belgium at present. For instance, while the database structure and maturity is evolving, one hopes a growing number of variables will hit the quality threshold and be allowed to enter the construction for the aggregated quality index. As a result these indexes and the corresponding models for adjustment are expected to change over a period to come. Once the database structure is stable, one should be able to work with a given derived model for the next period without the need for rebuilding the model for each feedback application.

Notwithstanding the database limitations, certain conclusions emerge with some clarity. For instance, the efforts towards standardizing surgical treatment appear to have paid off among the selected surgeons contributing to the PROCARE database. This transpires as the variability in centre-adjusted performance is relatively low in the surgery domain. This is in contrast with results for Domain 4 (diagnosis and staging) and Domain 8 (histopathology) where much more variation is observed. Five and 2 centres in each domain respectively have indicators that score significantly below the P25. This underpins the room for improvement in these domains. An educational system as put in place for surgery could be considered as a fruitful way forward.

WHAT THIS REPORT ADDS

Treatment of cancer is a multidisciplinary process, involving different long and complex clinical pathways to achieve better patient outcomes. Composite scores to benchmark centres already have been developed in the domain of cardiac surgery in US, but this refers to a pathology involving less complex and shorter treatment, and thus reflecting less complex process indicators. The innovative report proposes a methodology to deal with the multidisciplinary aspect of oncology, by working on domains of care, and combines process and outcome indicators into composite indexes, adjusted for patient characteristics when necessary.

CONCLUSION

In conclusion, we emphasize that this report mainly presents a methodological approach to the problem of assessing quality of care. Furthermore, the application to the current PROCARE database demonstrates the practical feasibility of the selected methods in this context. On both accounts methodology as well as application are not 'the final answer', but several important steps were achieved and a feasible approach was presented for work that will continue to progress.

The application to and results from the PROCARE database inherit all the limitations of the current database. Selectiveness of the recorded patients, missing data, and uneven data quality all contribute to results that must be interpreted with caution.

We finally conclude that registration, analysis and feedback on quality of care indicators form an important and irreplaceable instrument for quality of care improvement. It is essential for evidence-based assessments of the performance of health care centres and for the design of intervention strategies to improve care.

RECOMMENDATIONS[a]

Generic methodological recommendations, applicable to sets of QCIs developed in oncology

- **Since standard outcome regression methods (called fixed effects models), adjusted for small sample sizes seem to perform best in situations where a large number of centres treat a small number of patients, which is typically the case in Belgian oncology centres, the KCE recommends adopting this statistical method in future studies involving risk adjustment.**
- **When the construction of a composite score is considered based on a pilot tested QCI set, the selection of the most appropriate QCIs should be based on the following factors:**
  - **Quality and reliability of individual QCIs;**
  - **Clinical importance;**
  - **Potential of discriminating centres;**
  - **Predictive ability for important outcomes.**
- **The *all-or-none approach* should be used to construct the composite score.**
  - **For visualising the variability in quality between centres, the KCE recommends to use funnel plots as a tool for first inspection. Caterpillar plots of excess probabilities, adjusted for case mix when necessary, should be used for the benchmarking of centres.**
  - **For the benchmarking of centres, the KCE recommends to use external benchmarks / target values.**
  - **To develop an effective and long-lasting quality system, as recommended in KCE report 152, the data registration has to be as complete as possible. Data managers available in hospitals in the context of the National Cancer Plan should be dedicated to this.**

Specific recommendations applicable to the PROCARE registry

- **In order to diminish the actual burden of the PROCARE data registration, the KCE recommends a three-fold strategy:**
  - **From the present list of 44 QCIs, key indicators should be selected that result from the construction of the composite score;**
  - **The PROCARE data registration should be limited to the information needed to calculate the selected QCIs and to allow for risk-adjustment;**
  - **When possible, the administrative database (AIM/IMA data) should be linked to the PROCARE database.**
- **In view of the low coverage of the PROCARE registry, and to allow a fair and unbiased comparison between the participating centres, the KCE recommends that a participating centre should be obliged to register all eligible patients. Feedback to the centres should exclude data from centres where enrolment is selective.**
- **Unless the coverage increases, the results of the PROCARE database are not suitable for international benchmarking at the moment.**
- **Based on the PROCARE results, a broad debate including all stakeholders is necessary, to guarantee the good continuation of the project while maintaining its quality.**

a The KCE is the only responsible for the recommendations given to the public authorities

Agenda for future research

- **The effect of socioeconomic factors on the modelling of outcome QCIs and the benchmarking of centres should be evaluated.**
- **It should be tested whether the distance from a patient residence to the hospital would improve the statistical methodology to adjust outcome QCIs for patient characteristics.**

Scientific Summary

Table of contents

LIST OF TABLES **3**
LIST OF FIGURES **5**
LIST OF ABBREVIATIONS **6**
ABBREVIATED LIST OF AVAILABLE QCIS **7**
1 INTRODUCTION AND RESEARCH QUESTIONS **9**
1.1 BACKGROUND 9
1.2 THE CURRENT PROJECT 9
2 CONCEPTUAL FRAMEWORK AND METHODOLOGICAL APPROACH FOR THE ESTIMATION AND BENCH MARKING OF CENTER EFFECTS ON QUALITY INDEXES **11**
2.1 PATIENT MIX ADJUSTMENT FOR QUALITY OF CARE INDICATORS 11
2.1.1 The causal inference framework for center effects 11
2.1.2 A motivated choice among different modeling approaches 12
2.1.3 Limitations due to restricted variable access 13
2.1.4 On the Instrumental Variables method 14
2.1.5 Outcome regression methods and propensity score methods 15
2.1.6 Scope of analysis 18
2.2 LITERATURE SEARCH FOR PROGNOSTIC FACTORS OF RECTAL CANCER 20
2.2.1 Methodological approach 20
2.2.2 Results 20
2.2.3 Discussion 21
2.3 DIMENSIONS OF CARE AND THEIR AGGREGATION 22
2.3.1 Goal 22
2.3.2 Key approach 23
2.3.3 Criteria for indexes 24
2.3.4 Other issues in the construction of quality indicators and quality indexes 31
2.4 REVIEW OF STATISTICAL METHODS FOR PROVIDER PROFILING 33
2.4.1 Goals 33
2.4.2 Methods for provider profiling in the literature 33
2.4.3 Visualizing variation between centers 35
2.4.4 Graphical presentation of center results 37
2.5 PROPOSED APPROACH FOR PROCARE 40
2.5.1 Choice of index 40
2.5.2 Adjusting for patient specific factors 40
2.5.3 Bench marking 40
2.5.4 Visualization 40
3 APPLICATION AND RESULTS ON THE PROCARE DATABASE **42**
3.1 INTRODUCTION 42
3.2 METHODS 44
3.2.1 Descriptive statistics 44
3.2.2 Identification of relevant prognostic factors 45
3.2.3 Estimation of unadjusted and case mix adjusted center effects 46
3.3 DESCRIPTION OF THE STUDY COHORT 47
3.3.1 Centers 47
3.3.2 Prognostic factors 49
3.3.3 Associations between prognostic factors 55
3.3.4 Joint missingness patterns of prognostic factors 58
3.3.5 Follow-up information 58
3.4 DESCRIPTION AND ANALYSIS OF INDIVIDUAL QUALITY OF CARE INDICATORS 60
3.4.1 The worked out case of Overall Survival QCI 1111 60
3.4.2 Tabulated results for the QCIs 69

3.5 SUMMARY POINTS 76
4 AGGREGATED QUALITY INDEXES 78
4.1 INTRODUCTION 78
4.2 METHODS 78
4.2.1 Evaluation of quality and reliability of the QCIs 78
4.2.2 Number of dimensions needed 78
4.2.3 Clinical importance and discriminating ability of selected QCIs per dimension 78
4.2.4 Construction of the Quality Index 78
4.2.5 Risk adjustment 79
4.3 AN AGGREGATED OUTCOME QUALITY INDEX 79
4.3.1 Quality and reliability of the QCIs 79
4.3.2 Concurrent validity of the selected QCIs 79
4.3.3 Clinical importance and discriminating ability of the selected QCIs per dimension 79
4.3.4 Constructing the quality index (QI) 80
4.3.5 Adjusting the outcome QI for patient characteristics 82
4.4 AN AGGREGATED PROCESS QUALITY INDEX 85
4.4.1 Quality and reliability of the QCIs 85
4.4.2 Concurrent validity of selected QCIs 85
4.4.3 Clinical importance and discriminating ability of selected QCIs per dimension 85
4.4.4 Construction of the quality index (QI) 85
4.5 AGGREGATED QUALITY INDEXES PER DOMAIN 90
4.5.1 Overview of results 90
4.5.2 Overview of variation per domain 98
4.6 AN OVERALL COMPOSITE QUALITY INDEX 98
5 MODEL BUILDING TO EXPLAIN EXCESS PROBABILITIES 101
5.1 INTRODUCTION 101
5.2 METHODOLOGY: FORWARD STEPWISE REGRESSION 101
5.3 EXPLANATORY DESCRIPTIVES FOR EXTREME CENTER PERFORMANCE 102
5.4 FORWARD STEPWISE REGRESSION 106
6 DISCUSSION 108
7 BIBLIOGRAPHY 112

LIST OF TABLES

Table 1: Proportion of PROCARE patients in BCR database. 43
Table 2: Proportion of PROCARE patients in BCR database – per cStage (2005 not included) 43
Table 3: Proportion of PROCARE patients in BCR database – per pStage (2005 not included) 44
Table 4: Frequency, cumulative frequency and cumulative percentage of the center sizes in the PROCARE database, up to centers with 10 patients. 48
Table 5: Distribution of the number of patients and number of centers over the different grouping levels based on center size. 49
Table 6: Distribution of cStage in the PROCARE database 49
Table 7: Relation between cStage and (y)pStage in the PROCARE database. 50
Table 8: Staging accuracy for all patients in the PROCARE database, overall and separately for patients that underwent neoadjuvant treatment or not. The (y)pStage is taken as the reference stage to assess under- or over-staging. 50
Table 9: Frequency and percentage of the level of the primary tumor in patients for whom the lower limit of the tumor is known 50
Table 10: Frequency and percentage of the categorized BMI levels for the patients in the PROCARE database for whom BMI is known (63%). 51
Table 11: ASA classification 52
Table 12: Frequency, percentage and cumulative distribution of the ASA co-morbidity score among patients for whom the ASA score is known. 52
Table 13: Distribution of the types of reported tumor complications during surgery among patients registered in the PROCARE database. 53
Table 14: Frequency and percentage of the mode of surgery for the 86.3% patients with a known mode of surgery in the PROCARE database. 54
Table 15: Frequency and percentage of the distribution of ventral tumors for the 75% patients with a known tumor location in the PROCARE database. 54
Table 16: Frequency and percentage of the distribution of cCRM positivity for the 18% patients with a cCRM in the PROCARE database. 54
Table 17: Frequency and percentage of the distribution of (y)pCRM positivity for the 64% patients with a (y)pCRM in the PROCARE database. 55
Table 18: Cross-tabulation of neoadjuvant treatment versus cStage. 56
Table 19: Cross-tabulation of mode of surgery versus cStage among all patients in the PROCARE database 57
Table 20: Cross-tabulation of cStage versus tumor level among patients in the PROCARE database with valid tumor level information. 57
Table 21: Cross-tabulation of ASA score versus BMI among patients with valid information for both BMI and ASA score in the PROCARE database. 57
Table 22: Cross-tabulation of ASA score versus age categories among the patients with valid information for both age and ASA score in the PROCARE database. 57
Table 23: Number of patients in the PROCARE database for whom follow-up information is available (per follow-up period) and the distribution of timing of the last available follow-up. 59
Table 24: The number of patients, events and censored patients, per level of cStage. 62
Table 25: Hazard ratio [95% Wald confidence interval] estimate and corresponding p-value from univariate cStage-stratified Cox regression models for QCI 1111 [OS]. For categorical variables, 'ref.' indicates the reference level. 63
Table 26: Hazard ratio [95% Wald confidence interval] estimate and corresponding p-value from the final multivariate cStage-stratified Cox regression model for QCI 1111 [OS]. For categorical variables, 'ref.' indicates the reference level. 65
Table 27: Number of patients, person years, number of events and event rate for each cStage-Gender-BMI missingness combination among patients eligible for QCI 1111 [OS] with a non-missing/X cStage. 66
Table 28: Row percentages of the cross-tabulation of cStage versus BMI missingness for patients with non-missing/X cStage in the PROCARE database, separately for male and female patients. 66
Table 29: Minimum, P25, P75, maximum and interquartile range of the center effects (%) for QCI 1111 [OS] as estimated by different methods. 66
Table 30: Criteria considered in constructing a QI for outcome of rectal cancer, split by data quality, clinical importance and discriminating ability. 70

Table 31: Criteria considered in constructing a QI for process QCIs related to quality of care of rectal cancer, split by data quality, clinical importance and discriminating ability..... 71
Table 32: Center effect for general (outcome) quality indicators: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level. [NA: not applicable; Center effects for these QCI's were not calculated]..... 72
Table 33: Center effect for (process) quality indicators related to diagnosis and staging: inter quartile range, number of centers significantly below P25 or above P75 and univariate prognostic factors retained at the 5% level. 72
Table 34: Center effect for (process) quality indicators related to neoadjuvant treatment: inter quartile range, number of centers significantly below P25 or above P75 and univariate prognostic factors retained at the 5% level. 73
Table 35: Center effect for (outcome) quality indicators related to surgery: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level. 74
Table 36: Center effect for (process) quality indicators related to adjuvant treatment: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level. 74
Table 37: Center effect for (process) quality indicators related to histopathological examination: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level..... 75
Table 38: Statistical summary measures for the "all or none" QI and simple averaged QI, both based on QCIs 1111 [OS], 1231 [%R0res] and 1234b [%Major_morb]. 81
Table 39: Criteria considered in constructing a QI for outcome of rectal cancer, split by data quality, concurrent validity (PCA), clinical importance and discriminating ability..... 84
Table 40: Statistical summary measures for the "all or none" QI and simple averaged QI, both based on QCIs 1217 [Time_histo-1ther], 1232a [%Defin_osotmy], 1273 [%Dist_Margin_Pos_Rep] and 1274 [#Nodes_Examined]. 86
Table 41 (next page): Criteria considered in constructing a QI for process QCIs related to quality of care of rectal cancer, split by data quality, concurrent validity (PCA), clinical importance and discriminating ability..... 89
Table 42: Summary of QCIs per domain, incl. the domain name, the number of outcome (#O) and process (#P) indicators available and the QCIs retained per domain. The last column shows the average achievement rate over the centers for the derived quality index which is an "all or none" score. 90
Table 43: Overview of quality index results per domain. Inter quartile range and number of centers significantly performing below P25 and above P75 in caterpillar plots of (un-)adjusted excess probability..... 98
Table 44: List of selected process QCIs with sufficient data quality..... 103
Table 45: Excess probabilities (%) for process QCIs for centers with lowest adjusted excess probability for 3 year survival..... 104
Table 46: Excess probabilities (%) for process QCIs for centers with highest adjusted excess probability for 3 year survival..... 104
Table 47: Average center size for the 5 centers with highest and lowest estimated adjusted excess chance of 3 year survival. 105
Table 48: Univariate associations with excess 3-year survival chance as an outcome..... 105
Table 49: Parameter estimates for the significant (at the .05 level) predictors in the model for excess probability of 3 year survival. 106

LIST OF FIGURES

Figure 1: Schematic presentation of the construction of an outcome and process-specific QI 22
Figure 2: Schematic presentation of the construction of domain-specific QIs 23
Figure 3: Decision tree for the first (clinical-statistical approach) suggested item selection strategy ... 28
Figure 4: Decision tree for the second (statistical -clinical approach) suggested item selection strategy 29
Figure 5 Funnel plot for the aggregate index of the process indicators..... 36
Figure 6: Caterpillar plot for QCI 1271 [%Path_Rep_Use] on the natural scale. 38
Figure 7: Caterpillar plot of excess chances for QCI 1271 [%Path_Rep_Use]. 39
Figure 8: Distribution of center sizes, i.e. number of patients registered in the PROCARE database.. 48
Figure 9: Distribution of the difference in recomputed event date and reported event date in the PROCARE database based on the formula above. 60
Figure 10: Kaplan-Meier curves per cStage stratum, estimating the (unadjusted) probability of surviving t years after the incidence of rectal cancer. The unstratified curve is shown in green. Numbers at risk for that curve are shown on the x-axis..... 61
Figure 11: cStage-stratified Kaplan-Meier survival curves with an extra survival curve (red dotted line) showing the probability of surviving t years for a random patient (with available cStage information) in the PROCARE database. The green line presents the survival probability after t years for patients with X/missing cStage. 62
Figure 12: Scatter plot of adjusted center effects (from both fixed- and random effects outcome regression) versus unadjusted center effects for QCI 1111 [OS]. 67
Figure 13: Adjusted excess chance (%) of 3-year overall survival 68
Figure 14: Caterpillar plot of the unadjusted center-specific 'excess' proportion of patients for whom the "all-or-none" outcome score was met. 81
Figure 15: Scatter plot comparing the "all-or-none" score to the simple averaged QI for QCIs 1111 [OS], 1231 [%R0res] and 1234b [%Major_morb]. 82
Figure 16: Caterpillar plot of the adjusted excess probabilities for the outcome "all or none" score. 83
Figure 17: Scatter plot comparing the "all or none" score to the simple averaged QI for QCIs 1217 [Time_histo-1ther], 1232a [%Defin_ostomy], 1273 [%Dist_Margin_Pos_Rep] and 1274 [#Nodes_Examined]. 87
Figure 18: Caterpillar plot of the unadjusted center-specific 'excess' proportion of patients for whom the "all or none" process score was met. 88
Figure 19: Caterpillar plot of the unadjusted center-specific 'excess' probability for the "all-or-none" score for domain 2 'Diagnosis and staging' 91
Figure 20: Caterpillar plot of the unadjusted center-specific 'excess' probabilities of QCI 1221 (as an index for domain 3 `Neoadjuvant treatment') 93
Figure 21: Caterpillar plot of the adjusted center-specific 'excess' probability for the "all-or-none" score for domain 4 'Surgery' 95
Figure 22: Caterpillar plot of the unadjusted center-specific 'excess' probability for the "all-or-none" score for domain 8 'Histopathological examination'..... 97
Figure 23: Caterpillar plot of the adjusted center-specific 'excess' probability for the "all-or-none" composite score..... 100
Figure 24: Model predictions for the adjusted excess probability of 3 year survival. 106

LIST OF ABBREVIATIONS

5-FU	5-fluorouracil
APR	Abdomino-Perineal Resection of the rectum
ASA	American Association of Anesthetists score
BCR	Foundation Belgian Cancer Registry
BMI	Body Mass Index
CABG	Coronary Artery Bypass Grafting
CBSS	Crossroads Bank of Social Security
CEA	CarcinoEmbryonic Antigen (test)
CI	Confidence Interval
cN	nodal status
CRM	Circumferential Resection Margin
cT	clinical Tumor
e_c	Excess probability of achievement
FU	Follow-Up
ICD	International Classification of Diseases
ID	Identification
IMA	Common Sickness Funds Agency (Intermutualistisch Agentschap/L'Agence Intermutualiste)
INSS	Identification Number for Social Security (IdentificatieNummer voor de Sociale Zekerheid/Numéro d'Identification de la Sécurité Sociale)
IQR	InterQuartile Range
KCE	Belgian Health Care Knowledge Centre (Federaal Kenniscentrum voor de Gezondheidszorg/Centre fédéral d'expertise des soins de santé)
MRI	Magnetic Resonance Imaging
N	number
NA	Not Applicable
O	Outcome (QCI)
ORM	standard Outcome Regression Model
p	probability
P	Process (QCI)
P25	1st quartile or 25th percentile
P75	3rd quartile or 75th percentile
PCA	Principal Components Analysis
PME	Partial Mesorectal Excision
PROCARE	PROject on CAncer of the REctum
PS	Propensity Score
QCI	Quality of Care Indicator
QI	Quality Index
QLI	Quality of Life Index
QQ-plot	Quantile-Quantile plot
R0 resection	margin negative resection
RC	Rectal Cancer
RT	Radiotherapy
RX	X-ray
SES	Socio Economic Status
SSO	Sphincter-Sparing Surgery
TME	Total Mesorectal Excision
TNM	Tumor Node Metastasis
TRUS	TRansrectal UltraSonography

ABBREVIATED LIST OF AVAILABLE QCIS

To facilitate reading of this report a list of all QCIs with their abbreviation and with a summary description is presented here.

QCI	Abbreviations	Description
1111	OS	Overall survival
1112	DSS	Disease-specific survival
1112b	RS	Relative survival
1113	LRFS	Proportion of patients with local recurrence
1113b	DFS	Disease-free survival
1211	%DocDist	Proportion of patients with a documented distance from the anal verge
1212	%CT_Preop	Proportion of patients in whom a CT of the abdomen and RX or CT thorax was performed before any treatment
1213	%CEA_Preop	Proportion of patients in whom a CEA was performed before any treatment
1214	%Preop_Bowel_Im	Proportion of patients undergoing elective surgery that had preoperative complete large bowel-imaging
1214b	%TRUS_cT12	Use of TRUS in cT1/cT2
1214c	%MR_cII/III	Use of MRI in cStage II or III
1215	%Preop_Im	Proportion of patients in whom a TRUS and pelvic CT and/or pelvic MRI was performed before any treatment
1216	%cCRM_rep	Proportion of patients with cStage II-III RC that have a reported cCRM
1216b	cM0_Acc	Accuracy of cM0 staging
1217	Time_histo-1ther	Time between first histopathologic diagnosis and first treatment
1221	%Preop_RT	Proportion of cStage II-III patients that received a neoadjuvant pelvic RT
1221b	%(C)RT_cCRM+	Proportion of patients with cCRM = 2 mm on MRI/CT that received long course neoadjuvant radio(chemo)therapy
1221c	%Preop_RT_cI	Proportion of patients with cStage I that received neoadjuvant radio(chemo)therapy
1224	%Preop_cont_5FU	Proportion of cStage II-III patients treated with neoadjuvant 5-FU based chemoradiation that received a continuous infusion of 5-FU
1225	%Completed_preop_RT	Proportion of cStage II-III patients treated with a long course of preoperative pelvic RT or chemoradiation that completed this neoadjuvant treatment within the planned timing
1226	%Surg<12w_after_Preop_RT	Proportion of cStage II-III patients treated with a long course of preoperative pelvic RT or chemoradiation that was operated 4 to 12 weeks after completion of the (chemo)radiation
1227	%grade4_Tox_Preop_RT	Rate of acute grade 4 radio(chemo)therapy-related complications
1231	%R0res	Proportion of R0 resections
1232a	%Defin_ostomy	Proportion of APR- Hartmann's procedure or total excision of colon and rectum with definitive ileostomy
1232b	%stoma1year	Proportion of patients with stoma 1 year after sphincter-sparing surgery
1233a	%Leak_PME	Major leakage after PME + SSO + reconstruction

1233b	%Leak_TME	Major leakage after TME + SSO + reconstruction (global)
1234	30d_mort	Inpatient or 30-day mortality after radical surgical resection
1234b	%Major_morb	Postoperative major surgical morbidity with reintervention under narcosis after radical surgical resection
1235	%Perfor	Rate of intra-operative rectal perforation
1235b	%Pos_Dist_margin	(y)p Distal margin involved (positive) after SSO or Hartmann for low rectal cancer (= 5 cm)
1235c	%Pos_CRM	Mesorectal (y)pCRM positivity after radical surgical resection
1241	%Adj-Chemo<3m	Proportion of (y)pStage III patients with R0 resection that received adjuvant chemotherapy within 3 months after surgery
1242	%Adj_RT<3m	Proportion of pStage II-III patients with R0 resection that received adjuvant radiotherapy or chemoradiotherapy within 3 months after surgery
1243	%Adj_Chemo<12w	Proportion of (y)pStage II-III patients with R0 resection that started adjuvant chemotherapy within 12 weeks after surgical resection
1244	%Adj_5FU	Proportion of (y)pStage II-III patients with R0 resection treated with adjuvant chemo(radio)therapy that received 5-FU based chemotherapy
1245	%grade4_Tox_Prostop_CT	Rate of acute grade 4 chemotherapy-related complications
1251	%cStage4_Chemo	Rate of cStage IV patients receiving chemotherapy
1261		Rate of curatively treated patients that received a colonoscopy within 1 year after resection
1263	%Late_Tox_RT	Late grade 4 complications of radiotherapy or chemoradiation
1271	%Path_Rep_Use	Use of the pathology report sheet
1272	%TME_Qual_Rep	Quality of TME assessed according to Quirke and mentioned in the pathology report
1273	%Dist_Margin_Pos_Rep	Distal margin involvement mentioned after SSO or Hartmann
1273b	%Dist_Margin_Rep	Distal tumor-free margin mentioned in the pathology report
1274	#Nodes_Examined	Number of lymph nodes examined
1275	pCRM_mm_Rep	(y)pCRM mentioned in mm in the pathology report
1276	TRG_Rep	Tumor regression grade mentioned in the pathology report (after neoadjuvant treatment)

1 INTRODUCTION AND RESEARCH QUESTIONS

1.1 BACKGROUND

In 2004, the Belgian Section for Colorectal Surgery, a section of the Royal Belgian Society for Surgery, decided to start PROCARE (PROject on CAncer of the REctum), a multidisciplinary, profession-driven and decentralized project with as main objective the reduction of diagnostic and therapeutic variability and improvement of outcome in patients with rectal cancer. All medical specialties involved in the care of rectal cancer established a multidisciplinary steering group in 2005. They agreed to approach the stated goal by means of treatment standardization through guidelines, implementation of these guidelines (workshops, meetings, training for total mesorectal excision [TME], pathology, radiotherapy and radiology) and quality assurance through registration and feedback.

In 2007, the PROCARE guidelines were updated (PROCARE Phase I) [1]. In 2008, a set of 40 process and outcome quality of care indicators (QCI) was derived and organized in 8 domains of care: General quality indicators (including survival), Diagnosis and staging, Neoadjuvant treatment, Surgery, Adjuvant treatment, Palliative treatment, Follow-up and Histopathologic examination. These QCIs have been pilot tested on the PROCARE prospective database and on an administrative (claims) database (PROCARE Phase II) [2]. Patients who may have been registered in multiple centers (for diagnostic and treatment purposes) were for feedback purposes assigned to one main center by the BCR and registered anonymously in PROCARE with a patient ID (details in Appendix 7). Data submitted for the present study included a single, blinded center assignment per patient, in addition to basic clinical variables and eligibility status along with calculated QCI values per patient whenever possible. Centers have been receiving feedback from the PROCARE registry on these QCIs with a description of the distribution of the unadjusted center-averaged observed measures and the center's position therein. To optimize this feedback, centers should ideally be informed of their risk adjusted outcomes and be given some bench marks. The PROCARE Phase III study launched by KCE in response to a call from the PROCARE group is devoted to developing the methodology to achieve the best possible feedback in this sense.

1.2 THE CURRENT PROJECT

The current study aims to develop a methodology to identify low and high-performing hospitals in the management of rectal cancer on the basis of the available set of QCIs. Notwithstanding its specific application**,** the methodology will be generic in nature and applicable to other cancers, and even other (chronic) diseases. Methods will be further deployed to examine whether and how the number of QCIs involved can be reduced while retaining the most important information per domain. In view of the primary goal, we specifically addressed the following research questions:

1. Which patient characteristics (risk factors) should be taken into account to obtain a fair evaluation among centers of the set of QCIs developed for treatment of rectal cancer?
2. What is the most suitable statistical methodology - taking into account practical feasibility - to quantify performance of centers, and to identify low and high-performing centers?
3. How to combine process and outcome indicators within the 8 domains of care (general QCI, diagnostic and staging, neoadjuvant treatment, surgery, adjuvant treatment, palliative treatment, follow-up, histopathologic examination)? How to define a composite score based on these 8 domains of care? Are there indicators that can be omitted, or alternatively, which indicators are essential to evaluate the performance of centers?
4. How can results be presented into a feedback to individual teams?

The sequence of steps performed towards this goal can be consulted in online appendices to this report, which are listed below and cover documents produced over time as deliverables. These show the initial results which built on a preliminary dataset. We eventually arrived at the full analysis of the dataset of patients diagnosed since 2006 and followed up until August 27th 2010, which is summarized in the present overall report. This final report aims to inform a more general audience of stakeholders rather than primarily statisticians and epidemiologists, who we refer to the more technical appendices for further detail. The first 2 chapters focus on methodology, while. chapters 3, 4 and 5 cover application of the methods to the PROCARE database. In chapter 2 we describe the basic methodological concept and options taken in this project. In section 2.1 we elaborate on how to adjust center effects for patient-specific factors, while section 2.2 reports on a literature study of known prognostic factors for outcome. In section 2.3 we propose an approach for the construction of aggregated scores. Section 2.4 discusses methods for provider profiling as well as bench marking of the quality indicators. Section 2.5 summarizes our proposed approach. In Chapter 3 we describe our database and present results of patient-mix-adjusted center effects for individual outcome QCIs and unadjusted center effects for process QCIs. Chapter 4 summarizes the construction of aggregated domain scores and presents center effects on the corresponding quality indexes. It ends with an exploration of the explanation of center effects in terms of center characteristics including their general performance on process indicators. In chapter 5 we end with a discussion and a preliminary exploration of the relation between center effects on outcome and on process quality indicators. Detailed figures, tables and additional results are provided in appendices. Appendices 1-6 complement Chapter 2, Appendix 7 and 8 contain additional data for Chapter 3 and Chapter 4, respectively.

- Appendix 1: Detailed discussion of the methodology with technical specifications and a simulation study
- Appendix 2: Protocol, results and discussion of the literature review
- Appendix 3: Case Record form of PROCARE
- Appendix 4: Original definitions of the QCIs defined by PROCARE and the latest working definitions used in this report
- Appendix 5: Physician survey on Clinical importance and Measurement Quality for the QCIs
- Appendix 6: Descriptives of prognostic factors (section 1.1) and QCI's (sections 1.2-1.8) and detailed results of analysis of individual QCIs (section 2)
- Appendix 7: Analysis of selectivity of the PROCARE Cohort: comparing the PROCARE and BCR database
- Appendix 8: Detailed approach and results for the construction of aggregate quality indexes

2 CONCEPTUAL FRAMEWORK AND METHODOLOGICAL APPROACH FOR THE ESTIMATION AND BENCH MARKING OF CENTER EFFECTS ON QUALITY INDEXES

2.1 PATIENT MIX ADJUSTMENT FOR QUALITY OF CARE INDICATORS.

2.1.1 The causal inference framework for center effects

In this section we develop a method that allows adjusting QCI for the patient mix treated by the center ultimately arriving at one or more global quality indexes with well understood bench marks. This adjustment for patient mix is anticipated to be more important in the outcome than process domain, since process QCIs have by definition been adapted to the patient type where needed. To reach the goal of identifying low and high-performing hospitals in the management of rectal cancer (RC) on the basis of the available set of QCIs, we first translated the question in a conceptual and operational framework.

The framework most relevant here is that of causal inference. We wish to evaluate not just an association between centers and outcomes, but the effect of hospital characteristics, in addition to the patient characteristics, on the patient's treatment quality or outcome. In other words, we aim to find out what would happen if a well-defined group of patients was treated by provider A rather than provider B. For this purpose we first wish to correct for patient-specific characteristics but not for hospital characteristics, since these are considered part of the package the hospital provides to the patient. Once patient-mix-adjusted center effects have been estimated, we will consider hospital-specific characteristics that may help explain any variation in center effects and point to ways of improvement.

To arrive at a patient-risk-adjusted measure of hospital performance, the project aimed to have access to data from two cohorts: the smaller but comprehensive PROCARE database and an administrative (claims) database. The original 40 process and outcome QCIs proposed earlier [2] can be derived from the combined data in those databases. Further information is available there on the patient's background and general health, which may be prognostic for the treatment process and outcome QCIs. As the project got launched, however, the PROCARE steering group refused linking the PROCARE database to other existing databases for this goal. As a result, some of the original QCIs are no longer measurable and few baseline covariates remain. Nevertheless, we did have access to clinical baseline variables. The former aspect is largely remedied through the proposed updated set of QCIs given in Appendix 4. The substantially limited access to potential confounders of the relationship between center and outcome appears much more serious. It has led to some modification of the methodological development plan and will ultimately weaken its current application in this setting as described in the Section 2.1.3

At both levels of the analysis, special attention will go to center sizes which are known to vary substantially. At the first level, we will need to consider that centers that provide data on just a few patients produce a very weak evidence base for the center's general effect measurement. If these few patients were selected among more, they carry the additional risk of selection bias. Confidence/credibility intervals of the center-specific QCI summary may then be so wide that they become non-informative and cover regions of excellent average and lower performance. Random effects models and/or Bayesian models are designed to overcome this in part by borrowing information from an assumed population distribution of center effects.

Center size may have an additional impact beyond the precision of our estimates. For instance, high-volume centers are likely to be specialized and hence perhaps subject to a more complicated case mix with better or worse comparative performance for that very reason. For the purpose of evaluating center-specific quality of care, we do not plan to adjust for center-specific covariates, but to consider them as part of the center package just as other center-specific covariates. Hence, in its potential role of prognostic factor, center size will only enter the analysis in a second round, i.e. to explain existing center effects. Equally, any interaction effects between center and patient-specific covariates would indicate that similar patients fare differently in different centers. For instance, a center specialized in geriatric medicine may care particularly well for older rectal cancer patients. We will not control for this in the primary analysis but propose to explore such mechanisms in a second round, when one wishes to explain differences seen in (patient-mix-adjusted) center performance.

With the above considerations in mind we consider three main methods for risk-adjustment:

1. Standard outcome regression methods (ORM), adjusting for available confounders and possibly incorporating random center effects.
2. Methods using the propensity score (PS), this is the estimated probability that a patient with a given set of risk factors was treated in each of the considered hospitals.
3. Instrumental variable (IV) methods where the IV, i.e. a predictor for the hospital which is not further predictive of the outcome, is used as a vehicle to estimate the hospital effect.

The vast majority of measured QCIs are binary measures. In addition, there are several important right-censored survival time measures (to be summarized in, for instance, overall 5-year survival probability, relative survival and disease-specific 5-year survival probability). Beyond this, there is a QCI describing the number of lymph nodes examined, which is most naturally to be approached as a continuous or count measure. Since treatment of continuous outcome measures tends to be the most straightforward, methodologically speaking, we will concentrate in this text on the development for binary and survival type outcomes. As a final note: since the QCI for 5-year survival is not mature in the current PROCARE database, which is restricted to patients diagnosed since 2006 and followed up until August 27th 2010, we will focus on 3-year survival for the database received.

2.1.2 A motivated choice among different modeling approaches

A detailed discussion on methods for patient mix adjustment can be found in Appendix 1. Here, we explain our general findings and options taken, which are further supported by developments in the text below as well as in an extensive technical section (Appendix 1, Chapter 9). In terms of the statistical paradigm followed, we have elected the frequentist (repeated sampling based) approach rather than the Bayesian (degree of belief) approach. In principle, both approaches can handle similar questions under corresponding models and assumptions on the data. Since the BCR – who will eventually adopt these methods in its routine feedback practice – is familiar with the frequentist approach, this has been a big consideration. We recognize that in their most straightforward implementation, frequentist analyses tend to avoid resampling techniques and rely on large sample considerations which are not automatically satisfied in a growing data register. On the other hand, Bayesian methods tend to rely more on (subjective) prior information in such instances, which is not entirely satisfactory in our setting either. Both approaches may struggle when fitting models on centers contributing few patients to the register. Knowing that the small sample centers are recognized as providing preliminary data to be treated cautiously anyway, and that they should eventually mature to providing a sufficiently rich database, our choice for the frequentist paradigm stands.

We thus set out to consider three classes of methods, from the most standard to the most state-of-the-art, for risk adjustment in the evaluation of causal effects: outcome regression methods, propensity score methods and instrumental variables methods. We conducted our evaluation considering both the general assessment of quality of care and the specific context of the PROCARE database as well as the data structure made available to us.

The first two approaches (ORM and PS) rely on the assumption of 'no unmeasured confounders' for estimation of the (causal) effect of center on quality outcome. In contrast, the instrumental variables approach allows for unmeasured confounders, but requires an instrumental variable instead, i.e. a variable which is associated with center but not otherwise with the natural outcome of the patient. Important limitations in light of these requirements result from the restricted access to baseline data in the PROCARE database. We had access to, for instance, age, gender, cStage at diagnosis and ASA score (on a 4 point scale), but not to variables such as

1. socio-economic status (SES),
2. specific co-morbidity, or
3. patient distance from the treatment center.

2.1.3 Limitations due to restricted variable access

We briefly explain the limitations entailed by missing the three above-mentioned variables and the methodological choices resulting from that. The three variables mentioned are representative of different types of information not directly available in the PROCARE database, but potentially available through linkage with other existing databases such as the IMA database.

1. SES represents a potential confounder for the center-quality relationship through the link with a specific natural risk profile (over and beyond what is contained in age-gender-cStage), while it may at the same time influence treatment quality, irrespective of the center, for instance because patients in a higher SES stratum more easily receive a more expensive or specific treatment [3 4, 5].
2. Specific co-morbidities could definitely change the risk profile and would justify or may even require an adapted treatment.
3. Distance, or a derived measure, such as distance to a given center relative to the nearest center distance, is likely to be a strong predictor of center choice and may be an instrumental variable if it does not further affect the quality outcome. In several studies a measure of distance, location or region was proposed in this sense [6-12]. Alternatively, if distance affects outcome because of its association with region and perhaps a particular local toxin or genetic type of cancer, or if it moderates treatment - for instance through reduced visits with a longer distance, or the choice of a closer center when more frequent visits are required - it is a confounder or mediator and not an instrument.

So, first, all three variables could be confounders, that is, a common cause of center choice and outcome quality, for which one needs to adjust if the pure center effect is to be measured. Second, both SES and co-morbidity may generate a different treatment response for otherwise similar patients (across all centers). In an optimal quality setting SES should not influence treatment while co-morbidity should. In light of this, some scientists believe that one should not adjust for SES when analyzing treatment effects in view of benchmarking. We argue that in a practical setting where SES does influence treatment across the board (for all centers) the most relevant effect measure for the patient as well as the most fair comparison of quality delivered by centers is obtained after adjusting effect measure for SES. The arguments for this are summarized in Section 2.1.3.1.

Third, if distance between patient and treatment center influences the treatment (schedule) received and hence outcome, it affects outcome directly and can no longer serve as an instrumental variable. The general implications of all three points for our analysis approach are described following the next Subsection.

2.1.3.1 Arguments for adjusting for factors such as patient's socio economic status

Many patient-specific characteristics (at diagnosis) influence the outcome of rectal cancer patients. Not all of these factors are known or can be carefully measured. Typically, we adjust for just a few pre-treatment patient-specific factors, including age, gender, cStage at diagnosis, possibly ASA score, etc... The implication is that we predict risks of individuals based on limited prognostic information and then evaluate how the observed risk in a center deviates from it. The question is, should or should we not in principle also adjust for factors as SES if possible (potentially obtained through a link with the IMA database), knowing that in practice:

1. different SES may be treated differently across centers: e.g. patients with a higher SES receive a more expensive and better treatment element [3], , and
2. different SES patients may present with different natural disease progression because of distinct environmental, genetic, co-morbidity conditions beyond what has been measured through cStage, ASA-score etc. in a necessarily limited prospective voluntary register.

Without adjustment we fail to correct for a possibly associated differential natural risk (which is always needed) as well as for SES-related differences in treatment (which we may or may not wish to adjust for, if depending on SES, the treatment adaptation happens irrespective of the treatment center). With adjustment, we adjust for both different risk levels and different treatment levels associated with SES and hence do not penalize centers who carry a heavier load of the 'worse treated patients'.

In conclusion, if our perspective is the one of the patient ('given who I am, where should I go to get the better treatment/outcome'), then the most relevant answer would be found after adjusting for SES. This is true whether or not we evaluate the centers for the population with their own typical patient mix or for a fixed population average outcome. Hence, one should adjust for SES (like) factors if possible at all, to get the most scientific and relevant answers as well as an honest comparison of *differential* performance between centers.

If we would simply wish to alert the center of the fact that it has worse outcomes than other centers (which may be due to its different patient mix which may or may not be well treated), then an unadjusted analysis is in order. Since our primary goal in this section is on adjusting for patient mix, we will adjust for SES whenever possible, even though unadjusted reports have their own contribution to make.

As we are unable to adjust the analysis for some known confounders, we must acknowledge that patient adjustments constructed (by regression and the propensity score method) will only partially correct and the residual center effects defined may result in part from differential representation of these factors in the center's patient mix. Whether or not this is the case, can only be examined once the additional set of covariates becomes available for analysis.

The propensity score approach might be weakened, as the distance, a likely strong predictor of center effect, cannot be included in the propensity score. This would be a special point of concern when the distance is also moderately associated with the outcome, for then it is an important confounder, although not otherwise.

2.1.4 On the Instrumental Variables method

For the combined set of reasons stated below, we will not use instrumental variables in this project.

- Lacking the measures on the patients distance to every center considered we are unable to involve it in the analysis as an instrumental variable. No other potential instrumental variables were recovered based on the literature search described in Section 3.2.
- If distance is associated with outcome or treatment (schedule), either because the schedule gets adapted to the distance or the other way around, instrumental variable property is violated and it becomes an invalid instrument.

- Preliminary results indicate that the presence of that many centers with a correspondingly small propensity makes that there is too little information about the causal effect of the centers if one wishes to allow for unmeasured confounders. This is translated into confidence intervals so wide they become unusable.

Even though the instrumental variables approach is unworkable in the current setting, there may be a future role for it. While we cannot recognize the actual identity of specific centers and hence have no direct information on center type, it is clear that certain centers differ from others in important aspects. For instance, University hospitals tend to differ in size (larger), in equipment and staff they can draw on (more state of the art, costly, highly trained) and in the population they attract (more difficult cases). As a cluster they tend to draw on more resources which would suggest they have their own standard to aspire to. They are centers specifically dedicated to the advancement of science and its implementation in practice. It might be worth having a secondary analysis of center effects confined to this cluster of fewer and larger centers, for the development of their own benchmark. Here the argument of tiny propensity scores would vanish and distance could again become a workable instrument on the condition the instrument is rich enough to avoid multicollinearity in a two stage regression and no serious confounding or mediation through the distance remains.

2.1.5 Outcome regression methods and propensity score methods

For our goal, we now focus on the outcome regression methods and propensity score methods in more detail. Notwithstanding the limitations in the current setting, both approaches a priori have their merit here and more generally when the full scale of confounders and prognostic factors for center choice are included in the analysis.

To arrive at a meaningful evaluation and the comparison of outcome regression and propensity score methods, several basic choices are made. Different methods concentrate on direct modeling of distinct target parameters. These involve patient-specific, center-specific or population-specific risk estimation. Patient-specific adjustments are the more standard direct focus of modeling and will form building blocks of our models. Here, population-specific risks express risk of a certain event if all patients in one chosen common study population were treated in a given center. In contrast, center-specific measures compare the observed risk for patients in a given center with the risk that these same patients would have experienced in some ‘average’ center. Evidently, from the measures conditioning on more detailed information the more averaged measures can always be derived, but not the other way around. It was found that center-specific treatment effects are best evaluated on the patient mix they themselves currently treat. Hence this will be our primary aggregated outcome measure, even though this means that different centers are judged on different patient mixes. This reference was seen to be particularly relevant in a stable landscape where the patient mix tends not to change much over the years. Drastic interventions in the treatment landscape could of course make this stability premise untrue.

The center-specific treatment effect will most easily be derived from outcome regression models (fixed or hierarchical). Current implementation of a (fixed effect) propensity score method naturally focuses on population averaged effects only. As indicated, such an effect measure has the great advantage that it constitutes a common reference outcome for all centers and can be derived from the results of all methods. Our comparisons of results of different approaches in this report will examine both measures before coming to a conclusion in this report. While a propensity score based matched analysis can in principle be developed, this is documented to be less reliable than what we obtain through the doubly robust propensity based methods, a version of the method which protects against misspecification of either the outcome regression model or the propensity score model for center choice, and will therefore not be pursued here.

Either approach and target parameter leaves the question: relative to which 'specific center' effect do we express our adjusted outcome measures? There are (at least) two basic options studied in Section 2.4.2: an external (international) reference or standard, and an internal (to the PROCARE database) reference. Here we briefly discuss the latter only – in view of the modeling choices to be made. The discussion on benchmarking and quality standards is left to Section 2.4.2. Standard regression models, involving a separate effect for each center in addition to the effects of patient-specific characteristics parameterize center deviations, typically on the log odds scale for binary data and log hazard scale for survival data, from either

- a single chosen reference center (the first, last, largest, best, or on a percentile) - through 'dummy coding'
- the average center effect, averaged over all centers (on the given scale) - through 'unweighted effect coding' or
- the average center effect, averaged over all patients - through 'weighted effect coding' (= the "national" mean).

With weighted effect coding, large centers get more weight in defining the reference which is not the case with unweighted effect coding. In what follows we took our definition of `average' center from the unweighted effect coding above (and a corresponding average when a hierarchical random effects model has been constructed). This means that for logistic regression models, for instance, the mean center level was calculated on the logit (i.e. log odds) scale. In the end, the probabilities generated for this `average' center were calculated by back transformation from the log odds scale. This construction has advantages and disadvantages. The variation between centers is typically more symmetrically or even normally distributed on the logit scale than on a probability scale. The probability of achieving a good quality outcome may indeed be close to 100% for some QCIs and hence tend to be skew. On the normal scale the mean is an intuitive and efficient measure of central location. This approach does however become complicated when some centers have an observed proportion of 100% and thus a logit equal to infinity which would inadvertently determine the mean. Because we do not involve centers with less than 5 patients in the logistic regression and because will use the Firth correction (as explained at the end of this section), and this tends to shrink results somewhat towards the mean, we do not have this problem with infinity. An alternative approach would consist of recalculating the average probabilities over all centers directly on the probability scale for each combination of baseline covariates whenever a center-specific reference value is chosen. The specific choice of reference center is not that crucial in our case since we will use relative bench marks (quartiles of the center specific excess probabilities).

With those choices in mind we have developed a number of modeling options below. We studied in detail the fixed effect outcome regression, random effects outcome regression and a doubly robust propensity score method. We focused here on models for the most important, most common as well as most challenging outcome types which are binary outcomes (success) and right censored survival type outcomes (time to event). As prototype cases we focused on outcome QCI 1111 (overall observed survival) and QCI 1232a (proportion of APR and Hartman procedures among patients who underwent radical surgical resection). Their theoretical properties were considered and – more importantly - their practical potential performance in the PROCARE setting was evaluated through simulation based on preliminary data made available on August 4, 2010. The simulations were deemed necessary because the presence of small centers (some with just a single patient entered) precludes an uncritical reliance on asymptotic properties of model parameter estimators and, a fortiori, of estimators of center-specific effects. Through a well chosen computational data generating mechanism, the simulations allow one to study the accuracy of a particular method in a particular setting before implementing it there.

The precise set-up of the simulations is given in more detail in the technical part of Appendix 1 (Chapter 9). Basically, they mimic the available database and first generate a random center choice in function of baseline characteristics based on a propensity score. Next, from the chosen center a random outcome is generated for the patient based on the outcome regression model. It is thereby assumed that center effects are themselves randomly distributed with some variation over the various centers in the database. Because the propensity scores are fitted on the original data, they reflect also the variation in center size seen in the database.

After fitting the various models, we display when possible both the estimated center-specific effects and population averaged center effects for the different centers in our preliminary database. Based on the repeated simulations we gained insight in the variation of the estimators as they vary from simulated dataset to simulated dataset. We were concerned specifically with bias, precision and coverage of confidence intervals. We further considered center-specific risks and population-averaged risks estimated over all centers.

In summary, for the 2 QCIs considered, with regard to the center-specific effects, which are not estimated by the standard propensity score methods, we saw no reason to distrust estimated center effects with confidence intervals for the fixed effects models, but could benefit for some centers substantially from the tighter confidence intervals around the random effects estimates when the model is correct. It was found that for population averaged effects, a comparison with the propensity score method results, which do not rely on the outcome regression model being correct, may also be prudent and worthwhile. In many cases the same qualitative conclusions should result from the different evaluations. If and when they do differ a more in depth examination will be required in the specific setting.

Finally, results under a misspecified random effects model shown in the technical Chapter 9 of Appendix 1 were rather encouraging in this instance and largely followed the lines above. For right censored survival data with necessary focus on 2 year survival in the preliminary database, results are more tentative due to few events in a sizeable number of centers. When 3 year survival becomes available in an updated dataset, we will be able to draw more firm conclusions for that setting.

In summary, results were encouraging for all methods on the 2 selected QCIs explored at this stage. We therefore carried all 3 methods forward to the next stage, where their performance would be judged on a much wider range of QCIs and data structures. As it turned out (see section 3.2.3) for the full set of QCIs which may include some that are less balanced or have substantially fewer eligible patients, only fixed effects models with Firth bias correction [13] produced consistent results in our set-up. This correction has been shown to be equivalent in the case of logistic regression to maximizing a penalized likelihood, where the penalization follows from a noninformative prior on the distribution of the true center effects. Specifically, `Jeffrey's prior' is involved here and leads to some mild shrinkage towards the center since this prior is symmetrical but allows generously for outlying effects. Because of shrinkage towards the mean, small centers which achieved a particular QCI in 100% of their registered patients, will show up with a non zero estimated risk of failing the QCI. This may look unfortunate at first sight but is sensible as the true risk of ever failing the QCI would not be absolutely zero. A similar phenomenon emerges more drastically from the more standard hierarchical models. It has nevertheless been accepted by the field and is used in the quality of care literature and evaluation of cardiac surgeons for instance.

Center effects estimated through the Firth corrected fixed effects method hence tended to show a reasonable spread over the centers, unlike hierarchical models which sometimes shrunk away any signal of deviations between centers. Keeping further in mind that 1) our goal is individual feedback for self improvement, and hence sensitivity rather than the specificity is for the detection of unusual center effects most important to protect in order not to miss centers that should be encouraged to examine their performance and 2) the difficulty in justifying a manageable hierarchical model, we focus on the Firth corrected fixed effects models in our present application.

We note at this point that our methods are straightforwardly adapted to a situation where one wishes to change the balance between sensitivity and specificity, for instance, if public disclosure were anticipated.

2.1.6 Scope of analysis

At this point it is worth stressing that before embarking on more complex modeling, descriptive statistics on outcomes, centers, and prognostic factors are important to help define the scope of analysis. Due regard is to be given to missing data at this level. While we are not planning to elaborate on the standard approach to this in any detail here, we simply point to some more important features to be examined in our setting.

For key survival outcomes, examination of the distribution of follow-up time in the dataset and over the centers, together with the observed numbers of events will give an indication of the amount of information in the dataset and each of the centers. It will for instance reveal whether 5-year survival chances are estimable with any degree of confidence, given the extent of follow-up. If updated yearly, such measures per yearly epoch may also yield a helpful description of the center progress over time in response to the monitoring and feedback. Further for this outcome type, it is important to consider whether censoring is or appears to be non-informative, possibly conditional on certain factors, before embarking on any analysis. If censoring is related to observed covariates, conditioning on those factors will be necessary in (cause-specific) survival models to avoid censoring bias. Alternatively, marginal survival models can be fitted in combination with methods for dependent censoring which involve these covariates [14]. Depending on the event (cause-specific or not) Kaplan-Meier Survival curves or the cause-specific cumulative incidence curves will non-parametrically describe the proportion of patients avoiding specific events over time.

A similar basic description of other QCIs is warranted: tables for discrete (binary) variables, box plots, and summary statistics for continuous outcomes and counts.

Regarding the centers, a first descriptive analysis should shed light on the variation in center size and the percentage of very small centers for which negligible information may be available. Secondly it will be important to recognize whether centers differ in amount of follow-up time (and therefore the censoring distribution) as well as more general completeness (missing data) over the centers. Finally, especially for sizeable centers, a brief inspection of covariates and correlation between covariates can help reveal whether some forms of center-specific characteristics, suggest special selection or measurement error and could be further examined. Detailed data quality control and a study of possibly systematic selective patient recording lies however in the hands of PROCARE and the Belgian Cancer Register who, unlike ourselves, have access to important background data in this regard (such as what percentage of its patients the center actually registered in the PROCARE database, and how the profile of its registered patients differs from that of those patients it did not register). It would certainly be interesting to examine the association between a center's degree of participation and reported QCI levels achieved. It would further be interesting to learn more directly about the reasons for not participating fully. This is beyond the scope of the current project and we will hence proceed with methods assuming we are dealing with a relevant sample of the observed patient population over the given treatment centers.

Finally, we will examine the distribution of patient characteristics observed in the database and over the centers. Again, missing data patterns, measures of location and variation plus correlation between and among QCIs as well as their prognostic factors could vary substantially between centers. This will reveal, among other things, the importance of adjusting for specific characteristics in the patient mix. If there turns out to be little or no overlap however, the adjustment for those covariates based on a general model fit may no longer be meaningful [15].

Summary points

- **Frequentist methods have the advantage of being relatively straightforward to implement in routine practice, but tend to rely on large sample properties unless resampling techniques are used. Bayesian methods automatically handle small samples, but then rely on parametric assumptions instead. Given our goal and the setting in which the methods will have to be operationalized, we opted here for the frequentist approach.**
- **A review of different techniques for risk-adjustment of binary and right-censored patient-level QCI measures is presented, considering fixed effects outcome regression, random effects outcome regression, doubly robust propensity score methods and instrumental variable methods. These four techniques are all considered within the causal framework in which we aim at estimating the effect of choice of center of care on the outcome (QCI).**
- **It was decided not to pursue the instrumental variables approach since the identified instrumental variables for this setting (distance and region/location) will not be available in the PROCARE database and preliminary results showed that the presence of many centers results in very imprecise estimated effects.**
- **A simulation exercise emulating 2 prototype QCIs showed that no single technique performed uniformly better on those data. We therefore carried all three analyses approaches forward to the implementation stage. When fitted on the large set of QCIs most methods ran into problems, including the simple logistic regression with fixed center effects. This happened especially when small centers (e.g. with less than 5 patients) with few events entered the model. The problems were overcome by adopting fixed effects models with Firth's bias correction and restricting the scope of estimation to centers with at least 5 patients (others may be grouped into one overlapping center).**
- **To allow for stable results and avoid too much shrinkage of observed center differences in the current database, we have focused on fixed effects regression models with Firth correction on the sizeable centers (with at least 5 patients contributing to the register).**
- **Issues related to the lack of access to known confounders (e.g. socio-economic status) are discussed. The risk-adjustment analysis will necessarily be restricted to age and gender plus the baseline clinical patient-specific confounders available in the PROCARE database.**
- **Missing data problems have been discussed. Methods will be implemented under the missing at random assumption, while acknowledging that this assumption may well be violated in the volunteer register. Depending on the setting, a separate category is assigned for missing covariate values or multiple imputation techniques have been used. A limited exploration of the selective nature of the register has been performed through a comparison with data from the Belgian Cancer Register.**
- **We have opted to focus on center-specific measures of performance. We thus evaluate each center on the type of patient mix they typically care for**

2.2 LITERATURE SEARCH FOR PROGNOSTIC FACTORS OF RECTAL CANCER

2.2.1 Methodological approach

The aim of the PROCARE project is educational in the first place, i.e. individual centers receive feedback on the outcome of their rectal cancer patients as compared to all participating centers (the entire PROCARE database serves as the benchmark). A fair comparison is only possible when the center's results are adjusted for all variables that may affect a patient's outcome irrespective of the therapy or therapies administered.

To conduct an objective search of the literature a rigorous protocol was adhered to and the following databases were searched: Medline through PubMed, Embase and the Cochrane Central Register of Controlled Trials. Details of the search strategy protocol are presented in Appendix 2. Details concerning these papers are summarized in Appendix 2B, while the global results from multivariate analyses are presented in Appendix 2.

The following outcomes have been studied:

- Overall survival
- Local recurrence

2.2.2 Results

In total, 152 papers were included. The main prognostic factor for overall survival clearly is related to the stage at presentation. Patients with bowel obstruction, perforation, serosal invasion, or peritoneal metastasis fare worse. Gender does not seem to represent an independent prognostic factor, while the prognostic significance of age is variable among studies. Several studies have shown that socioeconomic deprivation represents an adverse prognostic factor for colorectal cancer survival [4, 5, 16]. A wide array of pathological prognostic variables, macroscopic as well as microscopic and molecular, was identified. A number of recent studies have identified hospital volume as a prognostic factor in rectal cancer ([17], [18], [19], [20], [21]).

Clinical and demographic variables with an impact on local recurrence include T stage, presence of liver metastasis, and gender. The impact of tumor location within the rectum on the risk of local recurrence is unclear at present, since some authors found a higher risk of local recurrence with low lying tumors ([22]), while others reported the opposite ([23]). Treatment-related factors influencing the risk of local recurrence include preoperative (chemo)radiation, performance of a TME ([24]), and performance of abdominoperineal resection ([25]). Among the pathological factors that may impact on local recurrence, the circumferential resection margin is clearly prominent ([26], [27]). Finally, anastomotic leakage was shown in some reports to be associated with a higher risk of local recurrence ([28], [29]). Several other reports, however, concluded that anastomotic leaks don't have an impact on local recurrence rate ([30], [31], [32], [33]).

There is very scarce literature on QCIs identified in the setting of PROCARE other than survival or local recurrence. Some specific factors are reported separately in appendix 2. Appendix 2 further tabulates relevant prognostic factors for each QCI based on published evidence and on expert opinion from the participating clinicians. The most important factors for outcome indicators such as local recurrence and (cancer specific) survival are patient-related (Age, Bacterial translocation to lymph nodes, BMI, Gender, Health and physical sub-scale of QLI, Insurance status, Marital status, Poor general condition/ Co-morbidity, Recurrence, Socioeconomic status, Venous tromboembolism), tumor-related (Bowel obstruction, Circulating tumor cells, Focal perforation, Local peritoneal involvement, Lymphangitis carcinomatosa, Serosal invasion, Surgical curability, Tumor regressing grading) or related to clinical factors (BMI, Distance from the anal verge, Liver metastasis, Recurrence, Stage, T-stage/ depth of invasion).

2.2.3 Discussion

Several limitations apply to the interpretation of the present systematic literature search. First, most papers concern small patient numbers treated with a myriad of different therapeutic approaches and include colon as well as rectal cancer patients. The number of rectal cancer patients is usually not specified or a (small) minority of the overall population. This is relevant since the biological behavior of (low) rectal cancer and the paramount importance of surgical techniques in achieving the desired outcome are quite different compared to colon cancer. As we identified only 23 studies on rectal cancer alone, studies on colorectal cancer were nevertheless included. Second, almost all data were the result of retrospective studies. Studies not including some form of multivariate analysis were excluded. This criterion was maintained in order to guarantee a minimal quality of included studies.

It is important to note that most papers study prognostic factors through joint regression models, which contain the patient-specific variables available. Whether a particular variable enters as a significant predictor into a joint model will highly depend on which other variables are further included in the model. Indeed, both the magnitude and even the indication of the true effect on outcome may change depending on which other factors are entered. For some sets of variables only one may need to be appropriately corrected for the prognostic value involved, i.e. they can act as each other's surrogate in this sense. This could imply that as soon as a variable is entered, the other variables no longer have anything to add. Which of them actually enters may then be a matter of chance. This complicates the definition and role of the prognostic factors for reporting purposes. Beyond the magnitude of its systematic effect in the joint model, there is also the issue of precision. Whether a particular factor (in a joint or univariate model) is significant or not, not only depends on the magnitude of its systematic effect, but also on the precision with which it is estimated and hence on the sample size and covariate distribution in the studied population. In this light, and the fact that current and future sets of available covariates may rarely overlap exactly with what is reported in the literature, we will report here first on any variable found to be a significant prognostic factor. In the more detailed report we will indicate in what combination of covariates it occurred with what weight.

Summary Points

- **The primary search identified 981 articles. From this list, 308 articles were selected for full-text evaluation leading to 152 articles included in the final assessment. From these articles, an extensive list of prognostic factors for overall survival was obtained as well as a less extensive list of prognostic factors for local recurrence, cancer-specific survival and post-operative complications. There is very scarce literature on prognostic factors for other QCIs identified in the setting of PROCARE.**
- **The literature search imposed restrictions in terms of study design and patient population. Since a mere 23 studies considered just rectal cancer patients, also studies on colon cancer patients were eligible for our selection.**
- **Most papers study prognostic factors through multivariate regression models, hence the direction and magnitude of effect of a specific prognostic factor on the outcome depends heavily on the other factors included in the model.**

2.3 DIMENSIONS OF CARE AND THEIR AGGREGATION

2.3.1 Goal

In this third section on dimensions of care and their aggregation we aim at investigating the following research questions:

- How to combine process and outcome indicators within the 8 domains of care as defined in phase 2 of the PROCARE project [2]?
- How to define a composite score based on these 8 domains of care? (General quality indicators, Diagnosis and staging, Neoadjuvant treatment, Surgery, Adjuvant treatment, Palliative treatment, Follow-up, Histopathologic examination)
- Are there indicators that can be omitted, or alternatively, which indicators are essential to evaluate the performance of centers?

One may examine the possibilities of developing indexes that capture the quality of care in the various rectal cancer centers, based on data available in the PROCARE database. We will construct quality indexes (QI) on three levels: QIs for evaluating respectively clinical outcome and process of care, one separate QI for each of the 8 domains and finally one global composite QI of the domain-specific QIs.

Figure 1 and Figure 2 provide an example of the relationships among the various QCIs and QIs. The outcome- and process specific QIs ($QI_{outcome}$ and $QI_{process}$) and domain-specific QIs (QI_1 – QI_8) are constructed from the QCIs which are provided by the PROCARE database (Appendix 4). A global QI may be constructed from the outcome and process QI, from the full set of original QIs (as was done for the outcome and process QI) or starting from the domain-specific QIs (as reflected in Figure 2).

Figure 1: Schematic presentation of the construction of an outcome and process-specific QI

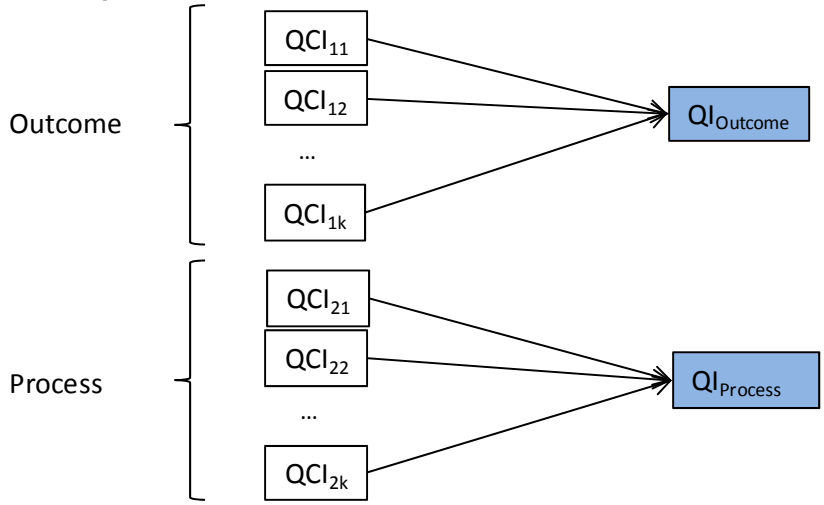
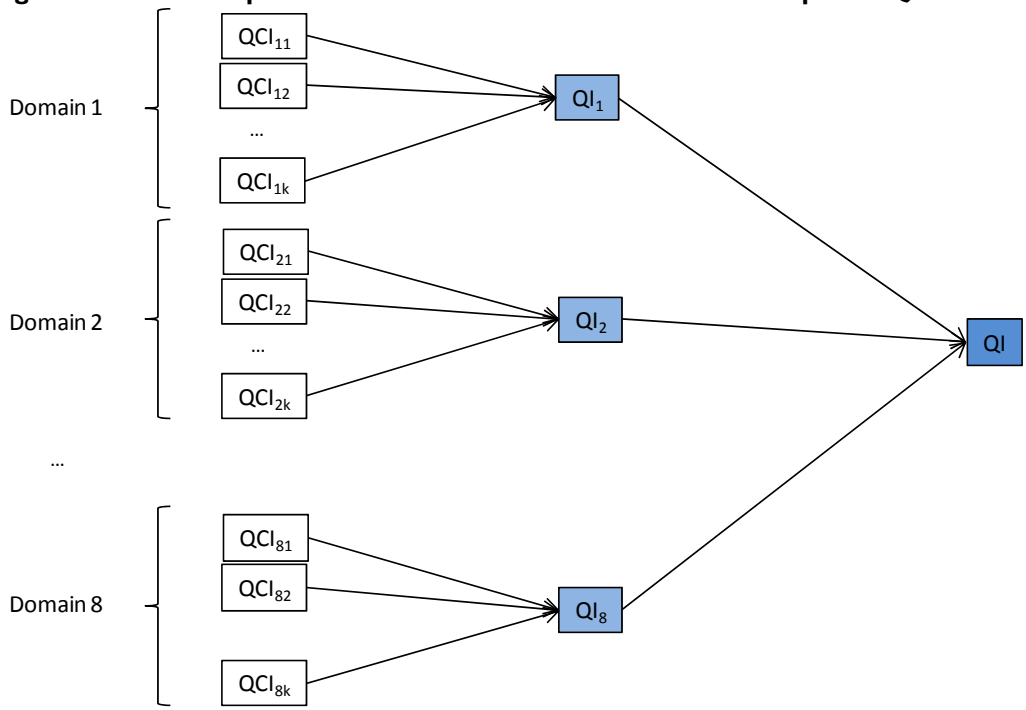


Figure 2: Schematic presentation of the construction of domain-specific QIs



The construction of each composite quality index will be guided by the following principles:

1. Involve indicators of clinical relevance/importance. For a hierarchy of clinical importance of the components of the composite quality index, we will rely on consensus in the literature and the judgment of our clinical team members and advisors.
2. Involve indicators which allow for meaningful discrimination due to substantial variation between centers, consider dropping those who do not.
3. Involve indicators for which we have or can obtain reliable data in the database, consider dropping those who do not. We seek to minimize undue influence of missing data and measurement error.
4. Involve indicators that empirically show a strong association with the underlying constructs, different dimensions covered by the QCIs as revealed by statistical ('factor') analysis.

As we consider building composite QIs based on subsets of the current QCIs, we will evaluate whether they are – or could be – close to as informative as scores built on the entire set. In doing so, we will acknowledge that if fewer QCIs need to be measured and fewer variables recorded, this may result in a lower percentage of missing data and more accurate measurements for the data recorded.

2.3.2 Key approach

In this chapter we focus on the composition of domain-specific QIs. The composition of an outcome and process QI as well as the global composite QI follow the same ideas.

We plan to proceed along the lines of previous research on composite outcomes used in Cardiac surgery in US [34] and will consider and compare simple or weighted averaging, with "all or none" scoring to combine QCIs within a domain in view of our four guiding principles.

Sum score methods make use of risk-adjusted center effects as we obtained from a (Firth-corrected) fixed-effect (Cox, logistic or linear) regression of the QCI on the appropriate risk factors and a factor for the center choice as described in detail in 2.1. The "all-or-none" scoring method starts from the raw, unadjusted QCIs and risk-adjustment is performed on the obtained "all or none" score.

An alternative for the stringent "all or none" score is using the number of targets reached (relative to the number that should be reached in theory).

We are well aware of the risk of data-snooping, formally or otherwise, and the involvement of clinicians and researchers (with a possible conflict of interest) in steering the choice of weights in the composite index. Before averaging or combining center summaries within domains, however, we suggest rescaling the QCI measures so that not only the directionality matches, but similar values indicate similar levels of lack of treatment or outcome quality. Practically, rescaling will be performed by standardization of the center effects with the between-center variation as estimated by a hierarchical regression model with at the patient-level the appropriate risk factors for the QCI in question. Note that this rescaling does not apply to the "all or none" scoring as this is one binary measure.

We acknowledge that in the construction of any composite score there may exist great diversity in preferences amongst different policy makers and stakeholders [14]. Ultimately, one could allow for different forms of calibration and present a simple software tool that allows certain users to produce their own set of weights and examine their relative score on that basis. This could offer "local policy makers the freedom to set their own priorities and to seek out improvements along dimensions of performance where gains are most readily secured" [35]. Caution is required with such application, however, since it opens the door to data snooping and to setting post hoc priorities which serve one's own (conflict of) interest.

2.3.3 Criteria for indexes

2.3.3.1 Definition

A quality index (QI) as used here is a composite indicator for a certain underlying construct, in this case, the quality of medical care in the various centers for rectal cancer. We distinguish two main types of indexes, those which are calculated as the linear (weighted) combination of a set of existing variables and those obtained from an "all or none" scoring method.

The first type of indexes is obtained by computing a sum score, i.e. combining some or all QCIs within a domain by a linear combination of the following form:

$$QI_i = \alpha_{i1} f_1(QCI_{i1}) + \alpha_{i2} f_2(QCI_{i2}) + \ldots + \alpha_{ik} f_k(QCI_{ik})$$

where QI_i is the quality index for domain i, $f_j(QCI_{ij})$ (j = 1, ..., k) the center effects for the j^{th} QCI, and α_j (j = 1, …, k) the weight given to the j^{th} QCI in the construction of the QI.[1]

The center effects for the k^{th} QCI are risk-adjusted center effects as obtained from one of the methods described in section 3.1. As we are interested in evaluating the center's performance on the set of patients actually treated in that center, we will focus on center-specific (in contrast to population-averaged) effects for computing QIs. Therefore the propensity score method, which currently only allows estimation of population-averaged center effects, is not considered further here. Since (Firth-corrected) fixed-effect regression models provide unbiased estimates of the center-specific effect, this in contrast to hierarchical regression models which are subject to shrinkage, we will focus on the former method for estimating risk-adjusted center effects. The set of risk factors to correct for will be determined separately per QCI.[2]

---

1 **Note** that no assumptions are made about the relations among the center effects ($f_j(QCI_{ij})$'s) and that the weights α_j (j = 1, …, k) are arbitrary in the sense that they are imposed upon the index rather than estimated during the construction of the index.

2 **Note** that centers with less than 5 patients eligible for a QCI will be discarded from the regression as no reliable estimates can be obtained for them. If needed for computation of the sum score, an appropriate imputation technique will be used to obtain values for these centers. The fact that several QCIs may jointly suffer from the same low sample size will complicate this exercise.
Note that the center effects will typically be obtained from (Firth-corrected) fixed-effect regression models. From this, one can either derive the risk-adjusted center-specific effects leading to the expected QCI in the center for the reference group (in terms of patient characteristics for which one adjusted). To

The different types of sum scores have different properties and underlying ideas. An important difference between them lies in the weights α_j (j = 1, …, k):

- Simple averaging: all selected QCIs are considered to be equally important (on their original scale) for the underlying construct and hence the weights are all equal (to 1/k if one wishes to keep working on the same scale as the original center effects).
- Weighted averaging: a different importance can be given to the different selected QCIs by giving each of them different weights. The weights are determined prior to construction and can be based on e.g. clinical importance of the QCIs.
- Composite opportunity model: provides a way of accounting for the fact that some patients may be ineligible for some measures, the weights j (j = 1, …, k) are the proportion of eligible patients for the respective QCIs. For our goals, this opportunity-based approach is not that interesting since the proportion of eligible patients for the QCIs does not have a specific clinical importance.
- Scaled combinations: center effects are inversely weighted by the amount of information captured in terms of discriminating ability (i.e. the inverse between-center variation as estimated from a hierarchical regression model), possibly in combination with a priori determined (clinical) importance of the QCIs used. This is a required step in domains with QCIs measured in different scales.
- In latent variable models (or item response theory [36]), multiple observed QCIs are assumed to be related to an underlying (unobserved) latent variable (e.g. quality of histopathological examination), the latter being the primary focus of interest. This type of model potentially allows quality to be estimated with high statistical efficiency by combining information from multiple observable measures into a single parameter. The relative weights for each observable indicator are determined iteratively from the model, obviating the need to make a priori weight assignments. These weights are rather abstract and therefore harder to interpret.

The clinical interpretation of these five sum scores can be rather different, and depends on the type of information captured within the domain. In practice, a combination of several of these methods may be appropriate to arrive at the final QI, e.g. scaling and weighting. After combining the QCIs, the resulting QI may be rescaled to allow for easier interpretation.

The second type of composite measure, "all or none" scoring, works differently. First, it is assessed for each patient whether he/she has reached all pre-specified targets for the QCIs within the domain. For binary QCIs the target is simply having the event (yes or no) while for continuous QCIs a target needs to be set by clinical experts, e.g. the time between first consultation and first treatment should not be longer than 30 days. Secondly, the "all or none" score per patient is set to 1 if the patient reached the targets for all the QCIs he/she is eligible for within the domain of interest and set to 0 otherwise. Finally, this "all or none" score is risk-adjusted in the same way as binary QCIs in section 3.1 The (rescaled) center effects obtained from this model represent the final composite score.

meaningfully combine these measures across QCIs one should refer to the same reference group for the different QCIs. The other option averages the predicted patient-specific outcomes over all patients within each center to arrive at the center effects and compares this to the predicted patient-specific outcomes we would observe if these patients were to be treated in the "average" center (excess outcome due to the center after adjusting for patient-specific confounders). This is likely more precise and can handle different adjustment models for different QCIs.

The choice between a (weighted) sum score or "all or none" score depends on the interpretation we wish to give to the composite measure. If missing any one is deemed (almost) as unfortunate as missing several QCIsan "all or none" scoring seems appropriate. On the other hand, if it matters how many QCIs (irrespective of which) one retains, a simple average may appear as the candidate summary (the number of targets reached relative to the number that should in theory be achieved). When different QCIs are considered to differ in importance, weights in the sum may accommodate this.

Compared to the simple averaging sum score measure in which all QCIs carry the same importance (averaging of the center effects per QCI), the "all or none" method has the advantage that it does not compensate when a center performs excellent on one QCI but poorly on another QCI (within the same domain). The fact that this "all or none" scoring is rather stringent can be seen as a disadvantage, but from a diagnostic perspective this scoring is very straightforward. An important issue that remains to be examined is whether the "all or none" score manages to discriminate between centers, which is unlikely if there are QCIs with very rare events but rather likely if all QCIs within a domain have a high prevalence.

2.3.3.2 Criteria for item selection

As the construction of QIs is more an art than a science, several decisions need to be made based on less scientific judgments. The choices are interrelated and will affect each other.

A very important decision to make is which items (QCIs) to include in the construction of the QIs. This decision will be based on the guiding principles as described in Section 0.

Theoretical and clinical importance

The most important criterion is whether the QCI is theoretically and/or clinically relevant for the underlying construct. In this case this seems unproblematic as the indicators in the PROCARE database were chosen for their clinical relevance [2]. However, some QCIs might be more important than others.

The clinical importance of each of the QCIs was assessed by performing a small survey among a group of clinical experts (Appendix 5). This survey helps to identify QCIs to be used in the construction of a valid QI. The validity of the QI refers to the extent that it measures what it is supposed to measure and can be evaluated in different ways:

- Face validity: Does the QI "looks to" measure what it is supposed to? This can be determined by examining whether the center effects that make up the QI are theoretically and/or empirically related to the underlying characteristic, and by evaluating whether they were combined sensibly in the composite index. For indexes this usually is the most important type of validity checking.
- Content validity: A more statistically rigorous extension of face validity, in which experts evaluate whether all aspects of the measured characteristic are represented in the measure and judge whether each item is useful for the underlying construct. For a proper content validity analysis a survey should be held among international experts in the field who should be asked to evaluate each of the QCIs and QIs on their validity and on the dimensions they cover.

To statistically examine the clinical importance of a QCI, overall survival (ultimately death from rectal cancer – data permitting) will be used as the 'gold standard'. The association between death from any cause (death from rectal cancer) and the QCIs will primarily be evaluated at the patient-level, using risk-adjusted Cox proportional hazards.

Ability to discriminate between centers

Another selection criterion considers whether QCIs discriminate between excellent and poorly performing centers, or less informed, to what extent each of the QCIs captures existing differences between the centers.

To examine whether a QCI shows at least variation between centers, caterpillar plots of the center-specific patient-mix-adjusted average effect of the QCI (called excess expectation or probability) with appropriate 95% confidence intervals (95% CI) were used. A descriptive table with statistical summary measures (minimum, 25%-tile, median, 75%-tile, maximum and interquartile range) of the distribution of the center-specific averages facilitated the evaluation.

To formally detect variation between centers beyond sampling variation in the QCIs, one may further test whether the variance of the center-level effect in the hierarchical regression model differs significantly from 0. To this end, a (pseudo-) likelihood ratio test with a mixture of 0 and 1 degrees of freedom was used.

It is to be acknowledged that a QCI that does not discriminate between excellent and poorly performing centers in the dataset, may still be an important QCI. If all centers score excellent – or indeed poorly - on such an important quality criterion, this should be recognized.

Data quality

Data quality is another important consideration in the selection of QCIs to be included in the QI. A high number of missing data for a QCI or its unreliable reporting disqualifies the QCI from being included in the QI. The frequency of missing data should easily be established from the final PROCARE dataset[3], the (presumed) reliability of the QCIs were included in the survey we wish to conduct among a group of clinical experts (Appendix 5). More formal measures of reliability of (the QCIs in) the PROCARE database (e.g. the percentage of patients actually registered per center) should be examined at the Belgian Cancer Registry (BCR).

Set of eligible partners

As seen from the denominators defined for the QCIs (Appendix 4), not all QCIs apply to all the patients registered in the PROCARE database. This has different consequences depending on the composite scoring measure that will be used. For QIs computed as the linear combination of center effects, it happens that for some QCIs a center had no eligible patients and hence no estimated center effect. In this case we suggest imputing a value for the missing center effect. One may for instance use a regression-based imputation method, allowing prediction based on the center's available QCIs. For the "all or none" method, a patient will obtain an "all or none" score by considering only the QCIs he/she is eligible for. As long as those are all met, the score stays 1.

Although the number of eligible patients per QCI is not directly visible in the QIs that result, QCIs applying to a very restrictive set of patients are not very interesting for inclusion in a QI that should be generally interpretable.

Concurrent validity

Concurrent validity checks can be used to select just a few QCIs that are close to as informative (with regard to explaining the underlying construct) as a larger set of QCIs within the same domain.

Here we examine associations among the various candidates for a QI though principal components analysis (PCA) and an exploratory factor analysis (or adapted latent variables model [36]). The latter was used to determine how strong each of the QCIs load on the underlying dimensions recognized by the PCA. High loading QCIs are prime candidates for inclusion in the indexes.

---

3 Based on the PROCARE database we currently have at our disposal, missingness is not very easy to assess.

To avoid issues with differential illegibility, these analyses were performed at the center level, using unadjusted effects for process QCIs and risk-adjusted effects for outcome QCIs. The number of factors to retain is based in part on the proportion of variance they account for and on the incremental variance accounted for between two subsequent factors. For the clinical interpretation of the obtained factors, clinical experts were consulted.

2.3.3.3 *Proposed strategy*

To finally select QCIs based on the four guiding principles described above we consider two item selection strategies. Decision trees describing these two approaches are presented in Figure 3 and Figure 4.

Figure 3: Decision tree for the first (clinical-statistical approach) suggested item selection strategy

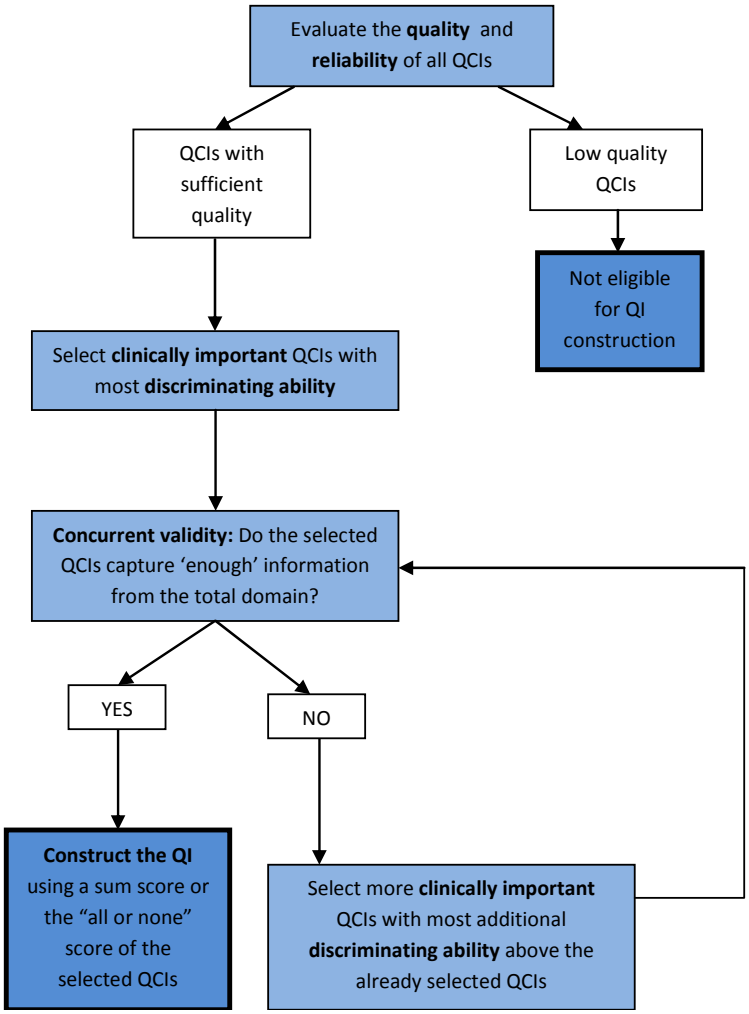
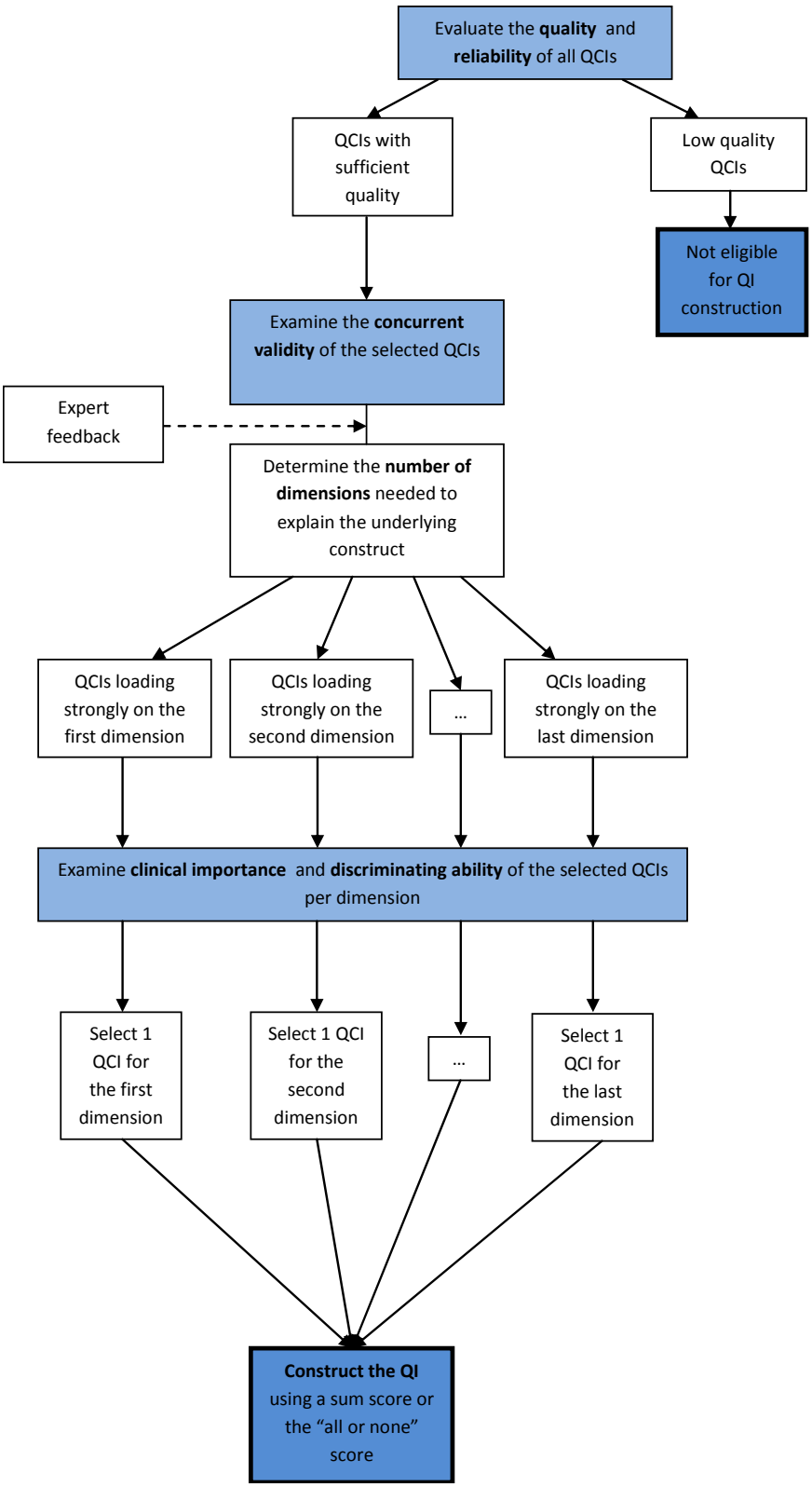


Figure 4: Decision tree for the second (statistical -clinical approach) suggested item selection strategy



The first step is common to both strategies. To allow for a fair and credible assessment, one should work with good-quality, reliable QCIs. Classification as a good-quality QCI follows from criteria to be imposed on the results of the second question in the survey to be conducted among clinical experts (Appendix 5) if in addition enough centers have at least 5 eligible patients for the QCI and the percentage of missing observations is limited. In an ideal world, this step is not needed, but in practice reliability and data quality must be monitored.

After this starting point, the two selection approaches diverge in the order in which they invoke arguments of clinical and statistical importance for selection.

First clinical-then-statistical selection

The first strategy starts from QCIs deemed clinically most important. This evaluation can follow from the QCIs mean score on the first question of the expert clinician survey (Appendix 5) and its association with 'overall' survival. Such candidate QCI will be selected if it further discriminates between centers, i.e. if the QCI shows important variation between centers beyond sampling variation and after adjusting for case mix. This is assessed following a likelihood ratio test for the random-effects variance, controlling for confounding variables.

In the next step, one evaluates through an PCA followed by an appropriate latent-variable technique (e.g. factor analysis for normally distributed QCIs) to what extent selected QCIs capture the full variation in (good quality) QCIs within their domain. The QCIs selected in step 2 which load high on the latent dimensions retained, will contribute to the QI constructed using a (weighted) sum score or an "all or none" score.

First statistical-then-clinical selection

The second approach starts from a factor analysis (or adapted latent variables technique) on the good quality QCIs. Based on the percentage of variation explained (through eigenvalues) an appropriate number of dimensions are retained per domain. Ideally these dimensions enjoy a clear clinical interpretation recognized by clinical experts.

Per dimension, highest loading QCIs in the factor analysis are evaluated next in terms of their clinical importance and ability to discriminate between centers. The clinical importance is assessed as above, from the QCIs mean score on the first question of the expert clinician survey (Appendix 5) and its association with 'overall' survival. Similarly, discriminating ability is assessed following a likelihood ratio test for the random-effects variance, controlling for confounding variables. The QCI that scores best on these criteria is then selected for construction of the QI.

Comparison of the two strategies

The first strategy favors the clinical importance of the QCIs (evaluation by experts, discriminatory power, predictive validity), while the second approach gives more weight to QCIs' ability to capture the variation within the domain QCI-set. The two approaches therefore may result in the selection of different QCIs for the final QIs.

In the second approach the clinical importance of the QCIs is but a secondary criterion for QCI selection, and clinically important QCIs may not be selected as just a single QCI remains per dimension. The second approach may further allow for clinically less important QCIs to be selected when none of the most important QCIs load highly on one or more of the dominant dimensions. When the first approach recognizes that the clinically most important QCIs fail to sufficiently capture the overall variance of the set, this may lead to the selection of additional QCIs for the final QI.

In the first approach, all QCIs deemed clinically important are considered for entering the QI (adding perhaps also a few less important ones). In the second approach the number of QCIs included in the QI depends on the number of dimensions that one distinguished in the domain since one QCI per dimension will be selected, preferably the clinically most important one loading high on that dimension.

2.3.4 Other issues in the construction of quality indicators and quality indexes

2.3.4.1 Attribution of cases

Clear rules need to be established for which center each of the patients in the PROCARE database is attributed to. As a standard, patients are attributed to the center in which they underwent surgery (and if no surgery was performed to the center where the pathology report was filled out or if this did not happen where chemo or radiotherapy was performed). This way of attributing patients to centers implies that surgical centers will be evaluated for the performance of e.g. the center where pathology was performed for a QI referring to the domain of histopathological examination. To avoid such 'misleading' evaluations for the domain-specific QIs, it should be considered attributing patients to the center where the therapy/surgery/follow-up/histopathology took place, depending on the domain. This more detailed information about what was performed where is not available in the PROCARE database we have currently available, but could in theory be provided.

2.3.4.2 Relative versus absolute benchmarks

Essential to the construction of indexes are benchmarks against which to measure the performance of each of the centers. There are two sources for benchmarks. One source derives benchmarks from the QCI distribution in the PROCARE database itself based on e.g. the 95% CI for the estimated (case-mix adjusted) center performance with the 25% and 75% percentiles of the distribution of these effects. This is suggested by the team of PROCARE experts for identification of poor and excellent performing centers, it has been our focus. The advantage of this approach is that these benchmarks are easy to obtain. The drawback is that for each indicator a similar prevalence of good and bad performers is generated by the relative QIs. These may be difficult to interpret in real or absolute terms and need not be comparable over time.

The alternative approach uses external benchmarks. These may be quality standards set by professional organizations or an expert panel, or may be performance criteria derived from the literature. Although it is much more difficult to obtain such external benchmarks, and it might even prove impossible to obtain them for all QCIs, this approach is to be preferred over the former one. By evaluating the centers against external benchmarks one achieves absolute QIs, which could well show that, despite variation between centers, most centers perform well and above the international standard, … or reveal the opposite and hence a shared opportunity for improvement. The benchmarks can take various forms, depending on the nature of the QCI and the available information in the literature. They may involve indicators of centrality (mean, median,…) or dispersion (standard deviation, interpercentile range,) or of upper and/or lower thresholds that indicate what performance is considered excellent, good and/or weak. Availability of benchmarks will affect the normalization of indicators and the production of the QIs.

2.3.4.3 Comparability across centers

A good QI must allow comparing centers in terms of the quality of care provided. This implies 1) that centers of equal quality should obtain identical scores, 2) that centers of different quality should get distinguishable scores, and 3) a one-to-one relation between the quality of the center and its score on the QI. Given that quality of care is multidimensional, it is unavoidable that different combinations of the QCIs will lead to identical QI scores. One needs to assure that identical scores obtained this way indeed represent an identical overall quality as measured by the QI. The determination of the weights (when the QI is computed as a linear combination of center effects) for each QCI will be crucial to achieve this comparability.

2.3.4.4 *Level and/or spread*

A related issue is whether the QIs should only capture the overall quality of care or also the distribution of quality of care over the patients. In the former, measure of central tendency and counts will be used in the construction of the indexes, while in the latter also the spread will have to be taken into account. For instance, a measure of the level of quality might make use of the mean, median, or mode of the pre-existing variables, where a spread measure would make use of the standard deviation, range, interquartile range, etc. Of course level and spread may be related: poorly or excellently performing centers may tend to have less spread among patients than moderately performing ones. With binary indicators spread and level are inherently linked, leaving only the relevant measure for non-binary QCIs.

2.3.4.5 *Comparability over time*

As the purpose of these QI includes tracking changes in the quality of care of centers over time, the QI should remain comparable over time. If two centers at two different moments in time (or even a single center) score identical on all QCIs they should score identical on the QI, and changes in the QCIs should be reflected in the QI. Practically, this means that not only the included QCIs should remain stable over time, but also that the method of combination and normalization of the QCIs should remain stable.

Summary Points

- **A quality index (QI) was constructed for the outcome QCIs, for the process QCIs, for each domain separately as well as a global QI.**
- **The QI may be a single QCI. Alternatively a set of selected QCIs was retained and summarized in either a (weighted) sum score (linear combination) or an "all or none" score for the selected set of QCIs.**
- **QCIs are selected based on properties of their center-averaged excess probabilities, this is their expected outcome beyond what is expected for their patients in the "average" center. For outcome QCIs this excess probability was risk-adjusted for patient-specific confounders of center choice and QCI, for process QCIs the excess probability is not adjusted.**
- **Selection of QCIs for the QI is based on four guiding principles: data quality, concurrent validity, clinical importance and ability to discriminate between centers. The order in which they are applied (order of importance) is first statistical, then clinical.**
- **How patients are attributed to centers may depend on the type of Q(C)I and is an important point of discussion.**
- **For labeling center performance as poor or excellent, ideally an objectively determined external bench mark is used on which a group of experts will agree. This can then – for instance - be used to judge evolutions over time. In practice such bench marks are not (yet) available.**
- **As suggested by PROCARE, benchmarking of centers may be based on the internally observed 25th and 75th percentiles of estimated center performances. Centers with confidence limits of the QI completely below the 25th percentile are then flagged with lesser results and centers with confidence limits of the QI completely above the 75th percentile are labeled with relatively higher performance. Properties such as sensitivity and specificity of such decision rule are yet to be determined in this context.**

2.4 REVIEW OF STATISTICAL METHODS FOR PROVIDER PROFILING

2.4.1 Goals

This section aims to review methods for provider profiling and builds on the three previous sections (methods for patient mix adjustment, literature review on prognostic variables, the construction of indexes) to arrive at a motivated proposal for benchmarking centers contributing to PROCARE. To this end, we take several steps and ultimately consider:

1. the indexes in terms of which centers will primarily be profiled,
2. the set of patient-specific covariates for which we will adjust the indexes (and the method applied),
3. the bench-marks chosen for the different indexes,
4. individualized feedback on these indexes. Which graphical presentations will be constructed? Why do we not wish to formally assign ranks to the different centers?

2.4.2 Methods for provider profiling in the literature

Provider profiling has become an active area of research. A web of science search on 'statistical methods for provider profiling' yields 16 papers of which some focus on very specific application areas not directly relevant to the current study. Other papers are more general and were reviewed in more detail for our purposes here. Methods and results are not only published in the statistical literature [37, 38], but also in the clinical literature [34, 38-41], the media[4] and even Consumer Reports in the US [as reflected upon in the New York Times, September, 2010]. The general approach of monitoring and profiling centers that deliver important services to the public is relatively new to the medical field. It has known more extensive use and experience in other areas, for instance education, where league tables have become popular as well as heavily criticized. In the clinical field, research on this topic was primarily generated and implemented in the domain of cardiac surgery. For an example in cancer, we can consider the work by the Dutch surgical colorectal audits [42], where practice is closely monitored and methods have been developed and applied to enable progress [43].

We start by reviewing the approach presented in a seminal 1997 paper 'Statistical Methods for Profiling Providers of Medical Care: Issues and Applications' by Normand et al. in the Journal of the American Statistical Association [37]. These authors recognize several steps in the profiling approach: 1) a risk-adjustment step, 2) the choice of an index and 3) benchmarking. They make a number of choices which overlap, but do not coincide with ours. This partial overlap is not unexpected and indeed anticipated in their discussion which states most relevantly:

'The methodologic issues confronting analysts in this area are not trivial, and simplistic "one-size-fits-all" approaches are not likely to work [44]. Major issues include data quality, detail, and availability; choice of performance measures, formulation of statistical analytic strategies; and development of approaches to reporting and interpreting the results of the profiling analyses' … 'Because profiling methods can serve a number of purposes, the choice of performance measures and analytic strategy will have to be customized. Data availability may also play a major role. For example if a reference standard such as a national guideline, is available, then it is reasonable to evaluate providers using an absolute cut-off level of performance. Unfortunately, such reference standards are often not available, and comparisons of providers will need to be made using relative measures of performance.'

4 In France, Le Point yearly publishes "Le palmarès des hôpitaux", ranking all French public hospitals.

Results based on [37] have served to broadcast quality measures for medical (cardiac surgery) centers in the media and have occasionally lead to sanctions. Correspondingly, a quite conservative standpoint has been taken in this approach, keen to reduce the type I error: erroneously labeling a center as underperforming. The consequences of that error are indeed grave and possibly irreversible. In our setting we are working in view of confidential individualized feedback towards the centers with the goal to help centers recognize where their own weak points lie and point them in the direction of further improvements. In our setting, therefore, relatively more weight is given to avoiding type II errors, i.e. failure to diagnose an area of true underperformance. A type II error prevents the center from recognizing an opportunity for improvement and thus deprives patients from the better treatment that might result.

In step 1, the risk adjustment step, the authors start from patient-specific data and fit two models: [16]

- The first model adjusts for patient-specific (baseline) factors only. To this end, patient characteristics are summarized in a severity index drawn from 34 baseline characteristics of the patients. The authors allow the effects of patient-specific factors to change over the hospitals. While our primary analysis is scheduled to adjust for patient-specific factors, interactions between center and patient-specific factors are not foreseen until a possibly tertiary analysis designed to explain any differences seen between centers in the primary analysis.
- The second model additionally adjusts for hospital-specific covariates, specifically: for rural (vs. urban), non-academic (vs. academic) centers, and the number of beds (categorized as: <= 100 small, 101-299 medium, >300 large).

This risk-adjustment is achieved here through hierarchical logistic regression. As is well known, and we have described and illustrated in Appendix 1 (Chapter 9, section 2.1.3) to section 2.1.5, such models tend to shrink estimates towards the average of all centers and hence produce conservative estimates for both the exceptionally weak and exceptionally good performers.[5]

In step 2, for the choice of an index, the authors chose to focus on a single performance index which is easily interpreted and evaluation is based on the probability of (30 day) mortality/survival adjusted for patient (and ultimately also hospital) characteristics. [16]

For step 3, benchmarking, the following specific measures are examined:

- An indicator of whether 30-day mortality exceeds 1.5 times the median of its distribution over the centers.
- The difference between the center-specific and center-averaged risk for the center - with its specific patient mix. This difference is compared with the following cut-off value: median difference plus 1.5 times the interquartile range of this difference. Median and quartiles are taken over the random effects distribution of the center effects. 1.5 times the median mortality rate is seen as a meaningful cut point for weak performance. Using the Bayesian paradigm they calculate each center's chance of crossing the cut point.
- The 'more standard Z-score', i.e. the center's observed minus expected risk divided by its standard error with 1.645 as cut point for relatively poor performance ('top 5 %'). The authors consider this relative measure of performance undesirable since a predetermined fraction of centers is expected to cross the cut point by design. That is true, even in the absence of between-center variation in true performance, purely on the basis of within-center sampling variation.

5 Note: to avoid contamination between center effect and patient effect on the outcome in the primary analysis, it is important that the patient-specific characteristics for which one controls are not themselves influenced by the centers. For certain measures, such as P-staging, this would be an issue.

The authors advocate the use of several measures of performance with an inspection of consistency across several measures. When a particular decision rule has been identified to label centers as (especially) poor or excellent it becomes important to evaluate the sensitivity and specificity of the ensuing diagnostic tool. Such examination is developed in Normand et al. [38].

In their recent reporting of 'online CABG report cards', Ferris and Torchiana [41] have labeled centers as performing significantly below (given one star) or above (three stars) the mean center performance based on 99% confidence intervals (99% CI). This resulted in 23 to 27% of the programs to be identified as outliers over the past 3 years.

In the Dutch 2009 report [45] one calculates percentiles of performance over the center distribution and compares the expected percentile based on the observed case mix with the observed percentile. One thus evaluates unusual performance aided by QQ-like plots.

Based on a survey, the team of PROCARE experts suggests to produce 95% CI for the estimated (case mix adjusted) center performance and to compare this with the 25% and 75% percentiles of the distribution of these effects over the centers for the confident identification of poor and excellent performance.

2.4.3 Visualizing variation between centers

2.4.3.1 Variation in case-mix

In the Dutch 2009 report [45], one depicts variation over the centers in case-mix through simple scatter plots of expected values of important prognostic factors. They produce, for instance, one plot of average ages over the centers, one plot for the percentage of ASA scores exceeding a threshold, … and one 'summary plot' showing how the expected 30-day mortality rate based on case-mix alone varies over the centers.

2.4.3.2 Variation in quality indexes

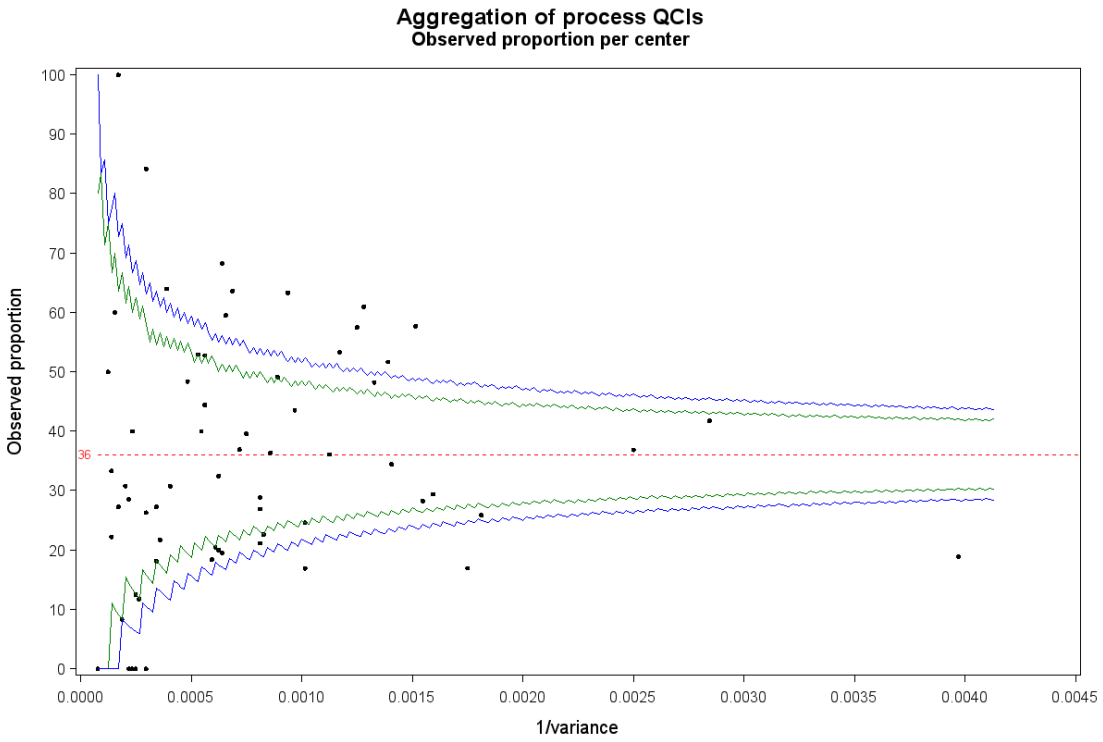
Revealing the variation in true quality index across the centers poses a special challenge. Since this true quality index is estimated with varying precision, one needs to 'subtract' sampling error from observed variation between estimated center effects. Indeed, variation in observed outcome is expected based on sampling variation alone, and tends to be larger for small centers than for larger centers. One therefore wishes to establish any evidence for variation in *true* center effects over and beyond the random sampling variation. The technique of hierarchical (random effects) models outlined in section 2.1 comes with a natural estimator of the variation in true center effects. A test can be derived to establish whether this variation is significantly different from zero as in van Houwelingen et al. 2002 [46]. If not, one lacks evidence of any systematic variation between centers and the profiling may end there. Once significant variation is established, however, it becomes important to engage in the evaluation of (relative) performance of the distinct centers. The final column in Table 30 and Table 31 refers to the significance (p-value) of such variation in the hierarchical model fit. If so desired one may add a measure of variation explained by centers along the lines of Efron [47] and van Houwelingen et al. [46]

Inspired by techniques developed for meta analysis, Spiegelhalter [48] developed the funnel plot which shows to what extent observed variation in estimated (possibly risk-adjusted) center effects lies within the boundaries of sampling variation - or not. To achieve this, the funnel plot starts from an estimated average effect. Assuming this effect is common to all centers, hence in the absence of variation in true effects between centers, one can calculate the sampling variation of the estimated center effect, and derive corresponding 95% and 99% confidence intervals in function of the center size. One may then test whether more than the expected number of estimated center effects lie beyond those boundaries. The typical plot shows estimated center effects either against the sample size or more directly against the information (1/variance) represented by each center. Plots in function of center size may not sufficiently protect center confidentiality. In some instances one may chose to present this type of graph also for feedback to centers, rather than the catterpillar plots.

Unlike caterpillar plots, they avoid a suggestion of implied ranking between the centers. Such suggestion is easily perceived from the order in which the estimated effects are shown in the caterpillar plots, even though confidence intervals emphasize substantial uncertainty on such ranks. The disadvantage of the funnel plot for our purpose is however that it does not focus on the relative bench marks that were chosen for our evaluation.

By way of example, Figure 5 shows a funnel plot for the all-or-none score indicating whether all selected quality indices in the process domain were achieved or not (more detail on this score is given in section 4.4). The central red line gives the percentage of all eligible patients who achieved this score, If all centers experience the same chance of achieving the score, this percentage produces an accurate measure of this probability of success. The dots on the plot show per center (of size 5 or larger) the percentage of patients for which this all-or-none score was actually achieved. The inverse of the variance of these estimates is the information measure on the x-axis. If all centers have indeed the same chance of achieving this score, then 5% (1%) of centers are expected to fall beyond the green (blue) inner (outer) lines. This amounts to respectively 3.4 and .68 expected centers of the 68 centers shown. In reality, respectively 26 and 18 centers were found beyond these bounds. The chance of this occurring if the probabilities of achieving the score were indeed identical are negligeable. Substantial evidence thus emerges of systematic variation between centers in achievement on the aggregate process score,

Figure 5 Funnel plot for the aggregate index of the process indicators.

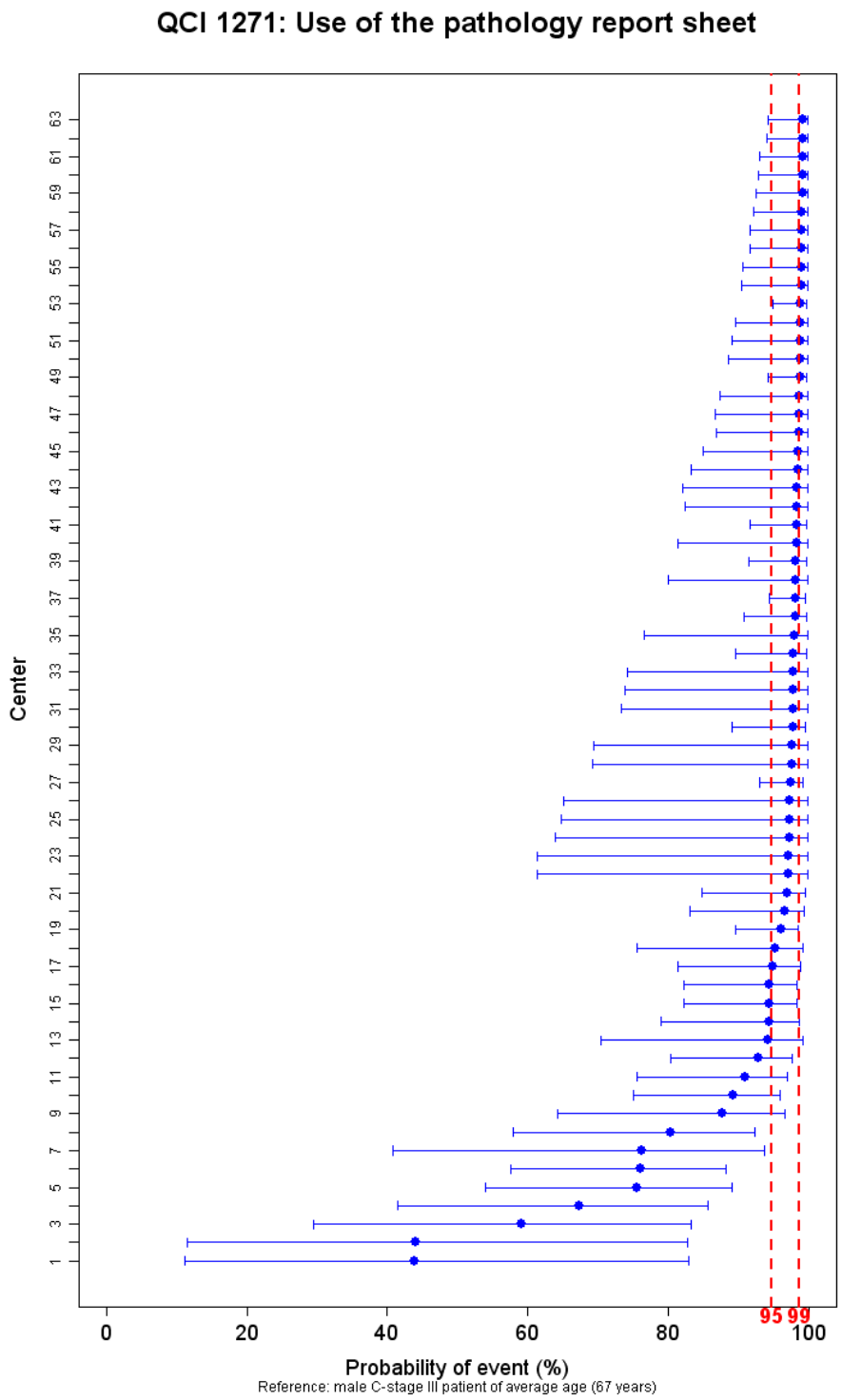


2.4.4 Graphical presentation of center results

Visualization of the estimated center effects with their precision using caterpillar plots was shown by Spiegelhalter [48, 49]. Caterpillar plots order centers by estimated effect size and plot them with their credibility/confidence intervals as illustrated for QCI 1271 (Use of the pathology report sheet) in Figure 6 below. Figure 6 shows an overview of case-mix adjusted center performances. Here, the adjustment was done for age, gender and cStage based on the preliminary PROCARE database. The expected center-specific outcome for the average aged male in C-stage III is shown. The red lines show the first and third quartile of the estimated excess probabilities as well as the median of these values over the center distribution. This helps evaluate centers in line with the boundaries identified as useful benchmarks by the PROCARE steering group who chose to label centers as performing poorly when the upper boundary of their 95% confidence interval stays below the 25% percentile. [6] A similar plot on the logit scale has the advantage that the 95% CI around the estimated center effects are symmetrical. This facilitates an understanding of their genesis and the estimation procedure. On the other hand, the plot on the original probability scale allows for an intuitive (clinical) interpretation of the center effects along with a subjective evaluation of their distance from the P25 and P75 boundaries. It was therefore preferred. Whether or not these P25 and P75 percentiles represent meaningful boundaries for poor and weak performance in absolute terms is indeed a question more easily judged by experts on this original probability scale. When performing this analysis for a whole range of QCIs, we chose to move to a relative measure, which we called excess expectation or excess probability: the difference between a center's expected outcome on his patient mix and the expected outcome for the same patient mix for the (logit based) average center (see Figure 6 below). We then also decided to focus on unadjusted measures for process quality indicators and adjusted measures for outcome indicators.

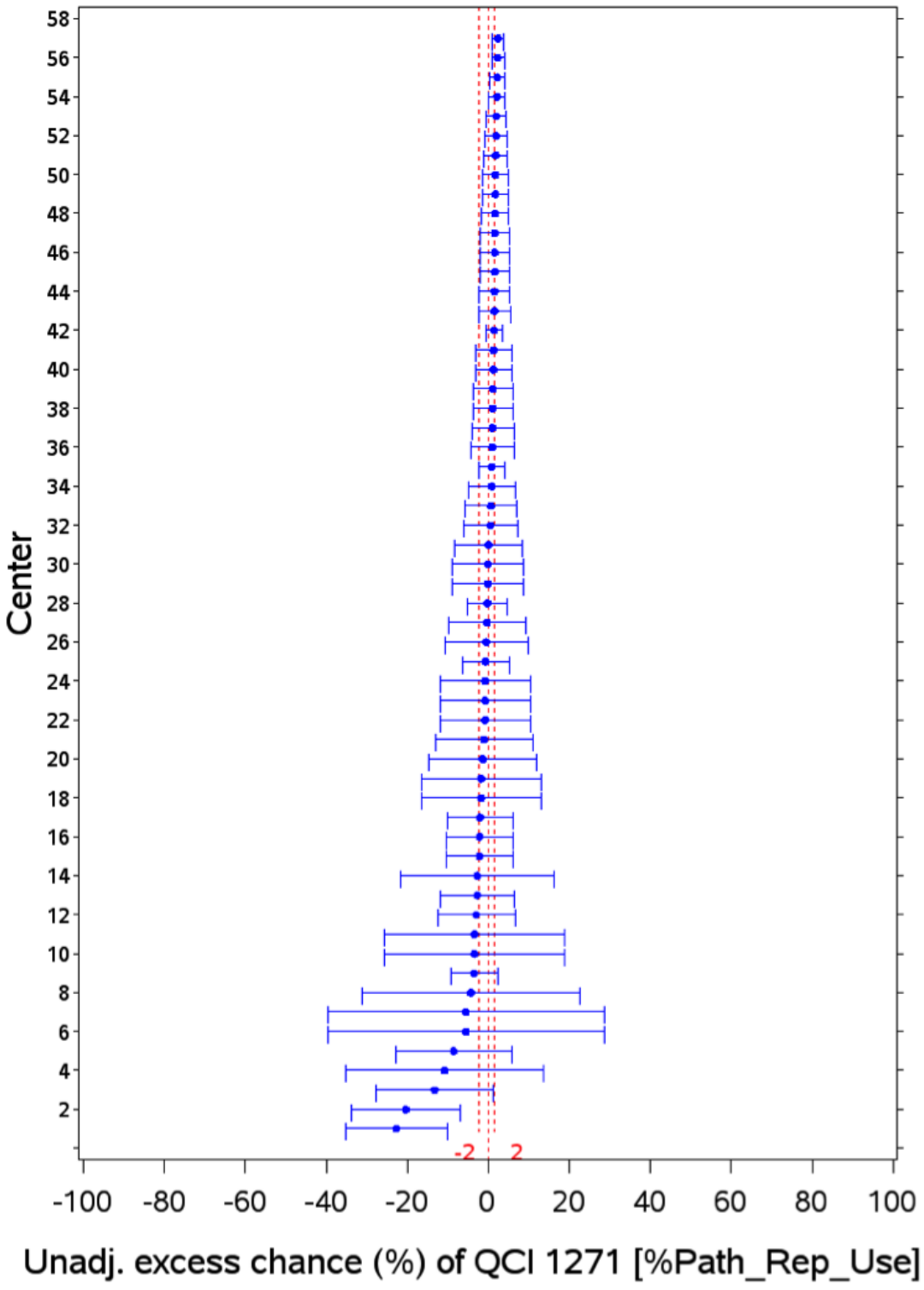
6 Note how the empirical quantiles are simple and transparent estimates of the corresponding quantiles of the true expected center effects. They are not necessarily optimal however. Since the estimated center effects suffer some imprecision their observed variation may generally be larger than that of the true center effect variation. This will be compensated to some degree in our case by the mild shrinkage achieved by the Firth correction. For a more comprehensive discussion on (Bayesian) optimal estimation of percentiles, see the work of Paddock and Louis, 2011 and Carlin and Louis, 2000,

Figure 6: Caterpillar plot for QCI 1271 [%Path_Rep_Use] on the natural scale.



The x-axis shows the estimated probability of use of the pathology report sheet after adjustment for the three baseline covariates gender, age and cStage. The probability is given for males, aged 67 in cStage III. Red lines show P25 and P75 of the estimated values.

Figure 7: Caterpillar plot of excess chances for QCI 1271 [%Path_Rep_Use]. The x-axis shows the estimated "excess" probability (beyond what is expected on the average center) of use of the pathology report sheet. Red lines show P25 and P75 of the estimated values.



2.5 PROPOSED APPROACH FOR PROCARE

2.5.1 Choice of index

This follows from the statistical-clinical approach outlined section 2.3 and we anticipate profiling in terms of each of the indexes developed there: one index for each of the 8 domains, one global index, one index for outcome and finally one for process.

2.5.2 Adjusting for patient specific factors

As argued in section 2.1 and established through the results of the literature review summarized in 2.2, one ideally adjusts for a whole range of patient-specific factors to account for case-mix. The list of prognostic factors may vary with the QCI considered. Especially for outcome QCIs (e.g. survival) several evidence-based prognostic variables have been identified, including some important ones which are either not present in the existing databases relevant to PROCARE or subject to denied permission for linking.

For the primary analysis we are therefore left with the more limited set of baseline covariates available in the PROCARE database. We note in this regard the also rather limited set of covariates adjusted for in the Dutch report [45]. These were age, gender, body mass index, Charlson score and ASA score for co-morbidity, history of abdominal problems, distance tumor anal verge, tumor complications, clinical stage.

In the PROCARE database we specifically have access to age, gender, height and weight (hence body mass index) and ASA score for co-morbidity (not the Charlson score). From the pre-operative data entry form we further consider: BMI with an indicator of its missingness, and the 9 categorical variables including a category for missingness: cSTAGE, ASA co-morbidity score, level of the tumor, mode of surgery, ventral tumor, cCRM, cT4, pre-operative incontinence and surgical technique.

2.5.3 Bench marking

We follow the suggestion of the PROCARE team and provide 95% CI for estimated case-mix-adjusted center effects and compare them with the relative P25 and P75 measures. Whether or not these relative measures also represent a meaningful delineation of areas of poor or excellent performance in absolute terms remains to be established. An expert consensus on this should allow complementing relative measures with more absolute markers of high and low performance. That clinically meaningful distances are captured in absolute terms is more generally one of the arguments pleading against formally ranking center performance. This was raised in the context of comparisons between institutions in the areas of health and education already in 1996 [49], and followed in our seminal reference [37]. We agree with this and for that reason do not pursue this avenue here. A well-known added complication is the difficulty in assessing uncertainty associated with ranks [50].

2.5.4 Visualization

We plan to use in sequence 1) funnel plots to indicate whether variation is present between centers beyond sampling variation: if this is the case, then 2) caterpillar plots will allow centers to position themselves at a glance in the distribution of observed center effects. The confidence intervals should caution against over-interpretation of ranks, while the vertical lines give relative and possibly also absolute markers of weak and good performance.

Summary points for the methods chapter

- **A more technical description of different techniques for risk-adjustment of binary and right-censored QCIs is presented, considering fixed effects outcome regression, random effects outcome regression, doubly robust propensity score methods and instrumental variable methods. These four techniques are all considered within the causal framework in which we aim at estimating the effect of choice of center of care on the outcome (QCI).**
- **It was decided not to pursue the instrumental variables approach since the identified instrumental variables for this setting (distance and region/location) were not made available from the PROCARE database and preliminary results showed that the presence of many centers results in very imprecise estimated effects.**
- **An extensive simulation exercise has shown that there is no single technique that performs uniformly better than another. We therefore performed all three analyses, and evaluated the combined results in light of their described strengths and limitations.**
- **Convergence problems when fitting simple models with center choice as fixed predictor have been identified. These problems were most prominent when small centers (with e.g. less than 5 patients) with few events were entered in the model. To allow for reliable results, we restricted estimation of center effects to centers with at least 5 patients (other centers may be grouped into one overlapping center).**
- **Issues related to the lack of access to known confounders (e.g. socio-economic status) are discussed. The risk-adjustment analysis was necessarily restricted to age and gender plus the baseline clinical patient-specific confounders available in the PROCARE database.**
- **Missing data problems have been discussed and we suggest multiple imputation techniques for reconstruction of the database under the missing at random assumption, while acknowledging that this assumption may well be violated.**

3 APPLICATION AND RESULTS ON THE PROCARE DATABASE

3.1 INTRODUCTION

In this chapter, we apply the first set of methods described in Chapter 0 to the data available in the PROCARE register up to August 27th, 2010. After summarizing basic descriptive statistics on centers, prognostic variables and QCIs, we embarked on a multivariate risk adjustment of outcome QCIs and univariate risk adjustment for process QCIs in all 8 domains. Results per center are shown in caterpillar plots and outlying centers on either side of the spectrum are noted.

Since these data originated from a voluntary register which is in constant development and not yet fully matured, results should be interpreted with caution. First, one should examine to the extent possible the patient inclusion rates as well as any selectivity in the choice of patients who enter the data register. Secondly, certain variables show more missing data than one would normally expect, others still show some inconsistencies and for specific survival outcomes, few events could result in a weak information base and instability of the statistical methods used. Correspondingly, the implementation developed below serves first and foremost as an illustration of the application of the methods in this setting, for now adapted to the current version of the database (follow up until August 27, 2010). Limitations of the data become evident in the patient-mix-adjusted QCIs, and especially in the build-up of the QI, which explicitly foresees to eliminate variables with excess missing data or limited information content. In other words, as the database becomes subject to more quality control and events accumulate over time, the composition of these indexes may change drastically under the same index building process. This being said, a lot has been learned from this pioneering effort.

Before starting with the approach and the descriptive analysis of our observed cohort, we elaborate briefly on the selective nature of our cohort. More on this can be found in Appendix 7 on the representativeness of the observed cohort. As explained early on, participation in PROCARE is on a voluntary basis. This means that the PROCARE database has no full coverage of rectal cancer patients in Belgium, because:

- Some centers do not participate
- Within participating centers, not every specialist involved in the treatment of rectal cancer, participates
- A participating specialist may not include all rectal cancer cases in PROCARE

To get a better insight into the participation rate (completeness) and the degree of randomness in participation, the PROCARE database has been linked with and compared to the BCR database. The BCR database has full coverage of all new primary cancers in Belgium from 2004 onwards and in addition, this database provides information on topography, morphology, stage, and the center that delivered the data. Therefore, the BCR database is the most suitable source to investigate the completeness of the PROCARE database. The link is, however, not without problems as explained first.

First, the rules for patients to be included in PROCARE are the following:

1. The lower limit of the tumor must be situated between 0 and 15 cm above the margo ani,
2. Only adenocarcinoma are included,
3. Only invasive tumors are included,
4. Multiple synchronous tumors are left out in PROCARE if it is indicated on the PROCARE data entry form. Note that before January 2010, this was not mentioned on the data entry form. Therefore, this exclusion criterion will not be considered for the present study,

5. Patients for whom the INSS is not indicated cannot be included in the PROCARE database system. People not residing in Belgium are not included in the PROCARE database.

For the present analysis, only PROCARE patients with year of incidence between 2005 and 2008 were selected, because the BCR data after 2008 are not yet available. As a rule, the PROCARE database was linked to the BCR selection of only those patients who fulfill the PROCARE criteria in the more strict sense. A second complication for linking both databases results from the way patients are attributed to centers. In PROCARE, the decision rule to assign a patient to a center is based on the medical specialism that delivered the data. In case a patient has been operated, the center of surgery is chosen. If this is missing, the center in which the pathological examination is performed is chosen. If this still yields no center, the center of chemotherapy is chosen. If no surgery is performed, then the center of chemotherapy is chosen. In the BCR database, it is less clear which speciality has delivered data. BCR data comes from multiple sources. More detail on this is given in Appendix 7.

To examine the completeness of the PROCARE database, a final selection was made of those patients with ICD code C20-C21 and an adenocarcinoma in the BCR database (PROCARE inclusion criteria).

The completeness of the PROCARE database per year of incidence is presented in Table 1. When 2005 is ignored (the prospective registration has started in 2006, only few centers have retrospectively registered patients of 2005), the completeness of the PROCARE database is around 30%. Also note that up until now, cases of 2008 are still being registered, so the actual proportion of 2008 might be underestimated.

Table 1: Proportion of PROCARE patients in BCR database.

Incidence year	N in PROCARE	N in BCR database	% coverage
2005	239	2091	11.4
2006	792	2287	34.6
2007	717	2298	31.2
2008	618	2340	26.4
Total	2366	9016	26.2

Next, it was investigated which stages, as registered in the BCR were best represented in the PROCARE database (Tables 2-3). The cases from the incidence year 2005 as well as stage 0 cases were not included for the analyses per stage.

For cStage I and II the proportion present in PROCARE is little less than 40%, whereas for stage III, this increases to nearly 48%. From the stage IV patients, only 25% were present in the PROCARE database. This was expected given that PROCARE has a firm surgical background and many stage IV patients did not undergo surgery. Twenty percent of patients with an unknown cStage in the BCR database were registered in PROCARE.

For pStage, the picture is similar, although the distribution between stage I, II and III is more equal.[7]

Table 2: Proportion of PROCARE patients in BCR database – per cStage (2005 not included)

cStage	N in PROCARE	N in BCR database	% coverage
I	212	563	37.7
II	313	813	38.5
III	810	1690	47.9
IV	194	781	24.8
NA	8	54	14.8
X	588	3021	19.5
TOTAL	2125	6922	30.7

7 Patients who have ypStage 0 in the PROCARE database, are registered in the BCR database with pStage X (88.7%), pStage I (6.7%), pStage II (2.1%) and pStage III (2.6%) respectively.

Table 3: Proportion of PROCARE patients in BCR database – per pStage (2005 not included)

pStage	N in PROCARE	N in BCR database	% coverage
I	501	1327	37.8
II	472	1322	35.8
III	576	1605	35.9
IV	97	388	25
NA	8	54	14.8
X	471	2226	21.2
TOTAL	2125	6922	30.7

Next, both databases were compared on age and sex. The mean age of the patients in the PROCARE database (67.0 years) was lower than the mean age of the patients who were only in the BCR database and not in the PROCARE database (69.9 years). This difference was significant $t(9013)=10.14$ ($p<.001$).

Furthermore, there were 61.9% male patients included in the PROCARE database. For the patients who were not included in this database, the percentage of men was 58.5%. This difference was also significant $\square^2$ (1) =8.74 ($p<.01$).

In conclusion, in the PROCARE database, the coverage of stage IV patients is lower, patients are on average younger, and the percentage of male patients is higher. In Appendix 7 (see Table 7) one finds more detail on this investigation as well as a table with coverage rates per center. Ideally we would also need to compare survival curves per center between the BCR and PROCARE database to get a more outcome directed measure of selectivity. Unfortunately the BCR could not provide this under the current database constraints.

3.2 METHODS

3.2.1 Descriptive statistics

As announced in section 2.1.7, descriptive statistics were computed for center, baseline patient-level characteristics (potential prognostic factors) and QCIs before embarking on more complex modeling. Special attention was paid to missing data.

We first describe the distribution of patients over the centers and the pace at which new patients are registered in PROCARE. To be informative while protecting confidentiality, centers will be (sample) *size-grouped* based on the number of patients registered in PROCARE (grouping levels: [1-10[, [10-20[, [20-40[, [40-60[, [60-80[, [80-100[and 100+ patients registered in PROCARE).

We then describe the study sample in terms of baseline characteristics, first univariately, then bivariately, followed by an examination of joint missingness patterns and some information on the follow-up information that is available in the PROCARE database.

For discrete variables (e.g. gender, cStage, …) tables show the overall frequency distribution, as well as for the size-grouped centers. For binary variables a caterpillar plot (see section 2.4.4 for more details) shows for individual (and one merged with less than 5 eligible patients) centers c unadjusted proportions ($\hat{p}_c$) with corresponding asymptotic confidence intervals:

$$\left[\hat{p}_c - 1.96\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c}}, \hat{p}_c + 1.96\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c}}\right]$$

Currently, for centers with proportions equal to 0% or 100%, no confidence intervals are provided, while lower or upper confidence limits below 0% or above 100% are truncated at 0% or 100%, respectively. For small centers the confidence intervals are hence approximate and merely give an indication of the level of uncertainty in this exploratory analysis. Where needed, exact confidence intervals may be given instead.

Continuous variables (e.g. age, BMI, …) are described using the five-number summary (i.e. minimum, P25, median, P75 and maximum), mean and standard deviation, overall and per size-grouped centers. Additionally, a box plot per center illustrates how distributions differ between centers.

For all QCIs there is – overall and for the size-grouped centers – a descriptive table with the total number of patients, the number of patients with missing information for the QCI, the number (%) of eligible patients, and

- for binary QCIs the number (%) of eligible patients meeting the QCI,8
- for survival QCIs the five-number summary of follow-up times, total person years, number of events and event rates are listed as well as Kaplan-Meier curves shown.
- for continuous QCIs the five-number summary of observed values.

Additionally, we plot for binary QCIs an unadjusted caterpillar plot (as described above), for survival QCIs (cStage-stratified) Kaplan-Meier curves and for continuous QCIs a box plot per center.[9]

3.2.2 Identification of relevant prognostic factors

Three sources are considered for identifying possibly relevant baseline characteristics as prognostic factors for outcome QCIs:

1. prognostic factors identified in the literature review of Appendix 2;
2. prognostic factors identified in the PROCARE consensus on risk/confounding factors per QCI of January 5th, 2011 (Appendix 2),
3. empirical associations between the QCI and a fixed set of prognostic factors occurring in the literature review and/or the PROCARE consensus: gender, age, BMI, ASA co-morbidity score, cStage, level of the tumor (high, mid or low), mode of surgery (elective or scheduled versus urgent or emergency), ventral tumor, cCRM positivity, cT4, preoperative incontinence and surgical technique used (PME versus TME).

For both outcome and process QCIs, statistically significant (p-value < 0.10) univariate associations with abovementioned baseline variables are reported.

To obtain an appropriate risk-adjustment model for the outcome QCIs, in a first step a *backward* modeling strategy (p-value for exclusion > 0.05) is applied to obtain a (multivariate) risk-adjustment model with only main effects. Next, in a *forward* modeling strategy (p-value for inclusion < 0.10), we consider all possible two-way interactions between main effects that were retained after the first step.[10]

---

8 In PROCARE II, this % was called a weighted mean, it corresponds to the QCI results for the cohort as a whole. At some point we will also talk about unweighted means, being the average of the center-specific proportions for that QCI.

9 **Note** In the caterpillar plot and box plot as described above, centers with less than 5 available patients have been grouped into one overlapping center.
For the binary QCIs funnel plots are produced, but not shown in this report to protect confidentiality of the centers.

10 **Note** The modeling strategy as described above is performed in a fixed effects outcome regression model with Firth's correction.
All risk-adjustment models for outcome QCIs will adjust for the prognostic factors gender and age, even if they are not significantly associated with the QCI.
When building these models, centers with less than 5 available patients for the QCI of interest are grouped into one overlapping center.
Handling of missing data
For categorical variables a separate level is created for the missing value.
For continuous variables, an interaction between the original variable and a dummy variable indicating missingness will enter the model. With very few (less than 5) missing values, we will discard these patients from the analysis.
Caution, using a missingness indicator to allow all patients in the model has some drawbacks; by conditioning on the missingness indicator one does not avoid the selection bias among the available data and adjusts only partially for confounding, it avoids deleting observations by making some assumptions on

3.2.3 Estimation of unadjusted and case mix adjusted center effects

As discussed in section 2.1, three methods have been explored for estimating the center effects for the outcome QCIs: fixed effects outcome regression and random effects (hierarchical) outcome regression for center effects related to the center's own patient population, and a doubly robust propensity score method for center effects related to a common reference population (i.e. the patients registered in the PROCARE database). The same (multivariate) risk-adjustment model, as obtained from the modeling strategy described in Section 3.2.2, is used in all three methods.

For ease of presentation on the larger scale, center effects are expressed as an 'excess' (relative to the average center) probability or outcome value, i.e. the obtained center-specific mean will be 'standardized' by subtracting from it the probability or mean outcome value one would expect to observe if all patients of that specific center have been treated in the average center. Results for the 'average center' are implicitly obtained by taking the unweighted (case-mix adjusted) mean of all centers considered for a QCI. In other words, this 'excess' probability or outcome value can be seen as the probability or outcome value associated with the specific center a patient was treated in. We refer to Appendix 1 (Chapter 10, section 1) for more detail on these 'excess' center effects.

For process QCIs, the center effects primarily remain unadjusted for prognostic factors, but statistically significant associations with prognostic factors are reported.

To visualize results, for outcome QCIs, unadjusted as well as adjusted center effects are estimated and compared in scatter plots. In addition, a caterpillar plot (see section 2.4.4 for more details) with the unadjusted center effects and a caterpillar plot with adjusted (from the fixed-effects model) center effects are produced. Equally, caterpillar plots with unadjusted center effects are reported for all process QCIs. Appendix 1 (Chapter 10, section 1) gives more detail on the computation of the error bars in these caterpillar plots.

3.2.3.1 Technical notes

While the propensity score analysis has desirable properties in large samples, we have found that the small size of various centers in this setting does not enable accurate assessment of the propensity score, i.e., the probability to attend a given center in function of patient characteristics. It must therefore at present be viewed as a more experimental analysis that is only included for illustrative purposes (in a more detailed analysis for QCI 1227 [%grade4_Tox_Preop_RT]). In particular, if we were to adhere to the heuristic 'rule of 10' according to which 10 events are needed per covariate in a logistic regression model to prevent important finite-sample bias, then each covariate that is included into the propensity score model for a given center would require at least 10 patients in that center. This imposes severe limitations on the number of covariates that we can realistically deal with in the propensity score model, or indeed any non-penalized regression model.

Future research will examine more closely a number of strategies that may help improve the performance of propensity score methods when the number of patients per center is relatively small. Some approaches are suggested in the Appendix 1.

Similarly, the hierarchical regression approach encounters problems in estimating (adjusted) center effects with standard pseudo-likelihood estimation techniques, often when there is a very low or very high event rate, for instance QCIs 1231 [%R0res], 1233a [%Leak_PME], 1234 [30d_mort], 1235 [%Perfor], 1235b [%Pos_Dist_margin], 1245 [%grade4_Tox_Prostop_CT]. Of course, there is little information to support or refute the shape of the random effects distribution assumed in those instances. The hierarchical regression model was then fitted using a maximum likelihood procedure with a Laplace approximation, which resolved the problematic model fit for all abovementioned QCIs except for 1235b and 1245.

---

the missing data. A multiple imputation strategy might be a more solid strategy for handling missing data, but this method then requires strong distributional assumptions.

For QCI 1235b the event rate is extremely low (2%) and for QCI 1245 there are only 128 patients eligible, hence few information is available for fitting the model. For these latter two QCIs no center effects from a hierarchical regression are reported.

Taking these estimation issues with the hierarchical regression approach and the propensity score method into account, we have decided to rely on the risk-adjusted center effects as obtained from the Firth-corrected fixed-effects regression approach for evaluating center performance and aggregating QCIs into quality indices (QIs).

To be consistent, unadjusted center effects for both process and outcome QCIs have also been obtained from a Firth-corrected fixed-effects outcome regression with 'center choice' then as the only predictor.

3.3 DESCRIPTION OF THE STUDY COHORT

The PROCARE database contains data on rectal cancer patients registered on a voluntary basis by the multidisciplinary clinical team, including surgeons, radiotherapists, oncologists and pathologists. The following exclusion criteria for registered patients from the database are applied by the Belgian Cancer Registry (BCR):

- patients in which the rectal cancer type is not an adenocarcinoma (e.g. carcinoid),
- patients in which the lower limit of an invasive adenocarcinoma is above 15 cm from the anal verge,
- patients with pStage 0 (i.e. pTis and pT0),[11]
- patients with synchronous primary tumors outside of the rectum,
- patients who do not reside in Belgium.

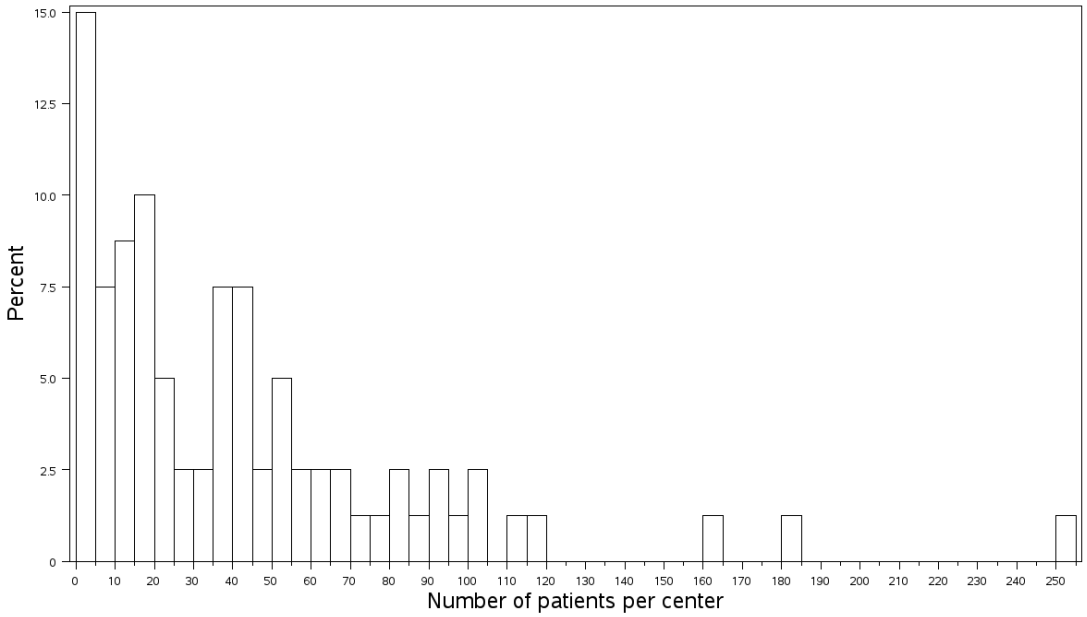
For the present study, patient inclusion was stopped on August 27th, 2010.

3.3.1 Centers

At the time of database closure, 3328 patients were included, involving 79 centers and 3 patients for whom the center of treatment is not known (even though the data must have been submitted in some way by 'a' center). These three patients are grouped into one 'artificial' center. The number of patients registered in the PROCARE database ranges from 1 patient in 5 centers up to 254 patients in 1 center. The median and average center sizes are respectively 32.5 and 41.6 patients. The distribution of the center sizes is shown in the histogram in Figure 8.

[11] **Note** that an ypStage 0 (ypTis and ypT0) is possible when the tumor is totally regressed after neoadjuvant radiochemotherapy and if the cT stage was cT1 or more and/or a biopsy or endoscopic resection confirmed the presence of an invasive cancer; these records are included in the database used for feedback.

Figure 8: Distribution of center sizes, i.e. number of patients registered in the PROCARE database.



As seen in Table 4, there are 12 centers with less than 5 patients registered in the PROCARE database. Note that one of these centers is the 'artificial' center representing the 3 patients for whom the center choice is not known.

Table 4: Frequency, cumulative frequency and cumulative percentage of the center sizes in the PROCARE database, up to centers with 10 patients.

Number of patients	Center frequency	Cumulative frequency	Cumulative percent
1	5	5	6.25
2	4	9	11.25
3	2	11	13.75
4	1	12	15.00
5	2	14	17.50
8	1	15	18.75
9	3	18	22.50
10	1	19	23.75
…	…		…
254	1	80	100

We found 20 centers with no patients registered in the PROCARE database after January 1st, 2008 (i.e. 2.5 years prior to the database closure). Eight of these centers have registered altogether less than 10 patients in the PROCARE database. This raises the following questions about centers that did not report any new patients 'recently':

- Are they still participating in PROCARE, and do they still expect feedback on their (or the overall) performance?
- Can these be centers that used to work independently but merged at a certain moment with another center? In this case, would it not be better to also include the patients from before the fusion into the merged center?

To protect confidentiality of centers with patients registered in the PROCARE database, no center-level descriptive statistics are provided, but instead descriptive statistics per grouping level as in Table 5, where grouping was based on the center sizes, e.g. grouping level [1-10[contains 68 patients from the 18 centers who registered less than 10 patients so far in the PROCARE database.

Table 5: Distribution of the number of patients and number of centers over the different grouping levels based on center size.

Grouping level	Number of patients	Number of centers
[1-10[	68	18
[10-20[	219	15
[20-40[	424	14
[40-60[	663	14
[60-80[	399	6
[80-100[	527	6
[100-]	1028	7

3.3.2 Prognostic factors

3.3.2.1 Gender and age

Of all patients in the PROCARE database, 61% are male, and the average age at incidence is 67 years with a standard deviation of 12 years. Forty-six percent of the patients in the database are older than 70 years. The age at diagnosis is not available for 198 (6%) of the patients in the database, due to a missing incidence date. This missingness occurs rather center-specifically: 165 (or 83%) of all missing incidence dates occur in 10 of the 80 centers, with more than 5 missing observations in each of these centers (details not shown to protect center confidentiality).

Appendix 6 (section 1.1, Tables 1-2) tabulates how age and gender are distributed in the size-grouped centers. Figures there show the percentage of male patients per center (A6-section 1.1, Figure 2) and a box plot of the age distribution per center (A6-section 1.1, Figure 3).

3.3.2.2 cStage

From Table 6 we can see how 46%, of the patients in the PROCARE database has cStage III, 13% has cStage IV, 15% has cStage II and 12% cStage I. There are 10 patients with cStage 0, but these are allowed in the database because they have a (y)pStage that is non-zero. However, for the statistical analyses per QCI the 10 patients with cStage 0 will be discarded.

For the remaining 392 (12%) patients, cStage is missing or could not be determined (X). The missingness (or X) of cStage is less center-specific than for age: there are 31 centers with a higher percentage missingness (or X) than the overall percentage of 14.9% (details not shown to protect center confidentiality).

Table 6: Distribution of cStage in the PROCARE database

cStage	Frequency	Percent
0	10	0.30
I	392	11.78
II	492	14.78
III	1520	45.67
IV	419	12.59
X	103	3.09
missing	392	11.78

In Table 7 the relation between the cStage and the (y)pStage is presented. It shows e.g. the number of (y)pStage I-II patients that had clinical stage III, and that the patients with cStage 0 do not have (y)pStage 0 (although missing for 2 patients). Beside this we observe substantial clinical over- and under-staging of patients. Note that for determining the (y)pStage with regard to metastases, the clinical M stage is used since there is no uniform determination of the pathological M stage. This explains the exact match of cStage IV and (y)pStage IV patients.

The distribution of the different cStages and (y)pStage over size-grouped centers is explored in Appendix 6 (Section 1.1, Tables 3-4).

Table 7: Relation between cStage and (y)pStage in the PROCARE database.

cStage	(y)pStage							Total
	0	I	II	III	IV	X	missing	
0	0	6	2	0	0	0	2	10
I	12	169	62	84	1	13	51	392
II	41	120	178	110	4	1	38	492
III	192	338	379	479	27	2	103	1520
IV	0	0	0	0	419	0	0	419
X	3	18	29	23	1	4	25	103
missing	6	46	59	49	9	158	65	392
Total	254	697	709	745	461	178	284	3328

The accuracy of cStaging relative to (y)pStaging is summarized in Table 8, overall and separately for patients who are known to have received neoadjuvant treatment and for patients for whom this is not known. From this we learn that 1245 of the 3328 patients are consistently staged, while 1082 have higher clinical stage (which might be related to rectal cancer regression after neoadjuvant treatment) and 296 lower clinical stage. For 705 patients the staging accuracy cannot be determined because either their cStage or (y)pStage is missing.

Note that for patients who are known to have received neoadjuvant treatment cStage and (y)pStage need not coincide due to the downstaging effects of neoadjuvant therapy. This is reflected in the high proportion (48%) of clinically over-staged patients in that group. In the group of patients with missing valid information on neoadjuvant treatment there is an overrepresentation of patients with missing cStage and/or (y)pStage. Hence, missingness of staging is associated with missingness of neoadjuvant treatment.

Table 8: Staging accuracy for all patients in the PROCARE database, overall and separately for patients that underwent neoadjuvant treatment or not. The (y)pStage is taken as the reference stage to assess under- or over-staging.

Clinical staging relative to (y)p-staging	Overall		Neoadjuvant treatment		No / missing neoadjuvant treatment	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Under-staging	296	8.89	92	4.88	204	14.15
Correct	1245	37.41	682	36.16+	563	39.04
Over-staging	1082	32.51	898	47.61	184	12.76
Missing c- or (y)pStage	705	21.18	214	11.35	491	34.05
Total	3328	100	1886	100	1442	100

3.3.2.3 Lower limit and tumor level

The lower limit of a tumor is – in order of priority - determined by 1) pretreatment rectoscopy, 2) pretreatment colonoscopy or 3) rectoscopy or colonoscopy at surgery and classified depending on the distance from the margo ani as shown in Table 9.

The distribution of the three tumor levels in the PROCARE database is presented in Table 9.

Table 9: Frequency and percentage of the level of the primary tumor in patients for whom the lower limit of the tumor is known

Lower limit tumor (cm)	Tumor level	Frequency	Percent
≤ 5 cm	**Low**	1292	42.94
>5 - ≤ 10 cm	**Mid**	1198	39.81
>10 cm	**High**	519	17.25
Available		3009	100
Missing		319	9.58
Overall total		3328	

The missingness of the lower limit of the tumor is quite center-specific: almost 80% of missings occur in 17 of the 80 centers, all with at least 10 patients registered in the PROCARE database (details not shown to protect center confidentiality).

Appendix 6 (section 1.1, Table 5) tabulates the more detailed distribution of tumor levels per size-grouped center.

3.3.2.4 BMI

BMI is known for 2084 (63%) of the 3328 patients in the PROCARE database. It ranges from 13 to almost 60 kg/m². The average and median BMI are respectively 25.7 and 25.2 kg/m².

BMI is categorized in four groups: underweight (BMI<18.5), normal (BMI < 30), obese (30 <= BMI < 35) and morbidly obese (BMI >= 35). The frequency distribution is given in Table 10. Eighty-one percent of the patients have a normal BMI, 12% is classified as obese, 70 patients (3.4%) as underweight and 74 patients (3.6%) as morbidly obese.

Table 10: Frequency and percentage of the categorized BMI levels for the patients in the PROCARE database for whom BMI is known (63%).

Categorized BMI	Frequency	Percent
Underweight	70	3.36
Normal	1691	81.14
Obese	249	11.95
Morbid_obese	74	3.55
Available	2084	100
Missing	1244	37.38
Overall total	3328	

Missingness of BMI occurs in all centers, and does not seem to be very much center-specific (details not shown to protect center confidentiality).

Appendix 6 (Section 1.1, Table 6 and Figure 4) provides a more detailed view of the distribution of BMI and frequency distribution of categorized BMI per (size-grouped) center.

3.3.2.5 ASA co-morbidity score

The ASA co-morbidity score can range from 1 to 5 with a higher value corresponding to worse general health, Table 11, [51].

Table 11: ASA classification

(Grade)	ASA classification	(Grade)	Revised classification
1	class 1	1a	Normal healthy patient.
		1b	Patient with mild systemic disease. Normal health patient, with operative or anesthetic risk(s).
2	class 2	2a	Patient with moderate systemic disease. Patient with mild systemic disease, with operative or anesthetic risk(s).
		2b	Patient with moderate to severe systemic disease that does not limit activity. Patient with mild systemic disease, with operative and anesthetic risks. Patient with moderate systemic disease, with operative or anesthetic risk(s).
3	class 3	3	Patient with severe systemic disease that limits activity, but is not incapacitating. Patient with moderate systemic disease that does not limit activity, with operative and anesthetic risk(s). Patient with moderate to severe systemic disease that does not limit activity, with operative or anesthetic risks.
4	class 4	4	Patient with an incapacitating systemic disease that is a constant threat to life. Patient with severe systemic disease that limits activity, incapacitated.
5	class 5	5	Moribund patient no expected to survive 24 hours with or without operation.

In the event of emergency operation, precede the number with an E.

Source: Reference [50]

The ASA score is known for 2637 (79%) of the 3328 patients in the PROCARE database. From Table 12 we learn that 28% of the patients have no co-morbidity, half of the patients experience moderate to severe systemic disease, 20% suffers from severe systemic disease that limits activity but is not incapacitating. Over 1% of the patients suffer from an incapacitating disease that is a constant threat to life and just 4 patients entered in a moribund state and did not expect to survive 24 hours, regardless of treatment.

Given the few patients with score 4 and 5, we have merged scores 3, 4 and 5 into one score. Missingness tends to be center-specific: over 80% of all missing ASA scores occur in 26 of the 80 centers, all with at least 10 patients registered in PROCARE (details not shown to protect center confidentiality).

More detail on the distribution of the ASA score over the size-grouped centers is provided in Appendix 6 (Section 1.1, Table 7).

Table 12: Frequency, percentage and cumulative distribution of the ASA co-morbidity score among patients for whom the ASA score is known.

ASA score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	742	28.14	742	28.14
2	1325	50.25	2067	78.38
3	534	20.25	2601	98.63
4	32	1.21	2633	99.85
5	4	0.15	2637	100.00
Available	2637	100		
Missing	691	20.76		
Overall total	3328			

3.3.2.6 *Tumor complications*

Preoperative tumor complications

Among the 2948 patients in the PROCARE database for whom information on tumor complications is available, 38 patients had tumor complications detected during pre-treatment coloscopy. No further detail about the type of complications is available in the database. The distribution of tumor complications over the size-grouped centers is provided in Appendix 6 (Section 1.1, Table 8).

Because of the very small proportion of reported tumor complications this prognostic factor is unlikely to be useful for risk-adjusting the outcome QCIs.

Operative tumor complications

In the operative part of the surgical PROCARE form, surgeons can report tumor complications before any mobilization. These complications are reported in four categories for 207 patients. There is no further indication on the status for the other patients (no complications or not reported), see Table 13.

Table 13: Distribution of the types of reported tumor complications during surgery among patients registered in the PROCARE database.

Operative tumor complications	Frequency	Percent
Free perforation	7	3.38
Peri-rectal abscess	31	14.98
Stenosis or obstruction	103	49.76
Other	66	31.88
Available	2948	100
Missing	380	11.42
Overall total	3328	

Because this is not a preoperative characteristic of the patients and it does not allow distinguishing between patients without a tumor complication and patients for whom the information on tumor complications is missing, it will not be used for risk-adjusting outcome QCIs.

3.3.2.7 *CEA performed*

The CEA serum before treatment is reported in the pre-treatment part of the surgical PROCARE form. Based on the reported value an indicator variable for CEA performed (yes or no) was constructed. For 2706 (81%) of the 3328 patients a CEA-level was reported. For the other 622 patients it is not sure whether CEA was either not performed or not reported.

The frequency distribution of CEA reported per size-grouped center is given in Appendix 6 (Section 1.1, Table 9). Because we are not able to distinguish between patients without CEA performed and patients for whom it is not known whether CEA was performed, this variable is not deemed reliable as prognostic factor.

3.3.2.8 Mode of surgery

The mode of surgery is known for 2872 (86.3%) of the patients in the PROCARE database and is either elective, scheduled, urgent or an emergency. Because the difference between elective and scheduled surgery and urgent and emergency surgery is not always very clear we merged these categories two-by-two. The frequency distribution of the resulting groups is presented in Table 14.

Table 14: Frequency and percentage of the mode of surgery for the 86.3% patients with a known mode of surgery in the PROCARE database.

Mode of surgery	Frequency	Percent
Elective-Scheduled	2820	98.19
Urgent-Emergency	52	1.81
Available	2872	100
Missing	456	13.70
Overall total	3328	

There appear to have been 52 urgent or emergency operations. Missingness of the mode of surgery occurs mainly in a selection of the centers: 75% of the missings occur in 16 of the 80 centers, all with at least 10 patients in the PROCARE database (details not shown to protect center confidentiality).

The distribution over the size-grouped centers is given in Table 10 in Appendix 6 (section 1.1).

3.3.2.9 Ventral tumor

Of the 2494 (75%) patients for whom the location of tumor is known in the database, 779 (31%) has a ventral tumor. The frequency distribution is shown in Table 15.

Table 15: Frequency and percentage of the distribution of ventral tumors for the 75% patients with a known tumor location in the PROCARE database.

Ventral tumor	Frequency	Percent
No	1715	68.77
Yes	779	31.23
Available	2494	100
Missing	834	25.06
Overall total	3328	

Missingness of tumor location does not appear to be strongly related to specific centers much (details not shown to protect center confidentiality).

The distribution of the percentage of ventral tumors over the size-grouped centers is given in Table 11 in Appendix 6 (Section 1.1).

3.3.2.10 Circumferential margin

The clinical circumferential margin (cCRM) is available for 612 (18%) of the 3328 patients in the PROCARE database. When defining cCRM positivity as having a circumferential margin < 5mm, 448 (73%) of the patients with known cCRM have a positive cCRM, Table 16.

Table 16: Frequency and percentage of the distribution of cCRM positivity for the 18% patients with a cCRM in the PROCARE database.

cCRM positive	Frequency	Percent
No	164	26.80
Yes	448	73.20
Available	612	100
Missing	2716	81.62
Overall total	3328	

The pathological circumferential margin, (y)pCRM, is available in the PROCARE database for 2144 (64%) of the 3328 patients. When defining (y)pCRM positivity as having a circumferential margin < 1mm, 387 (18%) of the patients with known (y)pCRM have a positive (y)pCRM, Table 17.

Table 17: Frequency and percentage of the distribution of (y)pCRM positivity for the 64% patients with a (y)pCRM in the PROCARE database.

(y)pCRM positive	Frequency	Percent
No	1757	81.95
Yes	387	18.05
Available	2144	100
Missing	1184	35.58
Overall total	3328	

Missingness of cCRM and (y)pCRM occurs in almost all centers and is as such not center-specific (details not shown to protect center confidentiality).

The distribution over the size-grouped centers is given in Table 12 and Table 13 in Appendix 6 (Section 1.1).

3.3.3 Associations between prognostic factors

To examine associations between some of the above-described prognostic factors, we provide cross-tabulation of key prognostic factors in the PROCARE database.

Table 18 shows the distribution of the kind of neoadjuvant treatment given versus cStage. Remarkably, for 1442 (43%) of the 3328 patients in the PROCARE database it was not reported whether or which treatment they received. This is problematic since we cannot distinguish between patients who did not receive treatment and patients for whom the kind of treatment they received was missing. A recommendation for improving the quality of the PROCARE database is to link it to the administrative database to correct these kinds of gaps.

Other rather unexpected observations relate to the substantial number of cStage 0-I-II patients receiving long-course neoadjuvant treatment. Further investigation of these patients reveals that 63% of the cStage 0-I-II patients receiving long-course neoadjuvant radiochemotherapy have tumor located in the lower third of the rectum. This is substantially more than the overall percentage of patients with a tumor in the lower third of the rectum (Table 9).

Table 19 shows the frequency table of the mode of surgery versus cStage. Here we observe that for 86 (21%) of the 419 cStage IV patients, the mode of surgery is missing. This is proportionally higher than for lower-staged patients and might be an indication that these patients in fact did not undergo any surgery. It would be interesting if the database would allow distinguishing between missing mode of surgery and no surgery performed, which is currently not possible.

Table 20 cross-tabulates cStage and tumor level among patients with a recorded tumor level. A significant association is found between cStage and tumor level (p-value < 0.0001) with lower cStages 0/I/II overrepresented among high tumors (10 – 15 cm from the anal verge) and cStage III underrepresented among high tumors.

Table 21 cross-tabulates ASA score versus BMI level among patients with recorded ASA score and BMI level. There is no significant association between ASA score and BMI level (p-value = 0.11).

Table 18: Cross-tabulation of neoadjuvant treatment versus cStage.

cStage	Neoadjuvant treatment										Total
	Long course RT + chemo	Long course RT + no/unknown chemo	Short course RT + chemo	Short course RT + no/unknown chemo	Missing RT course + chemo	Missing RT course + no/unknown chemo	Other RT course + chemo	Other RT course + no/unknown chemo	No/unknown RT + chemo	No/unknown treatment	
0/I	40	2	0	21	2	4	0	0	0	333	392
II	180	23	4	68	23	6	0	1	0	187	492
III	954	68	6	87	81	10	0	0	1	313	1520
IV	114	10	2	28	14	3	1	0	9	238	419
X/miss	70	14	0	12	22	5	0	0	1	371	103
Total	1358	117	12	216	142	28	1	1	11	1442	3328

Table 19: Cross-tabulation of mode of surgery versus cStage among all patients in the PROCARE database

cStage	Mode of surgery			Total
	Elective / Scheduled	Urgent / Emergency	Missing	
0/I	356	3	43	402
II	455	11	26	492
III	1436	9	75	1520
IV	317	16	86	419
X/miss	256	13	226	495
Total	2820	52	456	3328

Table 20: Cross-tabulation of cStage versus tumor level among patients in the PROCARE database with valid tumor level information.

Tumor level	cStage					Total
	0/I	II	III	IV	X/miss	
Low	154	211	682	152	93	1292
Mid	148	165	613	161	111	1198
High	86	100	181	70	82	519
Total	388	476	1476	383	286	3328

Table 21: Cross-tabulation of ASA score versus BMI among patients with valid information for both BMI and ASA score in the PROCARE database.

BMI	ASA score			Total
	I	II	III	
Underweight	17	35	14	66
Normal	447	823	333	1603
Obese	66	129	45	240
Morbidly obese	10	45	17	72
Total	540	1032	409	1981

Table 22 cross-tabulates ASA score versus age categories among patients with recorded ASA score and age. A significant association is found between ASA score and age (*p*-value < 0.0001), with more frequent ASA score II or III in the older age categories.

Table 22: Cross-tabulation of ASA score versus age categories among the patients with valid information for both age and ASA score in the PROCARE database.

Age category	ASA score			Total
	I	II	III	
[20-40[	27	20	2	49
[40-50[	100	69	4	173
[50-60[	199	233	36	468
[60-70[	233	416	115	764
[70-80[	150	431	230	811
[80-[	31	156	183	370
Total	740	1325	570	2635

3.3.4 Joint missingness patterns of prognostic factors

The joint missingness of the prognostic factors is examined in Table 14 in Appendix 6 (section 1.1). This table can be read in different directions:

- Row-wise, e.g.
  - From the first row we learn that only 1632 (49%) of the 3328 patients in the PROCARE database has non-missing information for the basic prognostic factors (age, gender, BMI, ASA co-morbidity score, cStage, (y)pStage and tumor level),
  - the third row tells us that 126 patients (first column) had a missing value for just one prognostic factor (1 in last column): cStage (and 0 in column 'cStage'),
  - the second last row shows 170 patients (first column) with a missing value for 6 prognostic factors (6 in last column): age, (y)pStage, tumor level, cStage, ASA score and BMI,
  - the last row informs us about the total number of missing values per prognostic factor.
- Column-wise, e.g. from the 'age'-column we learn that there is one patient with just missing age and BMI, there are 2 patients with missing age, ASA score and BMI and that there is one patient with missing age, tumor level and BMI, but otherwise all values available. For all other patients with missing age, there are at least 3 other prognostic factors that are also missing.

3.3.5 Follow-up information

The quality of the general quality indicators (overall survival, disease-specific survival, local recurrence and disease-free survival) heavily relies on 1) the quality of the link with the Crossroads Bank of Social Security (CBSS) for the dead/alive status of all patients in the PROCARE database and 2) the follow-up information as (should be) provided by the clinical teams on a 6 monthly basis.

First note that for 198 of the 3328 patients in the PROCARE database no incidence date is available and no other information can be used to approximate the incidence of rectal cancer in these patients. As all survival QCIs start counting time from the incidence date, these 198 are automatically excluded from the analysis.

The dead/alive status is known for 3315 (99.6%) of the 3328 patients in the PROCARE database. The other 13 patients were not found in the CBSS and hence no information on their dead/alive status is available. For the patients who are known to have died, the corresponding date of death is available.

Follow-up information in the PROCARE database, on the other hand, is far from complete for most patients. Overall, there are 1878 (56%) patients in the PROCARE database who have at least one follow-up form recorded besides the incidence date. Table 23 provides more detailed information on the frequency distribution of recorded follow-up forms. Only these patients can be considered for survival QCIs related to recurrence of the disease (local recurrence or distant metastasis). We lack an indicator of whether the last available follow-up corresponds to the last follow-up as expected based on the incidence date.

Table 23: Number of patients in the PROCARE database for whom follow-up information is available (per follow-up period) and the distribution of timing of the last available follow-up.

Follow-up (months)	# patients	# patients with last FU	Cum. # patients with last FU
6	953	388	388
12	777	319	707
18	522	216	923
24	422	218	1141
30	287	165	1306
36	231	158	1464
42	148	125	1589
48	69	69	1658
54	0	0	1658
60	0	0	1658
Missing	220	220	1878

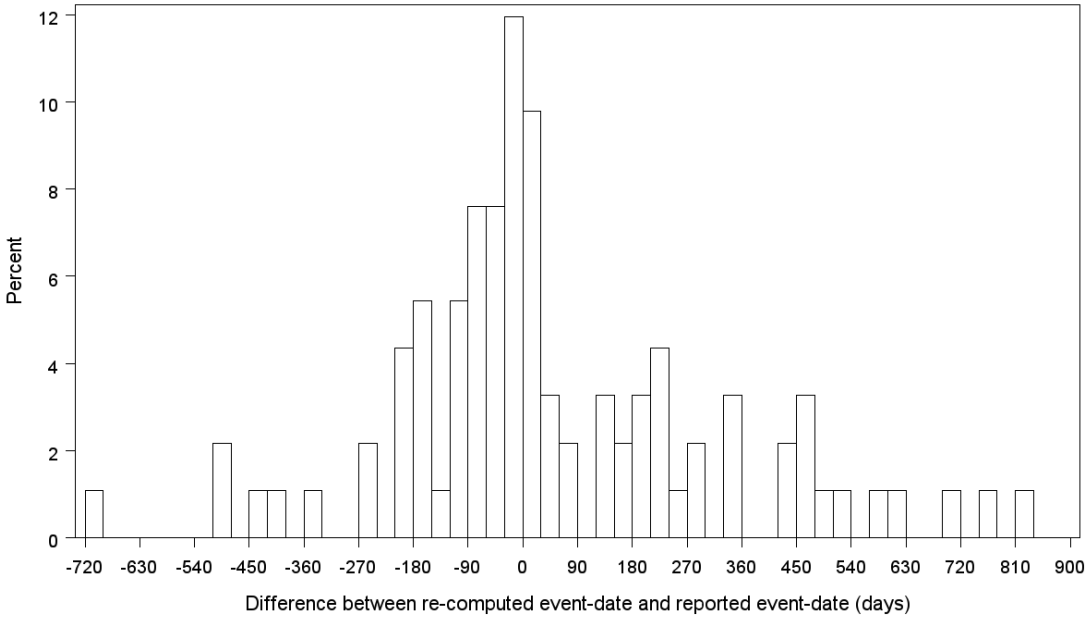
For 220 of these 1878 patients with some available follow-up information, it is not known to which follow-up time (6 or 12 or 18 or … months after incidence date) this information pertains. Among these patients are 137 patients with a reported local recurrence or distant metastasis and 83 patients without a recurrence. For 131 of the 137 patients with a reported local recurrence or distant metastasis, the length of their follow-up can in principle be computed from the reported date of local recurrence or distant metastasis, while for the other 6 no such date is available. For the 83 patients without a reported recurrence, the length of follow-up cannot be obtained from available information.

Hence, another 89 patients lack the follow-up information for recurrence-related survival QCIs. Additionally discarding 8 patients with missing dead/alive status, we arrive at 1781 (54%) patients in the PROCARE database who can be used for recurrence-related survival QCIs. Note that the missing subset is likely selective in terms of local recurrence.

For 34 of the 1878 patients with available follow-up information and with a reported local recurrence and/or distant metastasis, the date of this event is not available in the database, but can be recomputed based on the last follow-up (6, 12, 18, … months) as: incidence date + (last follow-up)*30.4.

If we compare this recomputed date of recurrence with the reported date of recurrence (for patients for whom this date is reported), we notice that it is not a very accurate surrogate, with differences ranging between 720 days (2 years) in one direction and 827.8 days (2¼ years), see also Figure 9.

Figure 9: Distribution of the difference in recomputed event date and reported event date in the PROCARE database based on the formula above.



3.4 DESCRIPTION AND ANALYSIS OF INDIVIDUAL QUALITY OF CARE INDICATORS

Appendix 6 (Section 2) shows the full results for the list of about 40 QCIs as presented in the PROCARE consensus of July 24th, 2010. By way of example, we run here through the analysis of arguably the most important outcome measure: `overall survival'.

3.4.1 The worked out case of Overall Survival QCI 1111

In principle, QCI definitions as provided in the PROCARE documents are clear, but for good practice these QCIs need to be translated into technically correct working definitions containing the observed time and endpoint, which may be either the event of interest, a competing event or censoring.

Censoring happens for patients for whom no event of interest was observed before closure of the database. The current PROCARE database was closed at the end of October 2010 and was linked with the death records of the version of the Cross Banks Social Security database of August 2010. All deaths up to July 31st, 2010 are registered in the database we have at our disposal. Therefore, July 31st, 2010 will be used as the administrative censoring date for the current PROCARE database.

PROCARE definition

N: Number of patients in denominator that survived 1-5 years.

D: Number of patients for whom the national registry number is known and have a follow-up of 1-5 years, respectively. Survival status was obtained through cross-link with the Crossroads Bank for Social Security (CBSS).

Working definition

Kaplan-Meier based probability of surviving at least 5 years after rectal cancer diagnosis. This involves follow-up time since incidence date (in years) and as endpoint the event (death) or censoring.

D: Number of patients for whom the national registry number is known and the incidence date is known to be before the administrative censoring date (July 31, 2010).

3.4.1.1 Description

All together, 3103 (94%) of the 3318 patients meet the requirements for the denominator of this QCI. Of the 215 patients who are not eligible, 13 have an unknown social security number, 198 had a missing incidence date and 4 had an incidence date after July 31st, 2010.

Of the 3103 selected patients, 697 (22%) patients have died since their incidence date over a median follow-up time of 2.5 years. More detail per size-grouped center on the follow-up time, person years and event rate can be found in Section 1.2 (Table 15) of Appendix 6. After merging centers with less than 5 eligible patients into one overlapping center, 69 centers remain to evaluate performance with regard to overall survival. The estimated survival curve for the study population is given as the green line among the cStage-stratified Kaplan-Meier curves presented in Figure 10, corresponding (y)pStage-stratified Kaplan-Meier curves can be found in Appendix 6 (Figure 5). Table 24 presents the number of patients, number of events and censored patients and the percentage of censored patients per cStage level.

From Table 15 in Appendix 6 (section 1.2) and Figure 10 we learn that the maximum follow-up time is less than 5 years for all cStage levels. This will be resolved in time when the PROCARE database matures. Given the currently available data, in this report we calculate 3-year survival probabilities instead of 5-year survival probabilities for all survival-related QCIs. When x-year survival is calculated based on modeled survival data, one may introduce administrative censoring at x-years to avoid unnecessary reliance on modeling assumptions beyond that time.

Figure 10: Kaplan-Meier curves per cStage stratum, estimating the (unadjusted) probability of surviving t years after the incidence of rectal cancer. The unstratified curve is shown in green. Numbers at risk for that curve are shown on the x-axis.

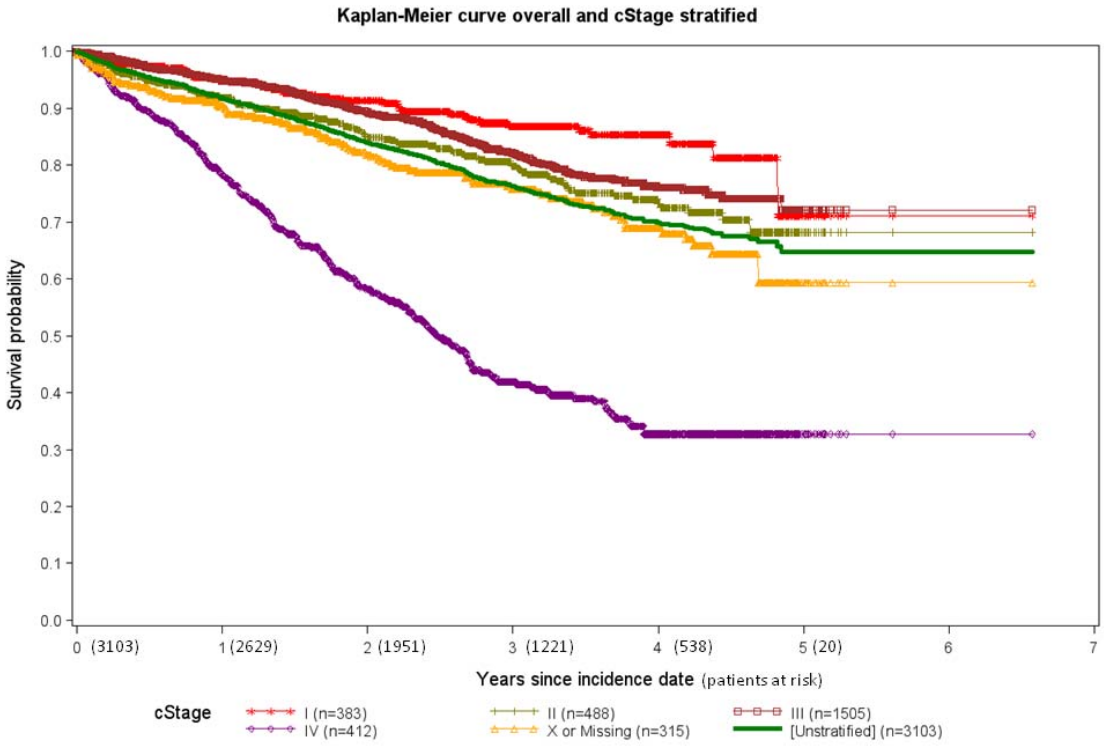


Table 24: The number of patients, events and censored patients, per level of cStage.

cStage	# patients	# Events	# Censored	% Censored
I	383	45	338	88.25
II	488	104	384	78.69
III	1505	247	1258	83.59
IV	412	217	195	47.33
X/Missing	315	84	231	73.33
Total	3103	697	2406	77.54

Unexpectedly, estimates in Figure 10 (and percentage censored in Table 24) show cStage II patients having lower survival probabilities than cStage III patients. This observation is borderline significant (hazard ratio [ref. = III] = 1.23 [0.98, 1.55], *p*-value = 0.08). Possible explanations could be a consistent clinical over-staging of actual stage II patients (see also Table 7) or that cStage III gets the better treatment.

As there is no significant difference in overall survival between patients with missing cStage and cStage X (hazard ratio [ref. = Missing] = 0.77 [0.47, 1.28], *p*-value = 0.31), we collapse these two categories into one overlapping category. To examine possible non-randomness of the collapsed X/missing category, the expected survival probability for a random patient with available cStage information in the PROCARE database is computed at each event time (as a weighted average of survival probabilities for cStage I, II, III and IV patients), and shown in Figure 11.

Figure 11: cStage-stratified Kaplan-Meier survival curves with an extra survival curve (red dotted line) showing the probability of surviving t years for a random patient (with available cStage information) in the PROCARE database. The green line presents the survival probability after t years for patients with X/missing cStage.

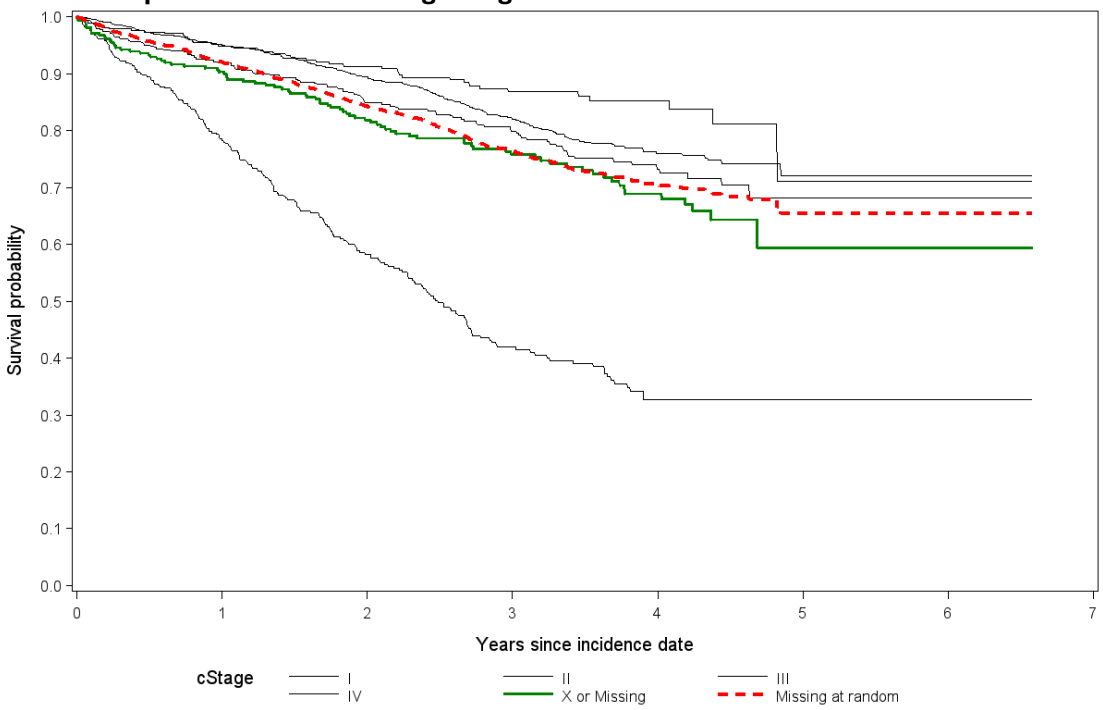


Figure 11 shows how the survival probabilities of patients with cStage X/missing are very similar to the weighted average of survival probabilities for cStage I, II, III, IV patients, the data is hence consistent with missingness at random at this level. This would leave a set of patients with available cStage information that is representative of the total PROCARE population. Whether this is actually the case, we do not know.

3.4.1.2 *Identification of relevant prognostic factors*

Prognostic factors identified in the literature review of Appendix 2

Appendix 2 contains an extensive list of prognostic factors that were found to be significantly associated with overall survival. Given our goals and methods as explained in section 2.1, we only consider pre-treatment prognostic factors of which BMI, co-morbidity status, age, tumor location and TNM-stage are available in the PROCARE database.

Prognostic factors identified in the PROCARE consensus on risk/confounding factors per QCI

Results from the PROCARE consensus survey in Appendix 3 suggest the following factors are to be used for adjustment: age > 70 years, gender, extraperitoneal location (i.e. tumor level), (y)pN+, no tumor response to long course radio(chemo)therapy.

Of these, the former 3 will be considered when building the risk-adjustment model for this QCI. The latter 2 are post-treatment factors, which can already be influenced by the quality of care provided by a certain center, and would hence 'cover' part of the center effect. Note that the (clinical) nodal status is partly considered through stratification by cStage. We will hence not adjust for it.

Empirical associations

Significant univariate associations between available prognostic factors and cStage-stratified overall survival are presented in Table 25.

Table 25: Hazard ratio [95% Wald confidence interval] estimate and corresponding p-value from univariate cStage-stratified Cox regression models for QCI 1111 [OS]. For categorical variables, 'ref.' indicates the reference level.

Prognostic factor	Level	Hazard ratio [95% CI]	*p*-value	Joint *p*-value
Age (years)	(continuous)	1.05 [1.04, 1.06]	< 0.0001	
BMI (kg/m²)	(continuous)	0.99 [0.97, 1.01]	0.33	< 0.0001
Missing BMI [ref. = Not missing]	Missing	1.28 [0.73, 2.26]	0.39	
Gender [ref. = Male]	Female	0.86 [0.74, 1.01]	0.06	
ASA score [ref. = I]	II	1.19 [0.94, 1.50]	< 0.0001	
	III-V	3.39 [2.68, 4.29]		
	Missing	2.67 [2.10, 3.41]		
Level [ref. = Low]	Mid	1.07 [0.91, 1.27]	0.01	
	High	1.01 [0.81, 1.26]		
	Missing	1.71 [1.25, 2.33]		
Mode of surgery [ref. = Elective/Sch.]	Urgent / Emergency	3.38 [2.33, 4.89]	< 0.0001	
	Missing	2.41 [1.95, 2.98]		
Ventral tumor [ref. = No]	Yes	0.78 [0.64, 0.95]	< 0.0001	
	Missing	1.54 [1.29, 1.83]		
cCRM positive [ref. = No]	Yes	0.96 [0.59, 1.54]	0.02	
	Missing	1.36 [0.90, 2.07]		
cT4 [ref. = No]	Yes	1.52 [1.23, 1.87]	0.0003	
	Missing	1.24 [0.91, 1.68]		

Preoper. incontinence [ref. = No]	Yes	1.43 [1.17, 1.75]	0.0006	
	Missing	1.42 [0.99, 2;03]		
Surgical technique [ref. = PME]	TME	0.83 [0.68, 1.00]	< 0.0001	
	Missing	1.94 [1.52, 2.49]		

interpretation of the hazard ratios obtained from univariate Cox regression models, illustrated using the results of Table 25.

Interpretation of the hazard ratio for continuous covariates:

- For each BMI increase of 1 unit (kg/m^2) in the group of patients with a valid BMI, the hazard (of dying) is estimated to 'decrease' with a factor 0.99; the 95% confidence interval [0.97, 1.01] accounts for finite-sample imprecision. This effect should be considered together with the missingness indicator for BMI: for the group of patients with missing BMI the hazard is estimated to be 1.28 times higher than for the patient group with a valid BMI.
- A similar interpretation holds for age; per age increase of 1 year in the group of patients with a valid age, the hazard (of dying) is estimated to increase with a factor 1.05 (95% CI [1.04, 1.06]). Note that all patients with missing age are automatically excluded from the denominator since they also have a missing incidence date.

Interpretation of the hazard ratios for a categorical covariate:

- E.g. for the ASA score: for the group of patients with ASA score III, IV or V the hazard (of dying) is estimated to be 3.39 times higher than for patients with ASA score I (95% CI [2.68, 4.29]).

3.4.1.3 Estimation of unadjusted and case-mix-adjusted center effects

The unadjusted center effects are estimated from a (Firth-corrected) fixed effects Cox regression model and expressed as 'excess' probabilities of surviving 3 years after the incidence of rectal cancer. This 'excess' probability (relative to the average center) can be seen as the center-specific deviation in probability of surviving 3 years after the incidence of rectal cancer associated with the population specific to the center under evaluation.

Unadjusted center effects range between -25% and 19% 'excess' probability of surviving 3 years after the incidence of rectal cancer. The P25 and P75 are located at -9% respectively 5%, hence the interquartile range is 14%.

A model building procedure was performed to identify joint associations between prognostic factors and overall survival, stratified by cStage. In this, the prognostic factors BMI, ASA co-morbidity score and level of the tumor (high, mid or low) were considered, while the adjustment for age and gender was always made.

A model with main effects for age (with a different slope before and after the breakpoint of 70 years), gender, ASA score and BMI was retained. A significant interaction between gender and BMI was found (p-value = 0.002), indicating that survival rates differ for different levels of BMI depending on the gender. In other words, in the PROCARE database BMI was missing for a subset of patients with increased hazard, to a different degree for men and women.

The hazard ratios (with corresponding 95% confidence interval and p-value) for the prognostic factors in the multivariate risk-adjustment model are presented in Table 26.

Table 26: Hazard ratio [95% Wald confidence interval] estimate and corresponding p-value from the final multivariate cStage-stratified Cox regression model for QCI 1111 [OS]. For categorical variables, 'ref.' indicates the reference level.

Prognostic factor	Level	Hazard ratio [95% CI]	*p*-value	Joint *p*-value
Age	(continuous)	1.00 [0.99, 1.02]	0.69	< 0.0001
Age (+ 70 years)	(continuous)	1.08 [1.06, 1.11]	< 0.0001	
BMI (for males)	(continuous)	0.96 [0.93, 0.99]	0.007	0.01
Missing BMI [ref. = Not missing]	Missing	0.41 [0.18, 0.91]	0.03	
Gender [ref. = Male]	Female	0.12 [0.04, 0.39]	0.0004	
Gender – BMI interaction [ref. = Male, (cont.)]	Female, (continuous)	1.07 [1.02, 1.12]	0.001	0.002
Gender - Missing BMI interaction [ref. = Male, Not missing]	Female, Missing	7.44 [2.23, 24.80]	0.03	
ASA score [ref. = I]	II	1.07 [0.82, 1.39]	< 0.0001	
	III-V	2.74 [2.06, 3.64]		
	Missing	1.98 [1.46, 2.69]		

Interpretation of the hazard ratios obtained from multivariate Cox regression models, illustrated using the results of Table 26

Interpretation of the hazard ratio for continuous covariates:

- Per age increase of 1 year in the group of patients younger than 70 years *with otherwise the same characteristics in terms of BMI, gender and ASA score*, the hazard (of dying) is estimated to 'increase' with a factor 1.00 (95% CI [0.99, 1.02]).

  Per age increase of 1 year in patients older than 70 years *with otherwise the same characteristics in terms of BMI, gender and ASA score*, the hazard (of dying) is estimated to additionally increase with a factor 1.08 (given 95% increase CI [1.06, 1.11]).

- For a BMI, there is a separate effect for male and female patients.
  - For males, per increase of 1 unit (kg/m^2) in patients with a valid BMI *and otherwise the same characteristics with regard to age and ASA score*, the hazard (of dying) is estimated to decrease with a factor 0.96; the 95% confidence interval [0.93, 0.99] accounts for finite-sample imprecision.
  - For females, per increase of 1 unit (kg/m^2) in patients with a valid BMI *and otherwise the same characteristics in terms of age and ASA score*, the hazard (of dying) is estimated to increase with a factor 1.07; the 95% confidence interval [1.02, 1.12] accounts for finite-sample imprecision.

  Note that these effects should be considered together with the missingness indicator for BMI. Also, the goal here is prediction correction and not causal effect estimation for baseline characteristics.

Interpretation of the hazard ratios for a categorical covariate:

- E.g. for the ASA score: for the group of patients with ASA score III, IV or V, the hazard (of dying) is estimated to be 2.74 times higher than for patients with ASA score I *with otherwise the same available or availability characteristics in terms of age, BMI and gender* (95% CI [2.06, 3.64]).

From Table 26 it appears that the association between missing BMI and overall survival is very pronounced in the multivariate risk-adjustment model. There is no immediate explanation for this observation, but it cannot be ignored. The fact that this selectivity is different for male and female patients makes it extra complex. From Table 27 we estimate that for women with missing BMI in most cStage levels, the event rate is twice that in women with valid BMI, while for men this ratio of event rates is much smaller. Table 28 shows how the stage distribution among patients with missing BMI is different for men and women. For male patients with missing BMI, cStage IV patients are overrepresented, while for female patients with missing BMI cStage I patients are overrepresented.

These two observations shed extra light on the selective missingness, but do not completely explain this difference between men and women.

Table 27: Number of patients, person years, number of events and event rate for each cStage-Gender-BMI missingness combination among patients eligible for QCI 1111 [OS] with a non-missing/X cStage.

cStage-Gender-Missing BMI	Total N (missing)	N (%) eligible	Person years	N (%) events	Event rate	Ratio
I-Female-Missing	79 (3)	76 (96%)	191.28	10 (13%)	0.05228	2.7
I-Female-Not Missing	90 (0)	90 (100%)	259.67	5 (6%)	0.01926	
I-Male-Missing	80 (4)	76 (95%)	183.50	13 (17%)	0.07084	1.4
I-Male-Not Missing	143 (1)	141 (99%)	347.96	17 (12%)	0.04886	
II-Female-Missing	68 (2)	65 (96%)	159.17	18 (28%)	0.11309	1.9
II-Female-Not Missing	130 (0)	130 (100%)	376.92	22 (17%)	0.05837	
II-Male-Missing	89 (0)	89 (100%)	231.99	28 (31%)	0.12070	1.9
II-Male-Not Missing	205 (1)	204 (100%)	562.29	36 (18%)	0.06402	
III-Female-Missing	160 (5)	155 (97%)	357.33	40 (26%)	0.11194	2.4
III-Female-Not Missing	399 (3)	396 (99%)	1064.55	50 (13%)	0.04697	
III-Male-Missing	255 (3)	252 (99%)	662.70	48 (19%)	0.07243	1.2
III-Male-Not Missing	706 (3)	702 (99%)	1832.79	109 (16%)	0.05947	
IV-Female-Missing	63 (3)	60 (95%)	100.57	38 (63%)	0.37786	2.5
IV-Female-Not Missing	92 (0)	92 (100%)	203.81	31 (34%)	0.15210	
IV-Male-Missing	113 (3)	110 (97%)	195.40	69 (63%)	0.35313	1.4
IV-Male-Not Missing	151 (1)	150 (99%)	315.70	79 (53%)	0.25024	

Table 28: Row percentages of the cross-tabulation of cStage versus BMI missingness for patients with non-missing/X cStage in the PROCARE database, separately for male and female patients.

Row Percent	Male					Female				
	I	II	III	IV	Total	I	II	III	IV	Total
BMI Not Missing	11.87	17.01	58.59	12.53	69.17	12.66	18.28	56.12	12.94	65.7
BMI Missing	14.90	16.57	47.49	***21.04***	30.83	***21.35***	18.38	43.24	17.03	34.23
Total	12.80	16.88	55.17	15.15	100.00	15.63	18.32	51.71	14.34	100.00

Adjusted center effects are estimated from both a fixed effects outcome regression (i.e. a Firth-corrected Cox regression model) and a hierarchical outcome regression (i.e. frailty Cox regression model). As described before, center effects are expressed as the 'excess' probability of 3-year survival relative to the 'average' center.

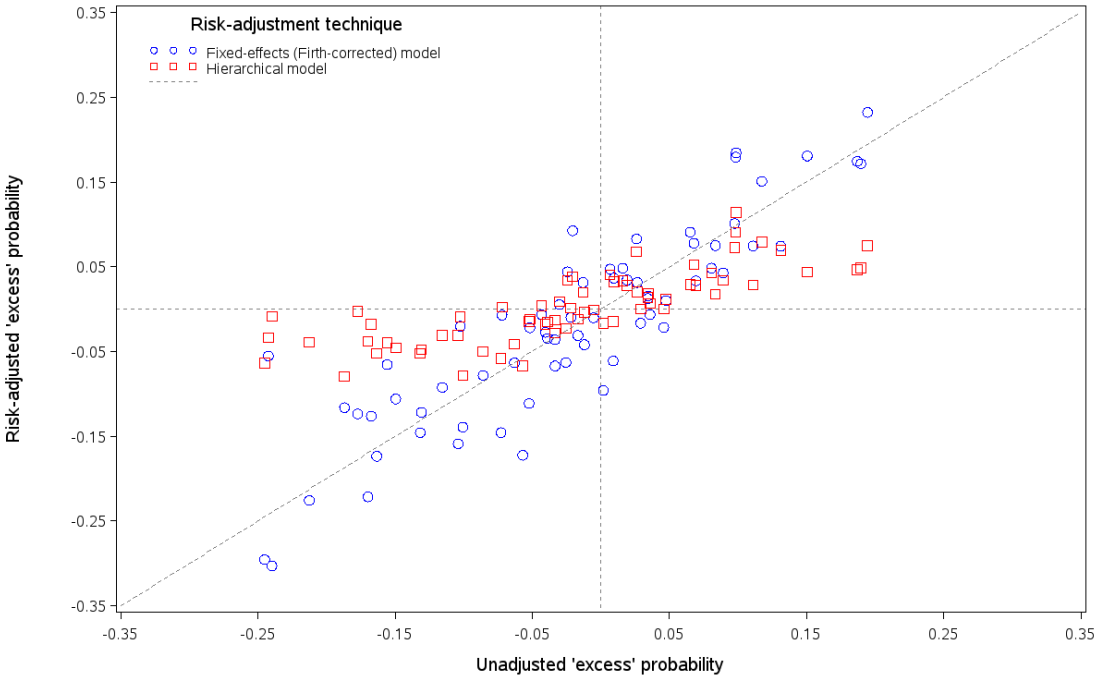
A basic summary of the distribution of center effects as estimated from the unadjusted analysis and the two adjusted analyses is presented in Table 29.

Table 29: Minimum, P25, P75, maximum and interquartile range of the center effects (%) for QCI 1111 [OS] as estimated by different methods.

Method	Min	P25	P75	Max	IQR
Unadjusted	-25	-9	5	19	14
Fixed-effects regression	-30	-9	5	23	14
Hierarchical regression	-8	-3	3	11	6

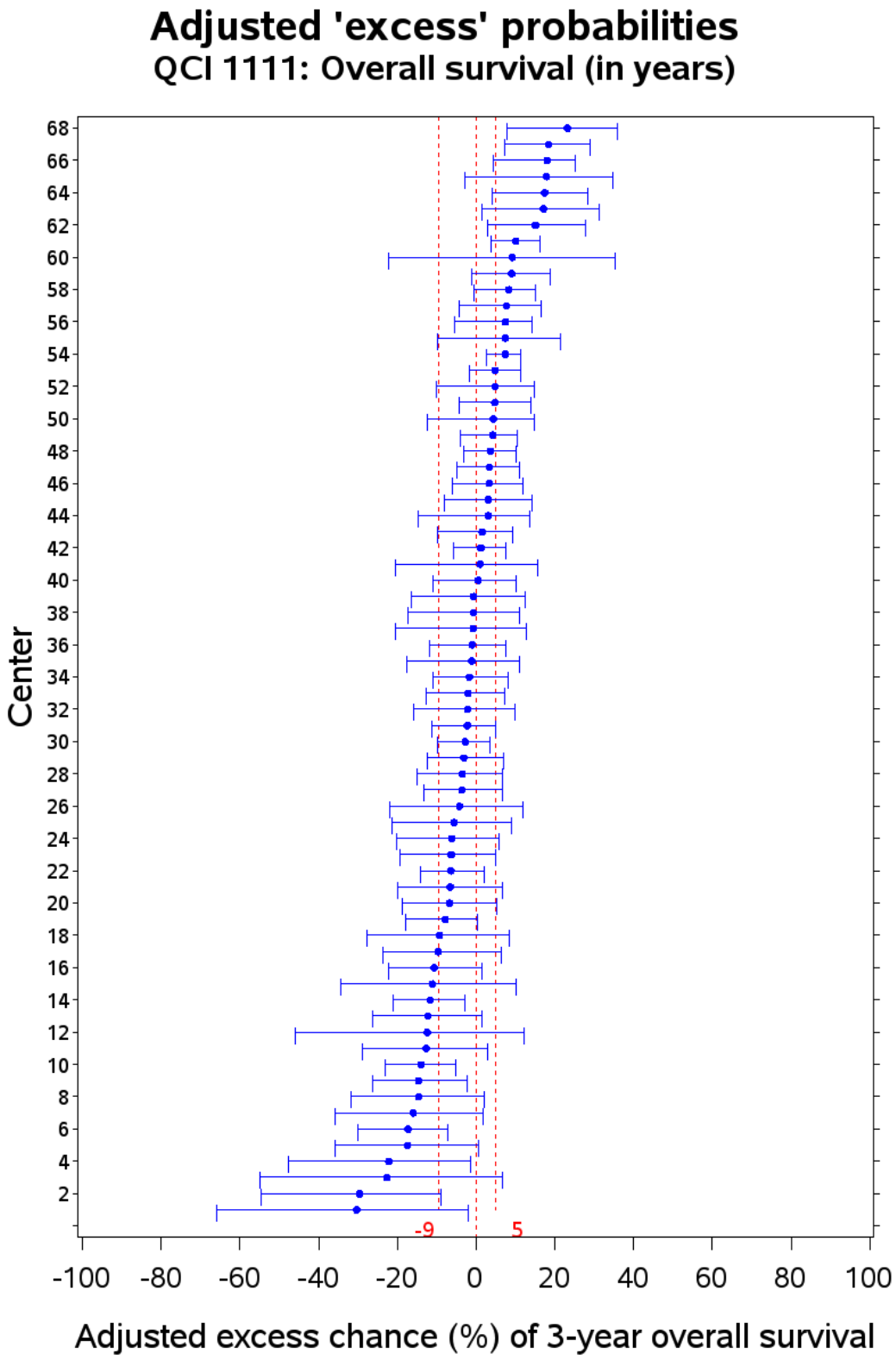
Figure 12 shows the relation between adjusted and unadjusted center effects. From this we learn how the center effects obtained from the fixed-effects adjustment model (blue circles) stay generally close to the unadjusted center effects, and that the center effects obtained from the random effects adjustment model (red squares) are – compared to the unadjusted and fixed-effects adjusted center effects - shrunken towards zero (as anticipated in section 2.1).

Figure 12: Scatter plot of adjusted center effects (from both fixed- and random effects outcome regression) versus unadjusted center effects for QCI 1111 [OS].



Center effects measured as adjusted excess probabilities range between 31% and 22%. Note that these excess chances are automatically constrained between -78% and 22% (since the average center has achieved this QCI in 78% of their patients). The P25 and P75 are located at -9.4% resp. 4.8%, hence the interquartile range is 14%. In the caterpillar plot, we observe no centers performing significantly below P25, i.e. with upper bounds of their confidence intervals below P25. We do observe 2 centers performing significantly above P75, i.e. the lower bound of their confidence interval is higher than P75.

Figure 13: Adjusted excess chance (%) of 3-year overall survival



3.4.2 Tabulated results for the QCIs

In Table 30 and Table 31:

Survey results are median scores over all responses
PROCARE consensus results are: 1: highly relevant, 2, secondary order, 3: not relevant
Visual discriminating ability is scored as: 1: good, 2. some, 3. None

N is the number of eligible patients and %events refers to the percentage of eligible patients with a recorded event of the given type in the database;

- for DSS the first number refers to deaths with local recurrence and/or distant metastasis while the second refers to deaths without reported recurrence.
- for LRFS the first number refers to recorded local recurrences as first events and the second to deaths without reported local recurrences.

Both the DSS and LRFS measure suffer from the fact that follow-up information was quite incomplete. Absence of a reported recurrence might therefore either indicate absence of recurrence or simply missing data. Because of this lack of reliability, DSS and LRFS have not been analyzed in any depth here. For their summary in the table below. deaths recorded without any indication of follow-up have been discarded and are not listed as either cause-specific death for the DSS summary, Specialized competing risks methods allowing for missing cause of death could be applied in this case provided every death is registered and the missing cause is either missing at random or there is some information on how much more likely deaths with than deaths without recurrence are to generate missing follow-up records (and hence missing cause-specific information) [52]. To be eligible for the LRFS variable required that at least one follow-up record was filed.

The final column gives the p-value for testing for between center variation in a mixed effects model with normal random effects.

Table 30: Criteria considered in constructing a QI for outcome of rectal cancer, split by data quality, clinical importance and discriminating ability.

							Data quality			Clinical importance			Discriminating ability		
QCI	**Short name**	**N patients**	**% events**	**N centers**	**N only 0/1**	**N ≥ 5**	**Survey**	**% patients**	**% ≥ 5**	**Survey**	**Predictive ability**	**PROCARE consensus**	**% only 0/1**	**IQR**	**Random effect variance – p**
1111	OS	3103	22	79	5	67	5	88	84	5		1	7	14%	
1112	DSS	2807	7/7	79	1	66	4	85	83	5		1	1		
1113	LRFS	1186	2/8	63	7	49	4	36	61	5		1	14		
1113b	DFS	1234	18	64	3	49	4	37	61	5		1	6	10%	
1227	%grade4_Tox_Preop_RT	544	11	51	9	30	2	16	38	4	0.295	1	29	12%	0.2492
1231	%R0res	2914	87	78	11	67	5	88	84	5	0.0001	1	16	11%	0.0001
1232b	%stoma1year	123	27	37	3	9	4	4	11	3	0.003	2	30	32%	0.0987
1233a	%Leak_PME	556	6	59	27	42	4.5	17	53	4.5	0.0284	1	63	16%	0.0444
1233b	%Leak_TME	1592	6	74	16	54	4	48	68	4	0.0003	1	29	7%	
1234	30d_mort	2919	2	78	35	67	5	88	84	5	0.0001	1	51	4%	0.0019
1234b	%Major_morb	2913	6	78	18	67	4	88	84	3	0.0001	2	26	6%	0.0319
1235	%Perfor	2900	8	78	14	67	4.5	87	84	4	0.0001	2	21	8%	0.0003
1235b	%Pos_Dist_margin	516	2	64	21	27	5	16	34	5	0.5033	1	75	6%	0.2716
1235c	%Pos_CRM	1932	19	73	6	60	5	58	75	5	0.0001	1	10	16%	0.0045
1245	%grade4_Tox_Prostop_CT	128	8	37	6	10	2	4	13	4	0.8356	1	55	5%	
1263	%Late_Tox_RT	1890	0	74	58	61	2	57	76	3	0.9681	2	94	3%	

Table 31: Criteria considered in constructing a QI for process QCIs related to quality of care of rectal cancer, split by data quality, clinical importance and discriminating ability.

							Data quality			Clinical importance			Discriminating ability		
QCI	Short name	N patients	% events	N centers	N only 0/1	N ≥ 5	Survey	% patients	% ≥ 5	Survey	Predictive ability	PROCARE consensus	% only 0/1	IQR	Random effect variance - p
1211	%DocDist	3318	91	80	29	68	4	100	85	4	0.0001	1	42	12%	0.0001
1212	%CT_Preop	322	79	41	13	22	4	10	28	3	0.7458	1.5	57	22%	0.0001
1213	%CEA_Preop	3318	81	80	7	68	4	100	85	2.5	0.0499		10	18%	0.0001
1214	%CEA_Preop_Bowel_Im	2811	98	77	48	67	4	85	84	3	0.4557	2	71	4%	0.0001
1214b	%TRUS_cT12	74	82	29	2	4	4	2	5	3.5	0.4604	2	40	24%	0.094
1214c	%MR_cII/III	103	91	25	4	9	4	3	11	4	0.0853	1	40	19%	0.0967
1215	%Preop_Im	435	61	49	5	25	4	13	31	5	0.3780	1	19	50%	0.0001
1216	%cCRM_rep	1950	26	74	21	62	3	59	78	4	0.0095	1	33	29%	0.0001
1216b	cM0_Acc	634	6	55	18	34	4	19	43	3.5	0.8077	2	51	5%	0.0009
1217	Time_histo-1ther	2687		76		65	4	81	81	2.5	0.9467	2			0.0001
1221	%Preop_RT	1830	74	74	4	60	5	55	75	5	0.0001	1	7	27%	0.0001
1221b	%(C)RT_cRM+	286	81	37	1	14	4	9	18	5	0.0325	1	7	27%	0.0001
1221c	%Preop_RT_cI	344	17	56	12	27	4	10	34	2	0.3072	2	43	18%	0.0002
1224	%Preop_cont_5FU	469	91	49	21	30	4	14	38	2	0.6439	2	68	7%	0.0001
1225	%Completed_preop_RT	1196	97	70	34	51	3	36	64	3	0.7915	2	65	5%	0.0001
1226	%Surg<12w_after_Preop_RT	1123	97	69	33	49	5	34	61	2.5	0.0683	2	66	5%	
1232a	%Defin_ostomy	2945	24	78	2	67		89	84		0.0001	1	3	19%	0.0001
1241	%Adj_Chemo<3m	28	86	14		1	4	1	1	4		1			
1242	%Adj_RT<3m	46	98	23		3	5	1	4	5		1			
1243	%Adj_Chemo<12w	58	93	23		4	4	2	5	3		2			
1244	%Adj_5FU	57	95	23		4	4	2	5	2.5		2			
1271	%Path_Rep_Use	1980	98	69	42	56	5	60	70	1.5	0.3454	3	74	4%	0.0001
1272	%TME_Qual_Rep	1572	68	67	6	55	4	47	69	5	0.0053	1	11	38%	
1273	%Dist_Margin_Rep	2051	86	75	15	63	5	62	79	4.5	0.0012		23	17%	0.0001
1273b	%Dist_Margin_Pos_Rep	2034	91	74	20	63	5	61	79	4	0.0548	3	31	14%	0.0001
1274	#Nodes_Examined	2714		77		64	5	82	80	5	0.0027	1			0.0001
1275	pCRM_mm_Rep	2615	74	77	6	66	5	79	83	5	0.6578	1	9	23%	0.0001
1276	TRG_Rep	1427	75	71	7	53	3	43	66	4	0.3149	1.5	13	39%	0.0001

There is a big diversity in ***eligibility rate*** in the PROCARE database among both process and outcome QCI's. For outcome QCI's we found 5 QCI's having eligibility rates below 30%, 5 between 30%-60% and 5 between 60% and 90% respectively. 1 QCI (QCI 1111) scored higher than 90%. For process QCI's we found eligibility rate to be very low (<30%) in about half of the QCI's (n=12), moderately low rates (between 30%-60%) were found in 7 QCI's, fairly high rates (between 60% and 90%) in 7 QCI's. 2 QCI's (QCI 1211 and 1213) score very high eligibility rates (>90%).
The ***achievement rate*** was uniformly high for all 16 outcome QCI's (mean: 90% +/- sd-dev.:8%) and 25 process QCI's (mean: 85% +/- sd-dev.:11%). As an exception there are 3 process QCI's with very low achievement rate, i.e. QCI's 1216, 1216b and 1221c (with 26%, 6% and 17% achievement rate respectively).

Table 32: Center effect for general (outcome) quality indicators: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level. [NA: not applicable; Center effects for these QCI's were not calculated]

QCI	Short name	IQR	# below P25	# above P75	Prognostic factors
***1111**	OS	14%	0	2	*Age, Age (+ 70 years), BMI, Missing BMI, Gender, Gender – BMI interaction, Gender - Missing BMI interaction, ASA score and cStage (the latter as stratification variable)
1112	DSS	NA	NA	NA	Age, Age (+ 70 years), Gender, ASA score, BMI, Missing BMI and cStage (the latter as stratification variable)
1113	LRFS	NA	NA	NA	Ventral tumor
1113b	DFS	10%	NA	NA	Age, Age (+ 70 years), BMI, Missing BMI, Gender, Gender – BMI interaction, Gender - Missing BMI interaction, ASA score

The clinical importance of all these indicators was rated very high (highest scores) but data quality was only deemed sufficient for QCI 1111 and QCI 1112. QCI 1112 is not included since it reflects a component of survival which is already part of QCI 1111 while - due to generally limited follow-up information on local recurrence and distant metastasis in the PROCARE database - its quality is judged to be weak with at present little hope for an improved signal in that QCI. The general quality indicators all have a high achievement rate (>78%).
QCI 1112b was added later and not considered in this report.

Table 33: Center effect for (process) quality indicators related to diagnosis and staging: inter quartile range, number of centers significantly below P25 or above P75 and univariate prognostic factors retained at the 5% level.

QCI	Short name	IQR	# below P25	# above P75	Prognostic factors
1211	%DocDist	12%	7	0	Age, Missing Age, BMI, Missing BMI, ASA score, cStage, Mode of surgery, Ventral tumor, cCRM positive, cT4, Preoper. Incontinence, Surgical technique
1212	%CT_Preop	22%	3	0	cT4, Preoper. Incontinence, Surgical technique
1213	%CEA_Preop	18%	8	3	Age, Missing Age, BMI, Missing BMI, Tumor level, ASA score, cStage, Mode of surgery, Ventral tumor, cCRM positive, cT4n Preoper. Incontinence, Surgical technique
1214	%CEA_Preop_Bowel_Im	4%	0	0	Age, Missing Age, BMI, Missing BMI, Tumor level, ASA score, cStage, Ventral tumor, cT4, Preoper. incontinence
1214b	%TRUS_cT12	25%	0	0	Missing BMI
1214c	%MR_cII/III	19%	0	0	cStage
1215	%Preop_Im	50%	3	1	BMI, Missing BMI, Tumor level, ASA score, cStage, Mode of surgery, Ventral tumor, cCRM positive, cT4, Preoper. Incontinence, Surgical technique

QCI	Short name	IQR	# below P25	# above P75	Prognostic factors
1216	%cCRM_rep	29%	0	5	Age, Missing Age, BMI, Missing BMI, Gender, Tumor level, cStage, Ventral tumor, Preoper. Incontinence, Surgical technique
1216b	cM0_Acc	6%	0	1	Age, Missing BMI, ASA score, cStage, Ventral tumor, Surgical technique
1217	Time_histo-1ther	14%	1	3	Age, BMI, Missing BMI, Tumor level, cStage, Surgical technique

QCI 1215 was most discriminating between centers (IQR of 50%). Clinical importance was rated highest for this QCI (1215) on both the survey and the PROCARE consensus score. QCI1217 was considered the least important clinical parameter. Only 4 out of 10 QCI's (1211, 1213, 1214 and 1217) reached sufficient data quality in terms of percentage of eligible patients (mean: 91%) and having at least 5 patients/center (mean: 84% of centers). The diagnosis and staging QCI's have a fair achievement rate (>60%) except for 1216 and 1216b (<30%).

Table 34: Center effect for (process) quality indicators related to neoadjuvant treatment: inter quartile range, number of centers significantly below P25 or above P75 and univariate prognostic factors retained at the 5% level.

The QCI indicated by a * is an outcome QCI and has multivariate prognostic factors listed.

QCI	Short name	IQR	# below P25	# above P75	Prognostic factors
1221	%Preop_RT	27%	5	1	Age, Missing Age, BMI, Missing BMI, Gender, Tumor level, ASA score, cStage, Mode of surgery, Ventral tumor, cCRM positive, Preoper. Incontinence, Surgical technique
1221b	%(C)RT_cCRM+	27%	0	0	Age, Gender, ASA score, cStage, Preoper. Incontinence
1221c	%Preop_RT_cl	18	0	0	BMI, Missing BMI, Surgical technique
1224	%Preop_cont_5FU	7	3	0	BMI, Missing BMI, ASA score, cStage
1225	%Completed_preop_RT	5	1	0	Age, Tumor level, Mode of surgery, Ventral tumor, Preoper. Incontinence, Surgical technique
1226	%Surg<12w_after_Preop_RT	5%	0	0	none
***1227**	%grade4_Tox_Preop_RT	12%	0	0	*Age, Gender

Discriminating ability was highest for 1221/1221b (27%). QCI 1221 and 1221b were rated with highest clinical importance for this domain. Except for QCI 1221c (17%) and 1227 (11%) achievement rate was fairly high (on average: 88%). Data quality was quite low (on average 20% eligible patients and 42% of centers with at least 5 patients) except for QCI 1221 (with 55% eligible patients 75% of centers having at least 5 eligible patients)

Table 35: Center effect for (outcome) quality indicators related to surgery: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level.

QCI	Short name	IQR	# below P25	# above P75	Prognostic factors
1231	%R0res	11%	0	2	Age, Gender, cStage, Tumor level, cT4, Mode of surgery
1232a	%Defin_ostomy	19%	1	1	Age, Missing Age, BMI, Missing BMI, Tumor level, ASA score, cStage, Mode of surgery, Ventral tumor, cCRM positive, cT4, Preoper. Incontinence, Surgical technique
1232b	%stoma1year	32%	0	0	Age, Gender
1233a	%Leak_PME	7%	-	0	Age, Gender, Tumor level, ASA score, cStage, Gender – ASA interaction
1233b	%Leak_TME	8%	1	0	Age, Gender, ASA score, Age – ASA interaction
1234	30d_mort	4%	0	0	Age, Missing Age, Gender, ASA score, Mode of surgery, Age – Gender interaction, Missing Age – Gender interaction
1234b	%Major_morb	5%	2	0	Age, Gender, Level, ASA score, Age – ASA score interaction, Gender – ASA score interaction
1235	%Perfor	8%	0	0	Age, Missing Age, Gender, Level, ASA score, cStage
1235b	%Pos_Dist_margin	6%	0	0	Age, Gender, ASA
1235c	%Pos_CRM	16%			Age, Missing Age, Gender, cT4

Discriminating ability for these QCI's was considered very low in the surgery domain except for QCI 1232a. The mean IQR is a low 13%, reflecting a fair degree of homogeneity in surgical practice.
Half of the QCI's (1231, 1232a, 1234, 1234b, 1235) have very good data quality parameters (>84% eligible patients with at least 5 eligible patients/center) At this stage it should be noted that some ambiguity surrounded QCI 1232a. First, not everybody had realized that the current definition implies brackets around the two final statements, so it should read: `Proportion of APR- Hartmann's procedure or (total excision of colon and rectum with definitive ileostomy)'. As a result the definition can refer to colostomy rather than ileostomy in some cases and the achievement rate is no longer unusual. Second, whether or not the colostomy/ileostomy is a sign of good quality treatment according to guidelines, depends on the underlying condition of the patient. In some conditions (e.g. sphincter invasive tumor) this is the outcome per guideline and hence this would become a process indicator. In other instances, this is not the result of targeted therapeutic efforts, but rather an unfortunate outcome of the disease process. We recommend that two different versions would be defined of this QCI, conditional on the tumor level. Conditional on a sphincter invasive tumor this would then become a process indicators and under the complement condition it would be an outcome indicator. Given the current ambiguity in interpretation, this QCI has been eliminated for further use at present.

Table 36: Center effect for (process) quality indicators related to adjuvant treatment: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level.

QCI	Short name	IQR	# below P25	# above P75	Prognostic factors
1241	%Adj_Chemo<3m		-	-	Not estimable
1242	%Adj_RT<3m		-	-	Not estimable
1243	%Adj_Chemo<12w		-	-	Not estimable
1244	%Adj_5FU		-	-	Not estimable
1245	%grade4_Tox_Prostop_CT	5%	0	0	Age, Gender

Lack of data quality (on average 2% of patients are eligible and 6% of centers have at least 5 eligible patients) does not allow further analysis of center effects in the "adjuvant treatment" domain.

Table 37: Center effect for (process) quality indicators related to histopathological examination: inter quartile range, number of centers significantly below P25 or above P75 and multivariate prognostic factors retained at the 5% level.

QCI	Short name	IQR	# below P25	# above P75	Prognostic factors
1271	%Path_Rep_Use	4%	2	0	cStage, cT4
1272	%TME_Qual_Rep	38%	3	2	BMI, Missing BMI, ASA score, cStage, Mode of surgery, cCRM positive, cT4
1273	%Dist_Margin_Rep	17%	4	0	BMI, Missing BMI, Tumor level, ASA score, cStage, Mode of surgery, Ventral tumor, cCRM positive, cT4, Preoper. Incontinence, Surgical technique
1273b	%Dist_Margin_Pos_Rep	14%	2	0	BMI, Missing BMI, Tumor level, ASA score, cStage, Ventral tumor, Preoper. Incontinence, Surgical technique
1274	#Nodes_Examined		3	4	Missing BMI, Tumor level, cStage, Surgical technique
1275	pCRM_mm_Rep	23%	4	1	BMI, Missing BMI, Gender, Tumor level, ASA score, cStage, Mode of surgery, Ventral tumor, cCRM positive, Surgical technique
1276	TRG_Rep	39%	6	0	Age, BMI, Missing BMI, cStage, cCRM positive, Surgical technique

Together with domain 2, diagnosis and staging, (average IQR of 20%), and domain 8, histopathologic examination, is most discriminatory between centers (IQR is 23% on average). Achievement rates were uniformly fair to high (68% and higher) for the QCIs in domain 8. QCIs 1272, 1274 and 1275 were rated with the highest clinical importance score in the survey and PROCARE consensus. Data quality was excellent for 1274 and 1275 (>79% patients eligible and >80% of centers have at least 5 eligible patients).

3.5 SUMMARY POINTS

On the methodology

- **Of the different methods, fixed effects outcome regression, random effects outcome regression and doubly robust propensity score methods, only the former – with Firth correction – fitted consistently well with the data.**
- **Hierarchical methods often shrink results dramatically to the point of eliminating any substantial differences. They work especially well in settings with exchangeable center effects and when protecting the null hypothesis of no center effect is of prime importance. In our setting of feedback aimed at learning for self improvement protecting against false negative results is most important since this would overlook substantial room for improvement.**
- **In light of both arguments above we opt for the fixed effects model with Firth correction as our primary patient-mix-adjustment approach.**

On the current database

- **Limited availability of prognostic factors for center choice and for outcome result in suboptimal adjustments for patient mix. Results shown are therefore likely still confounded to some degree.**

General remarks on the QCI's:

- **There is generally weak evidence of any differences between centers in performance except for the domains 2, diagnosis and staging, and 8, histopathologic examination, where substantial variation is seen, and the domain of surgery where fair homogeneity is found. The discriminating ability is expected to grow with more patient data and narrower confidence intervals. This effect may be counterbalanced by increasing homogeneity in practice.**
- **QCI's were not always defined reliably (ileostomy procedure in QCI 1232a, consistent definitions of denominators (e.g. QCI's 1263 vs.QCI 1245).**
- **There is a big diversity in eligibility rate in the PROCARE database among both process and outcome QCI's.**
- **The achievement rate was generally fair to high for 16 outcome QCI's and 25 process QCI's. 3 process QCI did not reach 30% achievement.**

Domain-specific summary points

- **GENERAL QUALITY INDICATORS**
  - **The general quality indicators all have a high achievement rate (>78%). Only QCI 1111 was found to have high enough data quality but relatively few events were seen and the discriminating ability between centers was low.**
- **DIAGNOSIS AND STAGING**
  - **Only 4 out of 10 QCI's (1211, 1213, 1214 and 1217) reached sufficient data quality in terms of percentage of eligible patients (mean: 91%) with at least 5 patients per center in 84% of centers. QCI 1215 was most discriminating between centers (IQR of 50%).-**
- **NEOADJUVANT TREATMENT**
  - **Data quality was very low (on average 20% of eligible patients with 42% of centers having at least 5 patients) except for QCI 1221 (with 55% eligible patients and 75% of centers having at least 5 eligible patients)**
- **SURGERY**

- **Half of the QCI's (1231, 1232a, 1234, 1234b, 1235) have high data quality parameters; QCI 1232a was found to have been interpreted ambiguously with respect to the ileostomy procedure. The IQR is quite narrow, only 13%, reflecting fair homogeneity in surgical practice**

- **ADJUVANT TREATMENT**
  - **Lack of data quality precludes further analysis.**
- **PALLIATIVE TREATMENT**
  - **There are no data available yet for evaluation.**
- **QUALITY INDICATORS RELATED TO FOLLOW-UP**
  - **The condition 'patients treated with neoadjuvant or adjuvant radio(chemo)therapy' cannot currently be assessed reliably in the PROCARE database because of the low quality of treatment data.**
- **HISTOPATHOLOGIC EXAMINATION**
  - **Together with domain 2, diagnosis and staging, (IQR of 20%), this domain 8 shows most variation between centers (with an IQR of 23%).**

4 AGGREGATED QUALITY INDEXES

4.1 INTRODUCTION

As described in Section 0, we will aggregate well defined sets of QCIs into a combined quality index (QI) as a summary measure of quality which may pertain to a specific domain. In this way, outcome and process QCIs will first be combined into an 'outcome quality index' and a 'process quality index', respectively. In the next section we will also construct aggregated quality indexes per domain of care and eventually an overall global quality Index. In line with our treatment of the original QCIs, our primary measure for evaluation will adjust the quality index for baseline patient characteristics when it involves an outcome QCI, but not otherwise.

4.2 METHODS

4.2.1 Evaluation of quality and reliability of the QCIs

Data Quality is first judged, using the following criteria:

- Median score should be at least 3.5 on the data quality survey.
- At least 80% of all patients in the PROCARE database should be eligible.
- At least 75% of all centers should have 5 eligible patients or more.

4.2.2 Number of dimensions needed

To determine the number of dimensions needed per domain we conduct a principal components analysis on the center-specific excess probabilities. Factor analysis is subsequently applied to find a (set of) high loading QCI(s) per selected dimension. In line with deliverable 6 excess probabilities for process QCIs remained unadjusted while for outcome QCIs excess probabilities were adjusted for baseline patient characteristics.

4.2.3 Clinical importance and discriminating ability of selected QCIs per dimension

For each principal component dimension we now chose the most 'clinically important' and 'discriminating' QCI among the sets suggested by factor analysis, based on the following criteria:

- Median score of at least 4 on the clinical importance survey.
- Significant *p*-value for the predictive ability with regard to overall survival.
- Score 1 (= highly relevant) in the PROCARE consensus.

4.2.4 Construction of the Quality Index

4.2.4.1 All-or-none-score

We have used the all-or-none-score here as the preferred aggregated quality index. The sum score was not calculated as it was shown in deliverable 6 to yield similar or less discrimination between centers.

4.2.4.2 Aggregated scores for domains with few reliable and/ or quality-proof QCIs

When there is at most 1 quality-proof QCI available for a domain, excess probabilities have been derived directly from the representative QCI. Appendix 8 gives more detail for the construction of binary indexes involving survival data for use in an all-or-none-score

4.2.5 Risk adjustment

Below we will adjust aggregated all-or-none scores for baseline patient characteristics when needed. To this end a stepwise procedure has been performed whereby we always adjust for main effects of age (continuous, with breakpoint after 70 years) and gender and then build a main effects model, allowing the following variables to enter: BMI with an indicator of its missingness, and the 9 categorical variables including a category for missingness: cStage, ASA co-morbidity score, level of the tumor, mode of surgery, ventral tumor, cCRM, cT4, pre-operative incontinence, surgical technique. After adding significant main effects in a stepwise forward manner (with significance level 0.05 for entering the model and 0.1 for leaving the model), significant interactions from among the main effects are added, again in a stepwise forward manner (with significance level 0.05 for entering the model and 0.1 for leaving the model).

For the construction of the aggregate score, we follow the method and different steps outlined in Section 0 and depicted in the decision tree of Figure 4. For further detail on the build-up we refer to Appendix 8.

4.3 AN AGGREGATED OUTCOME QUALITY INDEX

4.3.1 Quality and reliability of the QCIs

From Table 39, we first judge the data quality of all outcome QCIs. Based on the minimal quality requirements established in 2.3.3 we take QCIs 1111 [OS], 1231 [%R0res], 1234 [30d_mort], 1234b [%Major_morb], 1235 [%Perfor] and 1235c [%Pos_CRM] to the next step. For QCI 1235c, we have been slightly less stringent, based on 75% of the centers with at least 5 eligible patients we also included this QCI.

QCI 1112 is not included since it reflects a component of survival which is already part of QCI 1111 while - due to generally limited follow-up information on local recurrence and distant metastasis in the PROCARE database (see Section 3.3.5) - its quality is judged to be weak with at present little hope for an improved signal in that QCI.

4.3.2 Concurrent validity of the selected QCIs

To examine the concurrent validity we performed a principal component analysis on the risk-adjusted 'excess' probabilities (from the fixed-effects regression model) for QCIs 1111, 1231, 1234, 1234b, 1235 and 1235c.

The PCA confirms that there is not one single dimension underlying the six indicators. The main dimension (factor 1) accounts for only 41% of the overall variance for the six indicators. The second factor accounts for an additional 20%. Based on the default statistical selection criterion of eigenvalues greater than 1, 3 dimensions are retained which account for almost 80% of the information available among these QCIs.

From a factor analysis we learn that the first dimension is mainly related to survival with a natural different direction for QCI 1111 (overall survival) and QCI 1234 (30-day mortality). Dimension 2 appears to summarize the technical aspects of surgery, and dimension 3 consists mainly of QCI 1234b (major morbidities after surgery), i.e. the immediate patient outcome after surgery.

4.3.3 Clinical importance and discriminating ability of the selected QCIs per dimension

For each dimension we then chose the most 'clinically important' and 'discriminating' QCI, based on the criteria of section 2.3.3 and listed properties in Table 39.

For the first dimension, QCIs 1111 and 1234 both score very well on clinical importance, but QCI 1111 is much more discriminative between centers and is therefore selected for use in the outcome QI.

For the second dimension, both QCI 1231 and 1235b score very well on clinical importance and on discriminating ability, but since there are more centers with at least 5 eligible patients for QCI 1231, this QCI is selected for use in the outcome QI.

In the third dimension there is only one QCI (1234b) which does not score very well on clinical importance, but is selected anyway for use in the outcome QI.

4.3.4 Constructing the quality index (QI)

The outcome QI will hence be constructed using the following QCIs:

- QCI 1111: Overall survival
- QCI 1231: Proportion of R0 resections
- QCI 1234b: Postoperative major surgical morbidity with reintervention under narcosis after radical surgical resection.

4.3.4.1 "All-or-none" score

The "all-or-none" score for a patient indicates whether this patient reaches patient-level benchmarks for all QCIs for which it is eligible, in this case:

- whether he/she survived 3-years since incidence of rectal cancer,

For patients with a follow-up of less than 3 years, a model-based multiple imputation technique is used to construct the "all or none" score and corresponding confidence limits. For more detail, see Appendix 8.

- whether he/she had an R0 resection, and
- whether he/she did not have postoperative major surgical morbidity with reintervention under narcosis after radical surgical resection.

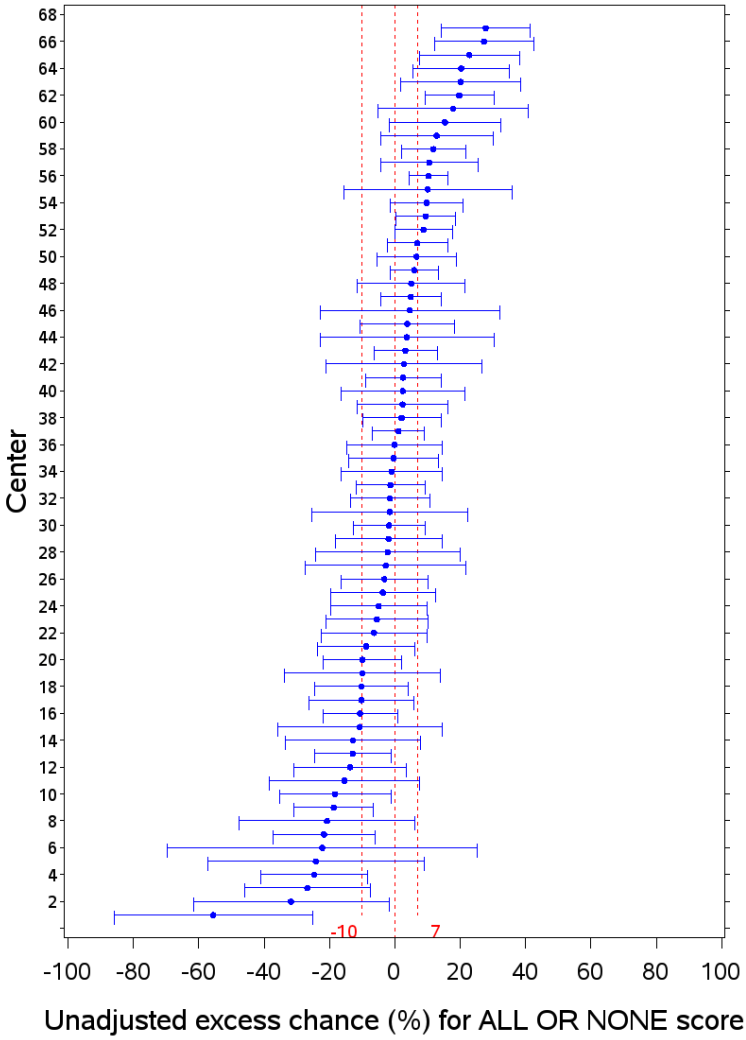
A first assessment of the discriminating ability between centers of this "all-or-none" score is made by looking at the caterpillar plot of the excess probability in Figure 14 and considering the statistical summary measures in Table 38. In this caterpillar plot we find one center performing 'significantly' below P25 and four centers performing significantly above P75.

4.3.4.2 Simple averaging

We computed the simple average of risk-adjusted center-specific excess probabilities of having QCI 1111, 1231 and 1234b into one global QI, statistical summary measures are provided in Table 38.

This amounts to averaging the excess chance of achieving each QCI over its eligible patients.

Figure 14: Caterpillar plot of the unadjusted center-specific 'excess' proportion of patients for whom the "all-or-none" outcome score was met.



4.3.4.3 *Decision*

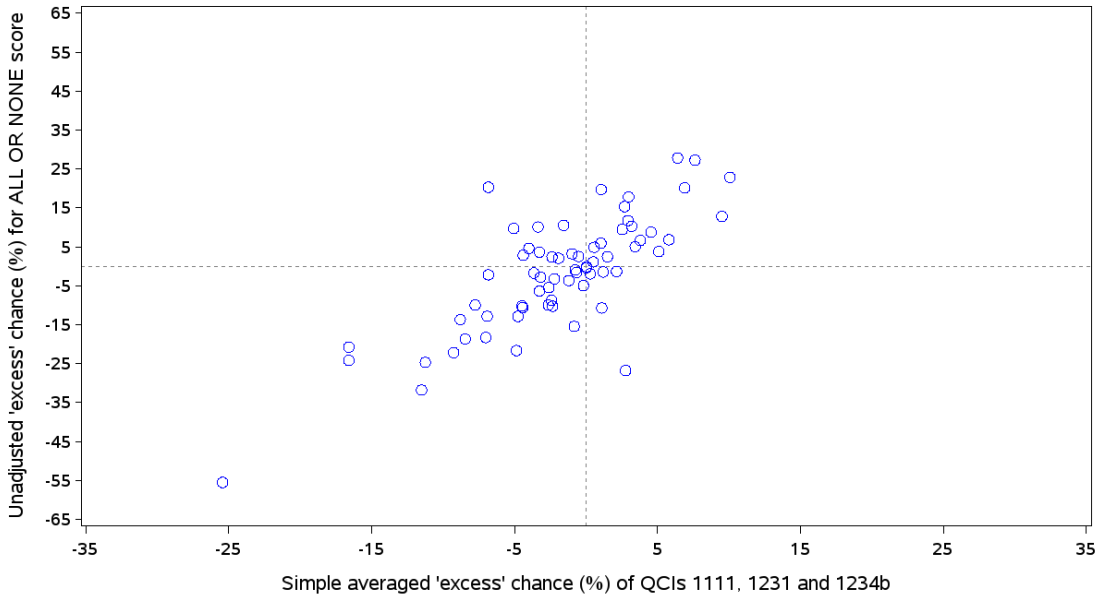
From Table 38, the all-or-none score appears more discriminating between centers than the simple average. Since it is also easier to interpret, we chose this as our quality index.

Table 38: Statistical summary measures for the "all or none" QI and simple averaged QI, both based on QCIs 1111 [OS], 1231 [%R0res] and 1234b [%Major_morb].

QI	Min	P25	Median	P75	Max	IQR
"All or none" (%)	-55	-10	-1	7	28	17
Simple average (%)	-25	-4	-1	2	10	7

Both QIs further show a relatively strong correlation in Figure 15, with a Pearson correlation coefficient = 0.78, suggesting that both may ultimately express a similar quality measure.

Figure 15: Scatter plot comparing the "all-or-none" score to the simple averaged QI for QCIs 1111 [OS], 1231 [%R0res] and 1234b [%Major_morb].



We decide to use the all-or-none score of QCIs 1111, 1231 and 1234b as outcome QI. The funnel plot for the outcome QI reveals systematic variation between centers in this (unadjusted) score.

4.3.5 Adjusting the outcome QI for patient characteristics

Forward stepwise model building leads to a model adjusting for main effects of age (with a different slope before and after the breakpoint of 70 years), gender, cStage, ASA score, cT4, mode of surgery and interactions between age and ASA as well as cStage and Mode-of-Surgery'.

The caterpillar plot of adjusted excess probabilities for the outcome "all or none" score is shown Figure 16. In this caterpillar plot we see 1 center performing 'significantly' below P25 (-9%) and 3 above P75 (8%).

Figure 16: Caterpillar plot of the adjusted excess probabilities for the outcome "all or none" score.

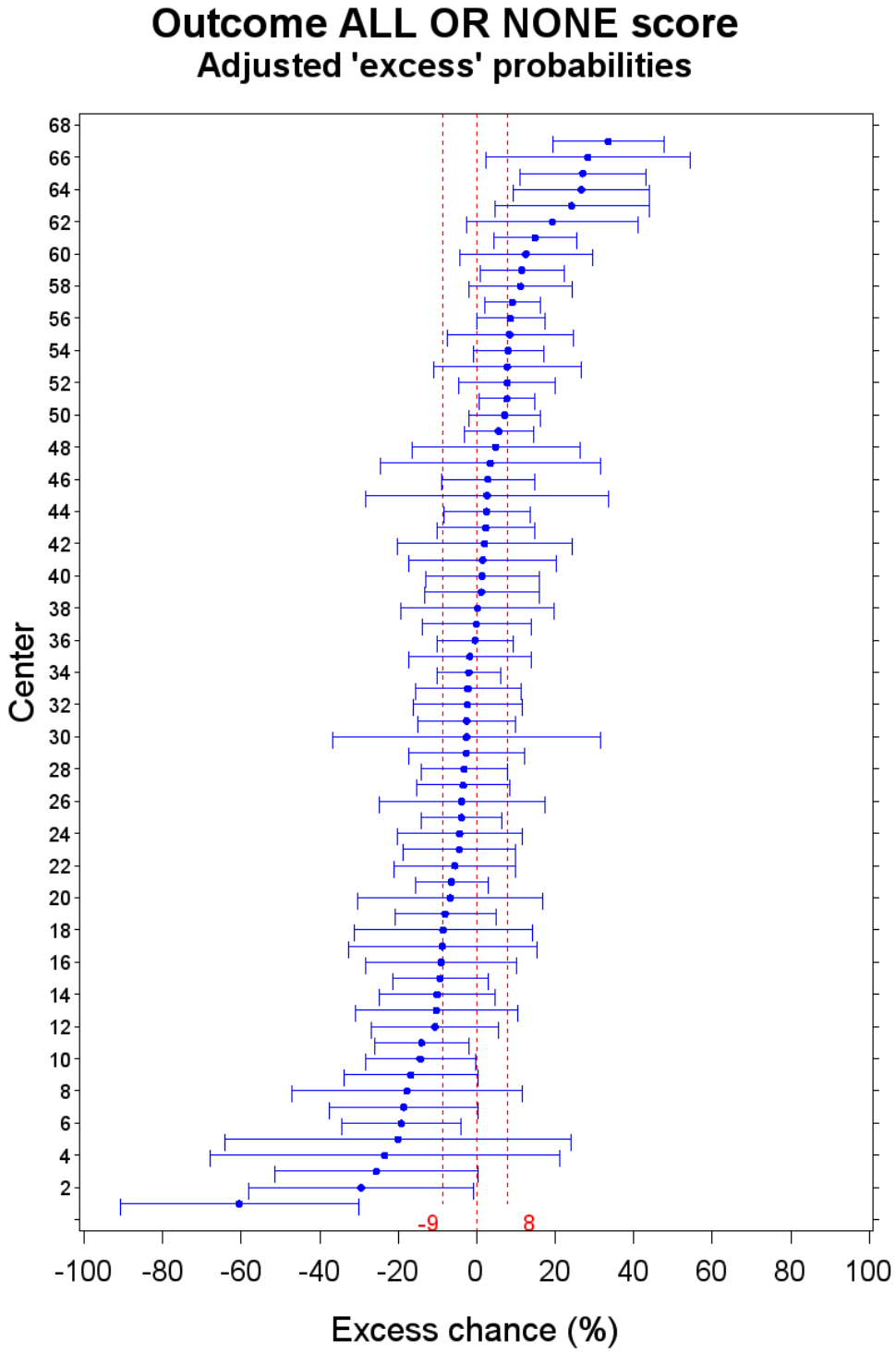


Table 39: Criteria considered in constructing a QI for outcome of rectal cancer, split by data quality, concurrent validity (PCA), clinical importance and discriminating ability

							Data quality				PCA		Clinical importance				Discriminating ability				
QCI	Short name	N patients	% events	N centers	N only 0/1	N >= 5	Survey	% patients	% >= 5	Include	Dimension	Loading	Survey	Predicitive ability	PROCARE consensus	Include	Visually	% only 0/1	IQR	Random effects variance	Include
1111	*OS*	3103	22%	79	5	67	5	88	84	x	1	-0.74	5		1	x		7	14%		x
1112	*DSS*	2807	7%/7%	79	1	66	4	85	83				5		1			1			
1113	*LRFS*	1186	2%/8%	63	7	49	4	36	61				5		1			14			
1113b	*DFS*	1234	18%	64	3	49	4	37	61				5		1			6	10%		
1227	*%grade4_Tox_Preop_RT*	544	11%	51	9	30	2	16	38				4	0.295	1		2	29	12%	0.2492	
1231	*%R0res*	2914	87%	78	11	67	5	88	84	x	2	0.94	5	0.0001	1	x	2	16	11%	0.0001	x
1232b	*%stoma1year*	123	27%	37	3	9	4	4	11				3	0.003	2		3	30	32%	0.0987	
1233a	*%Leak_PME*	556	6%	59	27	42	4.5	17	53				4.5	0.0284	1		2	63	16%	0.0444	
1233b	*%Leak_TME*	1592	6%	74	16	54	4	48	68				4	0.0003	1		3	29	7%		
1234	*30d_mort*	2919	2%	78	35	67	5	88	84	x	1	0.83	5	0.0001	1	x	3	51	4%	0.0019	
1234b	*%Major_morb*	2913	6%	78	18	67	4	88	84	x	3	0.99	3	0.0001	2	x	3	26	6%	0.0319	x
1235	*%Perfor*	2900	8%	78	14	67	4.5	87	84	x	1	0.75	4	0.0001	2		3	21	8%	0.0003	
1235b	*%Pos_Dist_margin*	516	2%	64	21	27	5	16	34				5	0.5033	1		3	75	6%	0.2716	
1235c	*%Pos_CRM*	1932	19%	73	6	60	5	58	75	x	2	-0.92	5	0.0001	1	x	2	10	16%	0.0045	
1245	*%grade4_Tox_Prostop_CT*	128	8%	37	6	10	2	4	13				4	0.8356	1		3	55	5%		
1263	*%Late_Tox_RT*	1890	0%	74	58	61	2	57	76				3	0.9681	2		3	94	3%		

4.4 AN AGGREGATED PROCESS QUALITY INDEX

4.4.1 Quality and reliability of the QCIs

From Table 39, we first judge the data quality of all QCIs. Based on the criteria established in 2.3.3 we decide to take QCIs 1211, 1213, 1214, 1217, 1221, 1232a, 1271, 1272, 1273, 1273b, 1274 and 1275 to the next step, being less stringent for QCIs 1221, 1271, 1272, 1273 and 1273b with regard to one or more of the proposed criteria.

4.4.2 Concurrent validity of selected QCIs

A principal component analysis is preformed on the unadjusted 'excess' probabilities for QCIs 1211, 1213, 1214, 1217, 1221, 1232a, 1271, 1272, 1273, 1273b, 1274 and 1275.

The PCA confirms that there is not one single dimension underlying the twelve indicators. The main dimension accounts for only 20% of the overall variance for the twelve indicators. The second factor accounts for an additional 19%. Based on the default statistical selection criterion of eigenvalues greater than 1, 4 dimensions are retained and account for 63% of the information available among these QCIs.

4.4.3 Clinical importance and discriminating ability of selected QCIs per dimension

For each dimension we now chose the most 'clinically important' and 'discriminating' QCI, based on the criteria established in 2.2.3:

For the first dimension, QCIs 1273 and 1275 both score very well on clinical importance and discriminating behavior between centers. Since QCI 1273 loads stronger on the first dimension, this QCI is selected as the first component of the process QI.

For the second dimension, both QCI 1211, 1221 and 1232a score very well on clinical importance. QCI 1211 is discarded because of the high percentage of centers with 100% documented distance from the anal verge, which does not allow discriminating between these centers easily. Finally, QCI 1232a is selected as the second component in the process QI, because of perceived problems with QCI 1221: one could not always distinguish between a missing value for RT and the actual absence of the RT treatment. In the third dimension, only QCI 1274 scores very well on clinical importance, and is therefore selected as third component for the process QI.

Finally, the fourth dimension is dominated by QCI 1272, but since only 47% of the patients are eligible for this QCI it is not considered for the QI. Since on the other hand QCI 1217 scores well on both data quality and discriminating ability, it is selected as fourth component for the process QCI.

4.4.4 Construction of the quality index (QI)

The process QI will be constructed using the following QCIs:

- QCI 1217: Time between first histopathological diagnosis and first treatment.
- QCI 1232a: Proportion of APR- Hartmann's procedure or total excision of colon and rectum with definitive ileostomy.
- QCI 1273: Distal tumor-free margin mentioned in the pathology report.
- QCI 1274: Number of lymph nodes examined.

4.4.4.1 *"All or none" score*

The "all or none" score for a patient is an indicator expressing whether the patient reaches patient-level benchmarks, in this case:

- whether the time between first histopathologic diagnosis and first treatment was no more than 30 days,
- whether he/she did not have APR- Hartmann's procedure or total excision of colon and rectum with definitive ileostomy,
- whether he/she had the distal tumor-free margin mentioned in the pathology report, and
- whether there were at least 10 lymph nodesl examined.

A first assessment of the discriminating ability between centers of this "all or none" score is done by inspecting the caterpillar plot of *unadjusted* excess probabilities in Figure 18 and considering the statistical summary measures in Table 40. In this caterpillar plot we find four centers performing 'significantly' below P25 and five centers performing significantly above P75.

4.4.4.2 *Simple averaging*

We compute the simple average of unadjusted center-specific excess probabilities of having QCIs 1232a, 1273, for QCI 1217 the unadjusted excess number of days between first histopathological diagnosis and first treatment (scaled between -1 and 1 by dividing by 50) and for QCI 1274 the unadjusted excess number of lymph nodes examined (scaled between -1 and 1 by dividing by 10) into one global QI, statistical summary measures are provided in Table 40.

Table 40: Statistical summary measures for the "all or none" QI and simple averaged QI, both based on QCIs 1217 [Time_histo-1ther], 1232a [%Defin_osotmy], 1273 [%Dist_Margin_Pos_Rep] and 1274 [#Nodes_Examined].

QI	Min	P25	Median	P75	Max	IQR
"All or none" (%)	-29	-11	-1	17	64	28
Simple average	-21	-5	-2	7	22	12

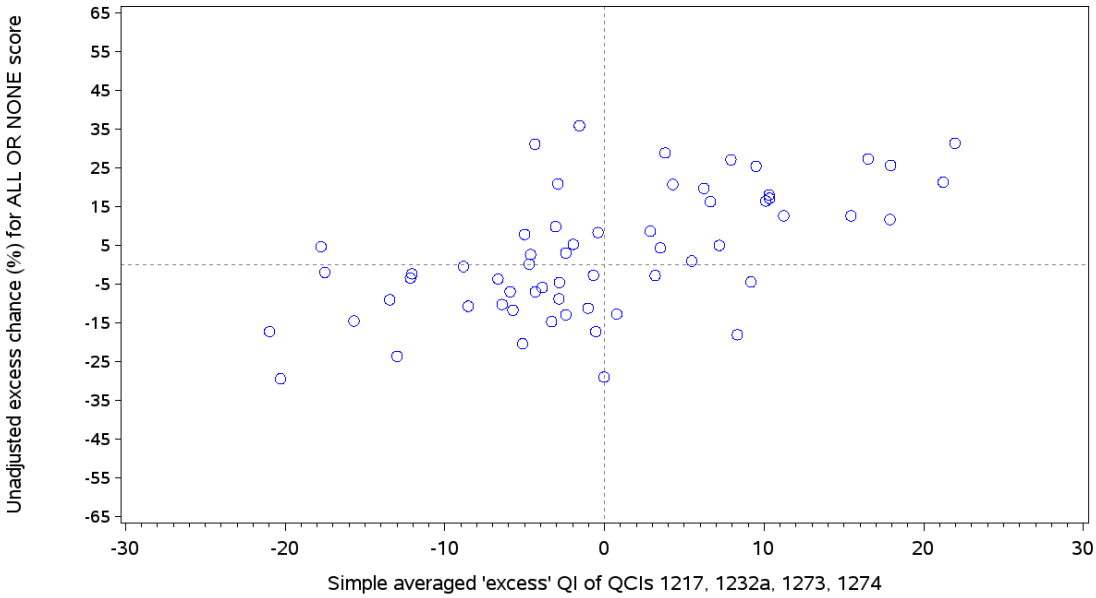
4.4.4.3 *Decision*

When comparing the two methods for combining QCIs 1217, 1232a, 1273 and 1274, we consider especially the discriminating ability which can be examined in Table 40.

Since the simple average is on a different scale (as one of its components is on a different scale) it is not possible to directly compare it with the "all or none" score based on this Table. The Pearson correlation between the two QIs is now 0.61 (Figure 17).

[1] We have chosen cutpoint 10 rather than 12 (used in the PROCARE II report) as a conservative compromise between common cutpoints available in the literature. This avoids a greater dominance in failures in the all-or-none score.

Figure 17: Scatter plot comparing the "all or none" score to the simple averaged QI for QCIs 1217 [Time_histo-1ther], 1232a [%Defin_ostomy], 1273 [%Dist_Margin_Pos_Rep] and 1274 [#Nodes_Examined].



The choice between the two measures is a matter of judgment, and preference is eventually given to the QI that is best interpretable, i.e. the "all or none" score which will lead to an unadjusted excess probability in this instance.

DECISION: Use the "all or none" score of QCIs 1217 [Time_histo_1ther] (with no more than 30 days between diagnosis and first treatment as threshold), 1232a [%Defin_ostomy], 1273 [%Dist_Margin_Rep] and 1274 [#Nodes_Examined] (with at least 10 lymph nodes to be examined as threshold) as process QI.

Over the centers this score is on average obtained in 40% of the patients. As for the other QI's (data not shown), the funnel plot reflected systematic variation between centers. This is also seen on the caterpillar plot: four centers fall significantly below the P25 (-11%) and 5 above the P75 (+17%) excess probability.

Figure 18: Caterpillar plot of the unadjusted center-specific 'excess' proportion of patients for whom the "all or none" process score was met.

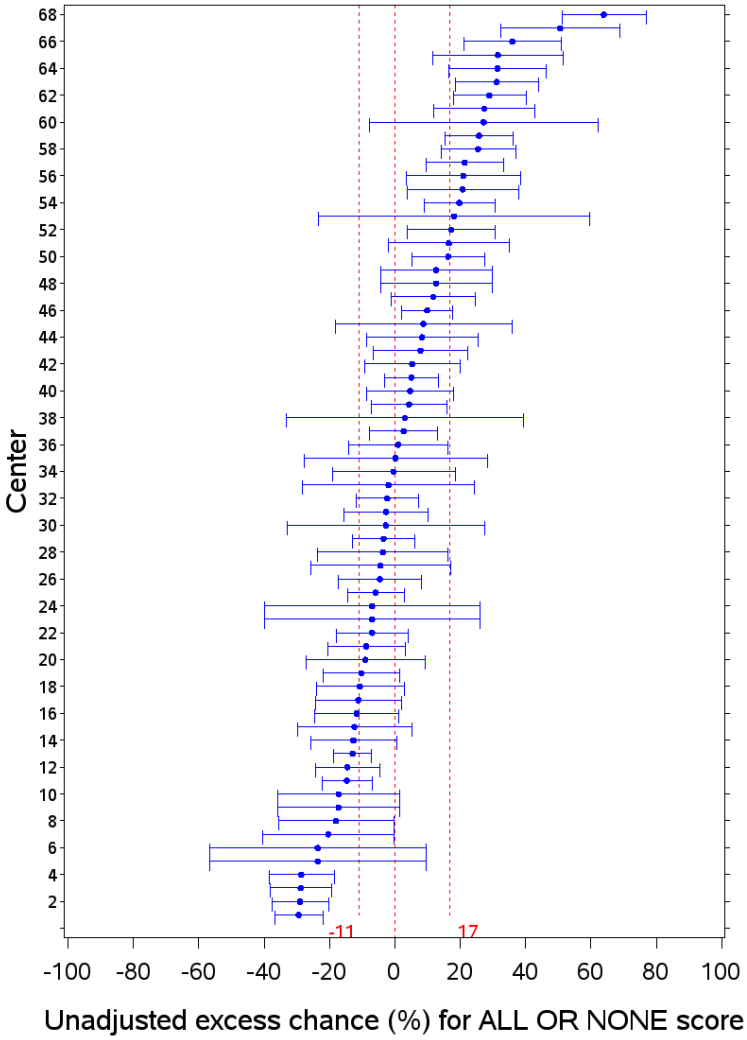


Table 41 (next page): Criteria considered in constructing a QI for process QCIs related to quality of care of rectal cancer, split by data quality, concurrent validity (PCA), clinical importance and discriminating ability

							Data quality				PCA		Clinical importance				Discriminating ability				
QCI	Short name	N patients	% events	N centers	N only 0/1	N >= 5	Survey	% patients	% >= 5	Include	Dimension	Loading	Survey	Predicitive ability	PROCARE consensus	Include	Visually	% only 0/1	IQR	Random effects variance	Include
1211	*%DocDist*	3318	91%	80	29	68	4	100	85	x	2	0.7	4	0.0001	1	x	1	42	12%	0.0001	
1212	*%CT_Preop*	322	79%	41	13	22	4	10	28				3	0.7458	1.5		2	57	22%	0.0001	
1213	*%CEA_Preop*	3318	81%	80	7	68	4	100	85	x	2	0.84	2.5	0.0499			1	10	18%	0.0001	
1214	*%Preop_Bowel_Im*	2811	98%	77	48	67	4	85	84	x	3	0.49	3	0.4557	2		2	71	4%	0.0001	
1214b	*%TRUS_cT12*	74	82%	29	2	4	4	2	5				3.5	0.4604	2		2	40	24%	0.094	
1214c	*%MR_cII/III*	103	91%	25	4	9	4	3	11				4	0.0853	1		2	40	19%	0.0967	
1215	*%Preop_Im*	435	61%	49	5	25	4	13	31				5	0.378	1		1	19	50%	0.0001	
1216	*%cCRM_rep*	1950	26%	74	21	62	3	59	78				4	0.0095	1		1	33	29%	0.0001	
1216b	*cM0_Acc*	634	6%	55	18	34	4	19	43				3.5	0.8077	2		2	51	5%	0.0009	
1217	*Time_histo-1ther*	2687		76		65	4	81	81	x	1 / 4	-0.48 / 0.5	2.5	0.9467	2					0.0001	x
1221	*%Preop_RT*	1830	74%	74	4	60	5	55	75	x	2	0.73	5	0.0001	1	x	1	7	27%	0.0001	
1221b	*%(C)RT_cCRM+*	286	81%	37	1	14	4	9	18				5	0.0325	1		2	7	27%	0.0001	
1221c	*%Preop_RT_cI*	344	17%	56	12	27	4	10	34				2	0.3072	2		2	43	18%	0.0002	
1224	*%Preop_cont_5FU*	469	91%	49	21	30	4	14	38				2	0.6439	2		2	68	7%	0.0001	
1225	*%Completed_preop_RT*	1196	97%	70	34	51	3	36	64				3	0.7915	2		2	65	5%	0.0001	
1226	*%Surg<12w_after_Preop_RT*	1123	97%	69	33	49	5	34	61				2.5	0.0683	2		3	66	5%		
1232a	*%Defin_ostomy*	2945	24%	78	2	67		89	84	x	2	0.53		0.0001	1	x	1	3	19%	0.0001	x
1241	*%Adj-Chemo<3m*	28	86%	14		1	4	1	1				4		1						
1242	*%Adj_RT<3m*	46	98%	23		3	5	1	4				5		1						
1243	*%Adj_Chemo<12w*	58	93%	23		4	4	2	5				3		2						
1244	*%Adj_5FU*	57	95%	23		4	4	2	5				2.5		2						
1271	*%Path_Rep_Use*	1980	98%	69	42	56	5	60	70	x	3	0.62	1.5	0.3454	3		2	74	4%	0.0001	
1272	*%TME_Qual_Rep*	1572	68%	67	6	55	4	47	69	x	4	0.85	5	0.0053	1		1	11	38%		
1273	*%Dist_Margin_Rep*	2051	86%	75	15	63	5	62	79	x	1	0.88	4.5	0.0012		x	1	23	17%	0.0001	x
1273b	*%Dist_Margin_Pos_Rep*	2034	91%	74	20	63	5	61	79	x	1	0.84	4	0.0548	3		2	31	14%	0.0001	
1274	*#Nodes_Examined*	2714		77		64	5	82	80	x	3	0.77	5	0.0027	1	x				0.0001	x
1275	*pCRM_mm_Rep*	2615	74%	77	6	66	5	79	83	x	1	0.59	5	0.6578	1	x	1	9	23%	0.0001	
1276	*TRG_Rep*	1427	75%	71	7	53	3	43	66				4	0.3149	1.5		1	13	39%	0.0001	

4.5 AGGREGATED QUALITY INDEXES PER DOMAIN

4.5.1 Overview of results

We now construct a quality index for each of 8 separate domains of care from their available QCIs as indicated in the table below. For the construction, we return to the statistical-clinical decision tree described in section 2.3.3.3. The resulting all-or-none scores are briefly described inTable 42. Outcome QCIs retained there are in bold and imply the primary QI measure will be a center-specific excess probability which has been adjusted for baseline patient characteristics. NAs show up as we could not assess performance of the centers for domains 5 (not enough eligible patients), 6 and 7 (insufficient data to date).

Table 42: Summary of QCIs per domain, incl. the domain name, the number of outcome (#O) and process (#P) indicators available and the QCIs retained per domain.[13] The last column shows the average achievement rate over the centers for the derived quality index which is an "all or none" score.

Domain	#O	#P	QCIs in QI	achieved
1 General quality indicators	4(+1)	0	1111 (OS)	72 %[14]
2 Diagnosis and staging	0	10(+1)	1211(%DocDist), 1217 (Time_histo-1ther)	60 %
3 Neoadjuvant treatment	1	6	1221 (%Preop_RT)	71 %
4 Surgery	9	1	1231 (%R0res), 1233b (%Leak_TME), 1232a (%Defin_ostomy)	70 %
5 Adjuvant treatment	1	4	NA	NA
6 Palliative treatment	0	0(+1)	NA	NA
7 Follow-up	1	0(+1)	NA	NA
8 Histopathologic examination	0	7	1273 (%Dist_Margin_Pos_Rep), 1274 (#Nodes_Examined)	62%
Total number of QCIs:	16 (+1)	28 (+3)		

For each of the obtained scores we show caterpillar plots of adjusted center-specific excess probabilities as soon as the quality index contains an outcome QCI and unadjusted excess probabilities otherwise. In the former case we also list the variables adjusted for in the model. We end with a table summarizing primary results over the different domains. Further detail can be found in Appendix 8.

4.5.1.1 DOMAIN 1`General quality indicators'

Since the single QCI 1111 is retained for this domain QI corresponding excess probabilities for surviving at least 3 years have been derived previously and are shown in section 3.4.1.3.

4.5.1.2 DOMAIN 2 `Diagnosis and staging'

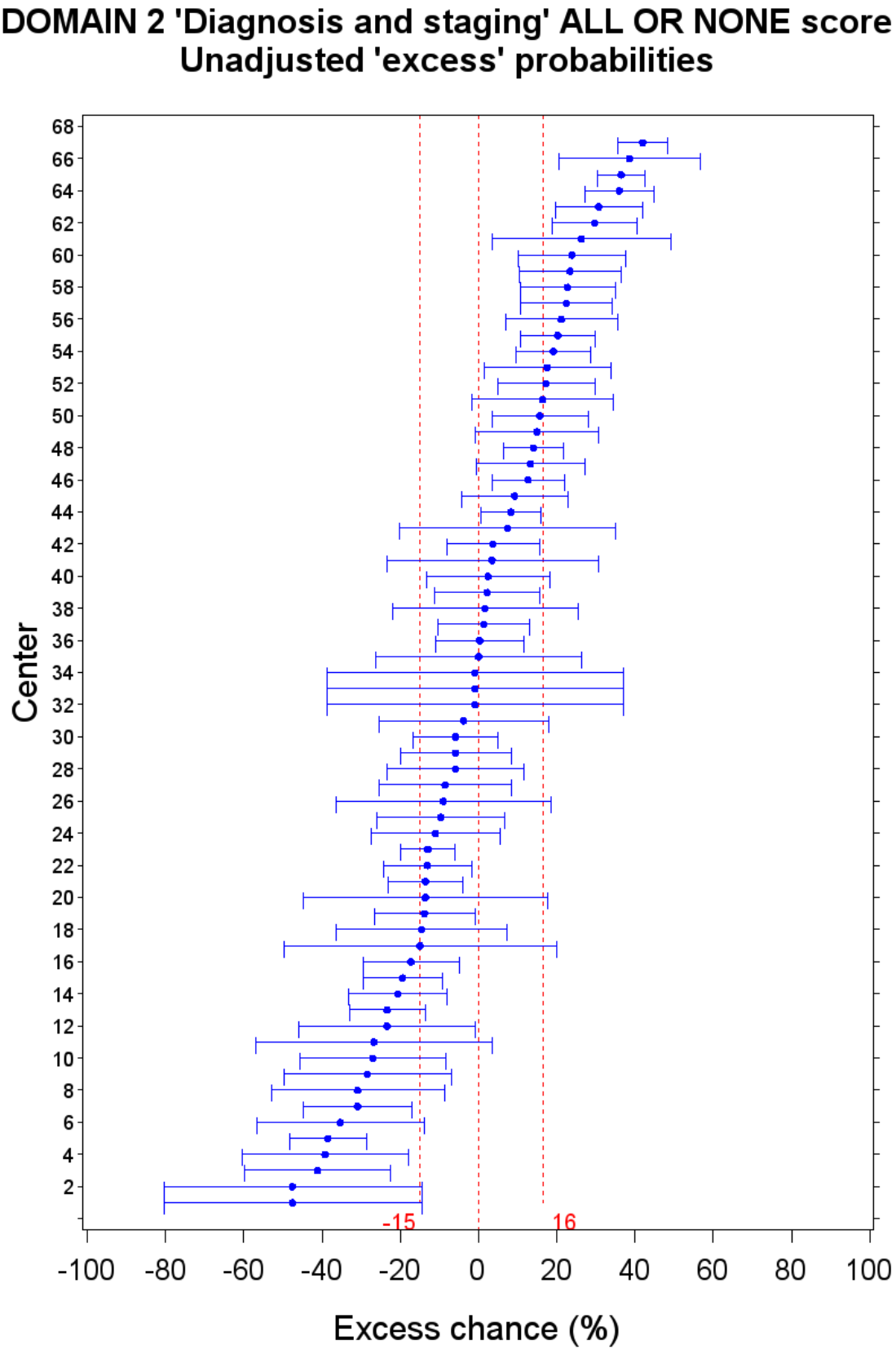
Since this index aggregates process QCIs, we show the unadjusted results. On average over the centers, it is achieved in 59.7% of the patients. In the caterpillar plot, we observe that four centers perform significantly below P25 (-15%) and six perform significantly above P75 (+16%). A multivariate regression model found significant main effects of age (with a different slope before and after the breakpoint of 70 years), gender, Tumor Level, cT4 and cStage with the 2 interaction effects "Age by Gender" and "Gender by cStage".

---

13 Notation: Y(+n): The number of indicators Y is given together with the number of recently defined new indicators n

14 Based on imputed score will be replaced by direct Cox model based value

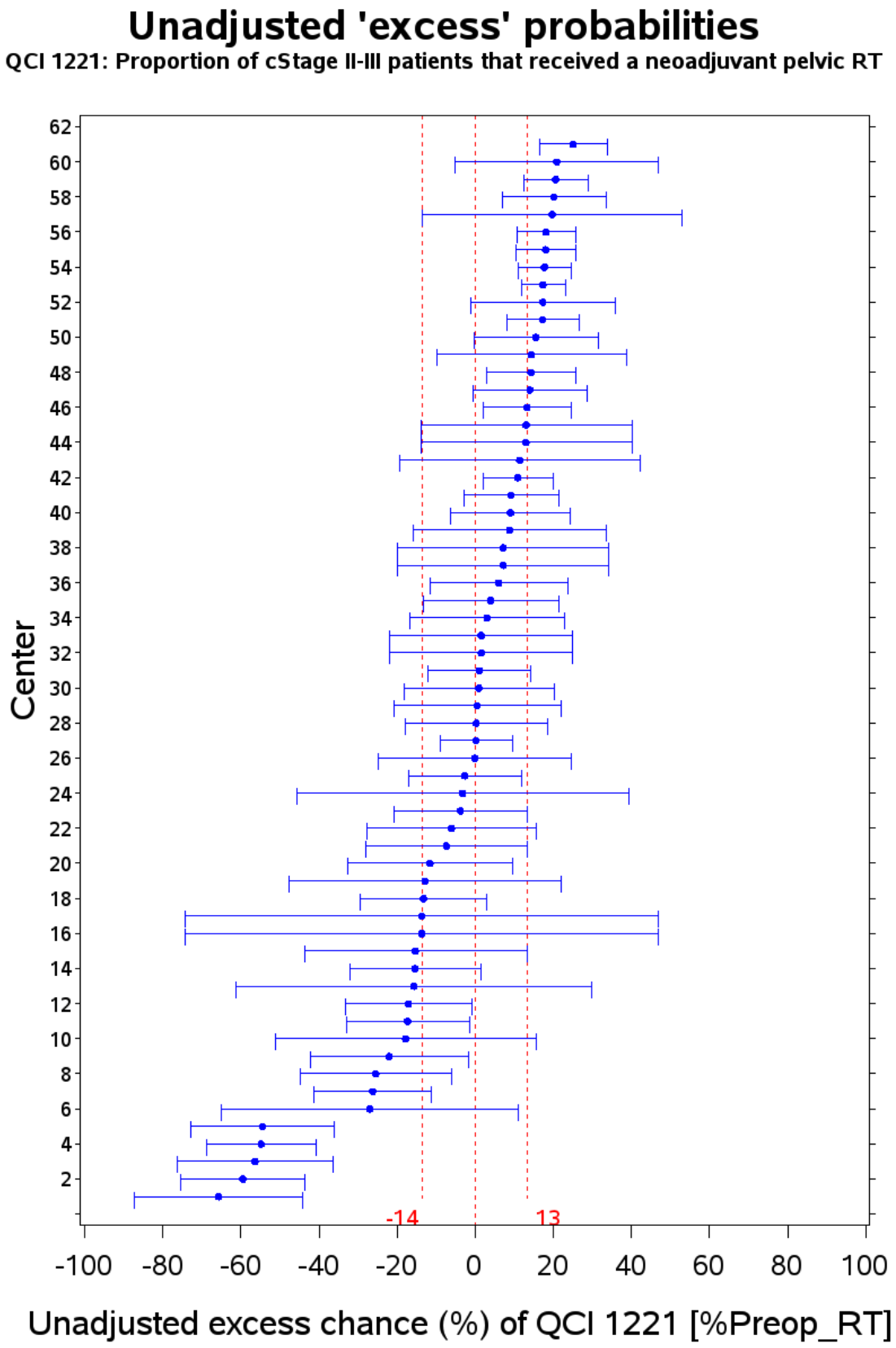
Figure 19: Caterpillar plot of the unadjusted center-specific 'excess' probability for the "all-or-none" score for domain 2 'Diagnosis and staging'



4.5.1.3 DOMAIN 3 `Neoadjuvant treatment'

A single process QCI has been retained here: QCI 1221 `Proportion of cStage II - III patients that received neoadjuvant pelvic RT'. On average over the centers, it is achieved in 74% of the patients. In the caterpillar plot, we observe that five centers perform significantly below P25 (-14%) and one that performs significantly above P75 (+13%).

Figure 20: Caterpillar plot of the unadjusted center-specific 'excess' probabilities of QCI 1221 (as an index for domain 3 `Neoadjuvant treatment')

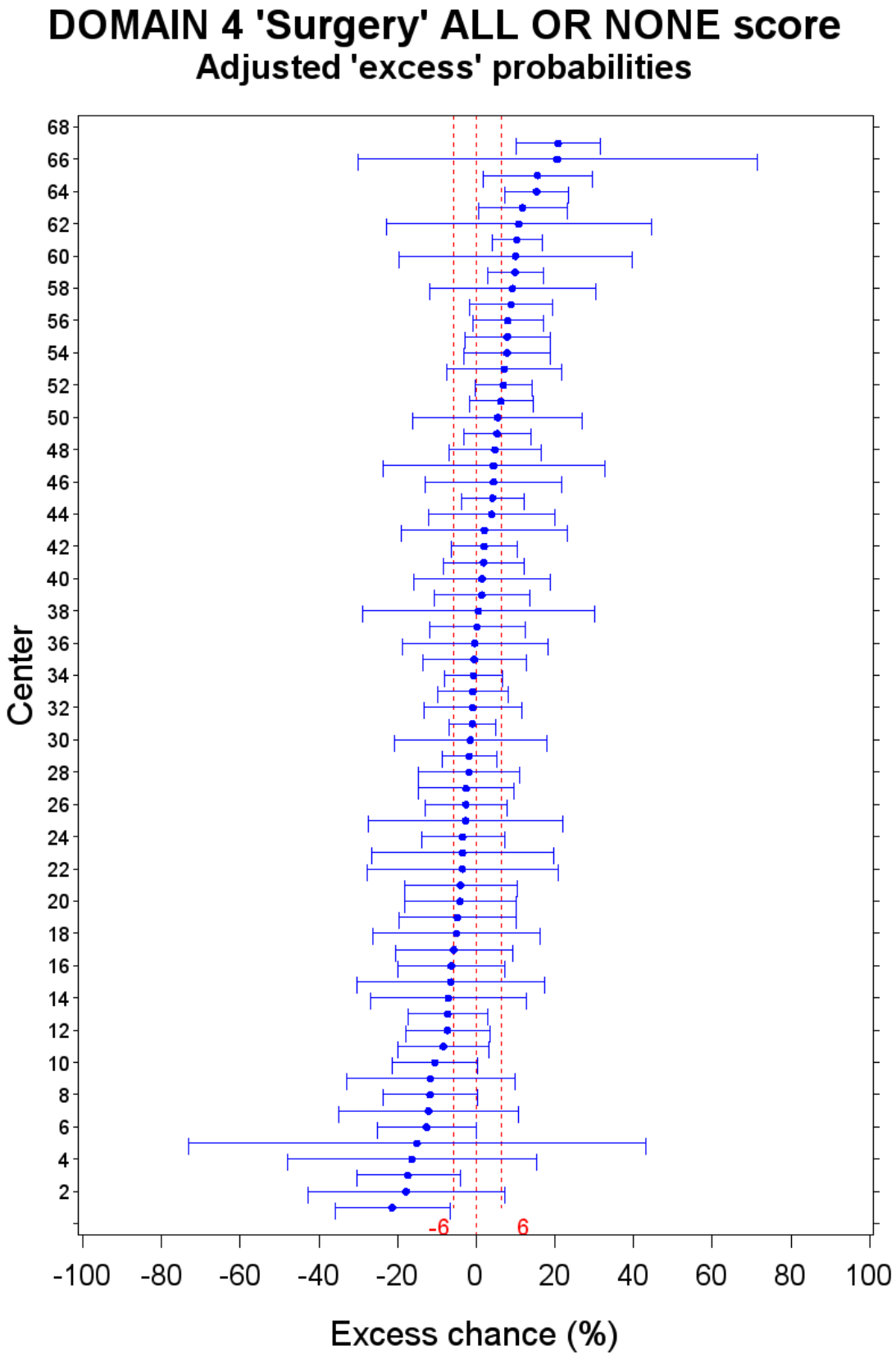


4.5.1.4 DOMAIN 4 `SURGERY'

On average over the centers, this "all or none" score has been obtained in 70% of patients. The model has adjusted for main effects of age (with a different slope before and after the breakpoint of 70 years), gender, Tumor level, surgery, mode of surgery, cT4, cCRM, cStage, Ventral Tumor, ASA score and 2 interactions "ASA score by Mode of surgery" and "cStage by ventral Tumor".

In the caterpillar plot of adjusted excess probabilities, we find one center performing 'significantly' below P25 (-6%) and two centers significantly above P75 (6%).

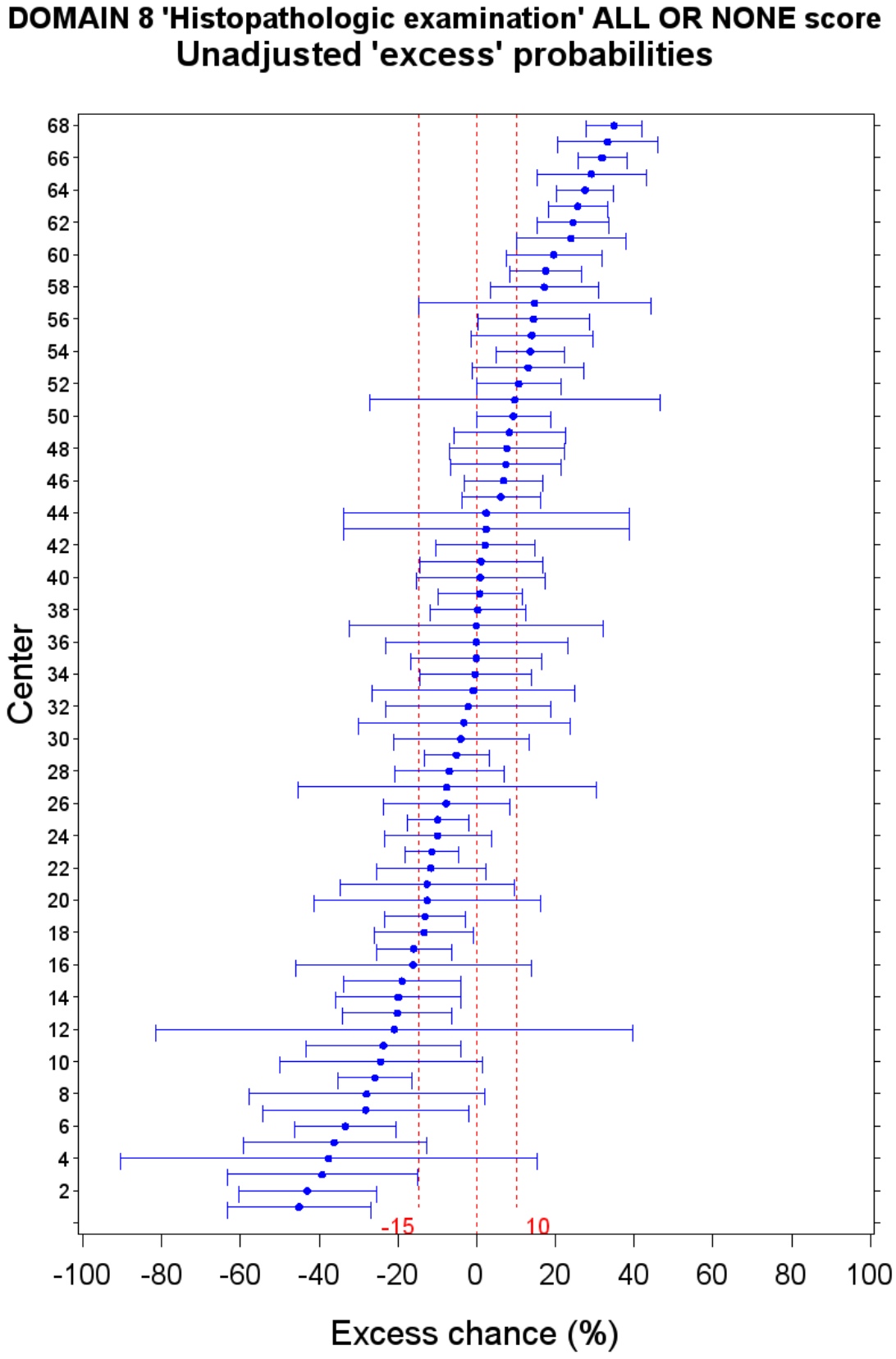
Figure 21: Caterpillar plot of the adjusted center-specific 'excess' probability for the "all-or-none" score for domain 4 'Surgery'



4.5.1.5 DOMAIN 8 `Histopathologic examination'

On average over the centers, this "all or none" score has been obtained in 61.7% of patients. Unadjusted center effects are shown in the caterpillar plot (Figure 22), where we find 5 centers performing 'significantly' below P25 (-15%) and 8 above P75 (+ 10%).

Figure 22: Caterpillar plot of the unadjusted center-specific 'excess' probability for the "all-or-none" score for domain 8 'Histopathological examination'.



4.5.2 Overview of variation per domain

Table 43 summarizes variation between centers expressed as the Inter Quartile Range of relevant excess probabilities as well as the number of centers with significant excess probabilities on domain quality indexes, i.e. the number of centers scoring significantly below P25 and above P75.

Table 43: Overview of quality index results per domain. Inter quartile range and number of centers significantly performing below P25 and above P75 in caterpillar plots of (un-)adjusted excess probability.

Domain	Description	unadjusted			adjusted		
		IQR (%)	<P25	>P75	IQR (%)	<P25	>P75
1	*General quality indicators*				**14**	***0***	**2**
2	*Diagnosis and staging*	**31**	**4**	**6**	25	5	5
3	*Neoadjuvant treatment*	**27**	***5***	***1***	-	-	-
4	*Surgery*	15	1	3	**12**	**1**	**2**
5	*Adjuvant treatment*	-	*NA*	*NA*		*NA*	*NA*
6	*Palliative treatment*	-	*NA*	*NA*		*NA*	*NA*
7	*Follow-up*	-	*NA*	*NA*		*NA*	*NA*
8	*Histopathologic examination*	**25**	**5**	**8**	24	2	5

NA: not assessed due to data availability or data quality; primary analysis is in bold (adjusted if outcome QCI is involved); italic: domains 1 and 3 are represented by single QCIs, analyzed as described in section 4.5

One concludes that domain 2 and 8 show most variation between centers. The efforts to harmonize surgical treatment appear to have paid off in this cohort, since less variation is seen in this domain (and quality achievement rates were quite high). For more extreme centers, the caterpillar plots warrant an investigation of the centre-specific covariates that contribute to their under/outperformance.

It is anticipated that fewer missing data will yield better discrimination in the future (i.e. there will be more significant outliers due to more narrow confidence intervals).

4.6 AN OVERALL COMPOSITE QUALITY INDEX

Here we report on the construct of a composite index for outcome and process QCIs combined. We appreciate this is a complex measure to interpret and produce it in response to the request to arrive at a single global quality index. The relative weight of outcome and process QCIs in this construction is to some degree a matter of choice. In an attempt to give similar chances to outcome and process QCIs to enter, a first approach started from the previously selected QCIs for respectively outcome and process aggregated scores. A second approach started instead from all available QCIs with sufficient data quality. For a number of reasons pertaining to data quality of measures involved and the discriminating ability of the resulting score, we opted for the second quality index and report on it below. Further detail can be found in Appendix 8.

Starting from 28 process and 16 outcome QCIs, and after imposing stringent data quality criteria (see Methodology section 2.1), we were left with 6 outcome QCIs (1111, 1231, 1234, 1234b, 1235 and 1235c) and 12 process QCIs (i.e. QCIs 1211, 1213, 1214, 1217, 1221, 1232a, 1271, 1272, 1273, 1273b, 1274 and 1275) to construct a composite index.

A principal component analysis confirmed that there is not one single dimension underlying the 18 indicators. The main dimension (factor 1) accounts for only 18% of the overall variance for the 18 indicators. Based on the default statistical selection of eigenvalues greater than 1, 6 dimensions were retained. These six factors account for 69% of the information available among the QCIs. Following a factor analysis and our criteria for inclusion, the composite QI is constructed using the following 2 outcome (O) and 4 process (P) QCIs:

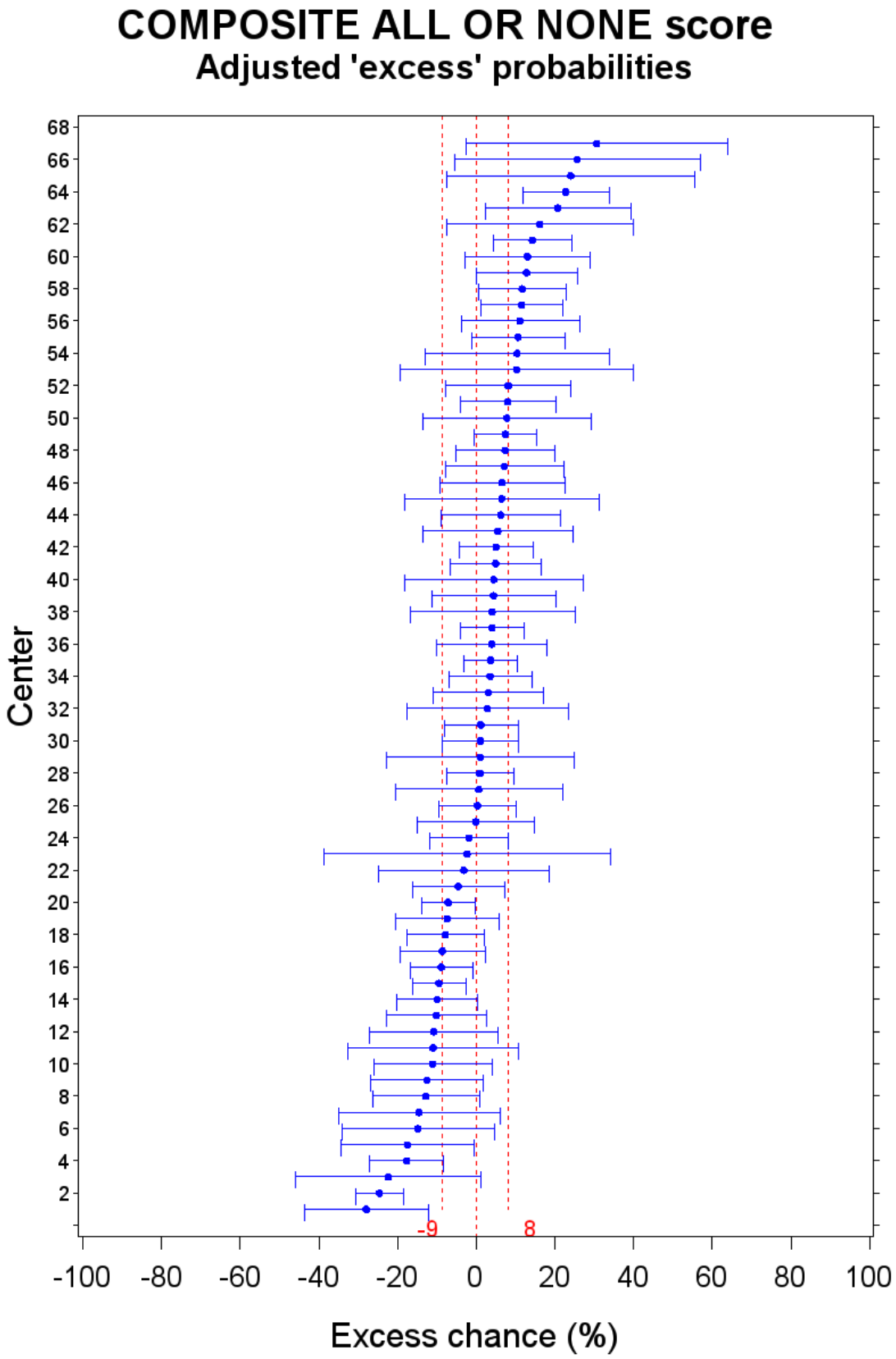
- QCI 1111 (O): Overall survival
- QCI 1231 (O): Proportion of R0 resections
- QCI 1211 (P): Documented distance from anal verge
- QCI 1273 (P): Distal tumor-free margin mentioned in the pathology report.
- QCI 1274 (P): Number of lymph nodes examined
- QCI 1272 (P): Quality of TME assessed according to Quirke and mentioned in the pathology report

Over the centers, the resulting "all or none" score is on average obtained in 29% of the patients. Since outcome QCIs are involved in this quality index, a model is built that adjusts for main effects of age (with a different slope before and after the breakpoint of 70 years), gender, ASA score, Tumor level, cStage, cT4, surgery and "cStage by Surgery". In terms of adjusted excess probabilities 2 centers are found to perform significantly below P25 (-9%) and one above P75 (8%).

Summary points

- **The methodology described in section 2 for aggregating quality indexes for the outcome domain, the process domain, for each of the 8 specific domains, and finally the overall composite index was applied to the available PROCARE dataset. In a first step principle components and factor analysis conducted on the center-specific excess chances lead to a selected set of QCIs to be included in the aggregated index. Because of limited data quality the most clinically relevant QCIs could not always be included in this construction. This was particularly apparent in the domain `Diagnosis and staging'.**
- **After considering discriminating ability and interpretation, the all-or-none scoring method was chosen to yield the aggregate quality index measure from the retained QCIs.**
- **Aggregated all-or-none scores which involved outcome QCIs were adjusted for baseline characteristics of the patients to yield center-specific excess probabilities. When only process QCIs were involved the all-or-none score was left unadjusted. The same method was used to calculate excess probabilities as for the original QCIs.**
- **Substantially more variation was seen in the achievement levels of the domains `Diagnosis and staging' and `Histopathological Examination' than in the `Surgery' domain suggesting there is still room for improvement in these former domains.**
- **A few centers were found to lie significantly below the P25 level for one or more of the aggregated outcomes. Given the present limitations of the database, especially the limited data quality and selectivity of registration, these results can not be seen as conclusive.**

Figure 23: Caterpillar plot of the adjusted center-specific 'excess' probability for the "all-or-none" composite score.



5 MODEL BUILDING TO EXPLAIN EXCESS PROBABILITIES

5.1 INTRODUCTION

As suggested by Normand et al.[37] centers were evaluated based on a comparison between the average expected outcome over the patient set they treated and the expected outcome these patients would have experienced had they been treated at the 'average' center. We called this measure the center-averaged excess outcome and denoted it with e_c for center c. For binary outcomes it was obtained through logistic regression and for survival outcome through the proportional hazards model (both with Firth correction, see Appendix 1, Chapter 10). In this section we propose a first explanatory analysis of these excess probabilities in terms of center-specific characteristics including excess probabilities on process QCIs as well as size. Specifically, we 1) regress excess probabilities for overall survival on the said predictors and 2) list average unadjusted excess probabilities on process QCIs for the 5 centers with the highest and lowest estimated adjusted excess probability for the 3 year survival outcome QCI. Given the current reservations on the data quality and derived excess probabilities this merely presents an illustration of such an approach referring to a limited set of possible center-specific explanatory variables and has no ambition of being a full-fledged explanatory analysis.

5.2 METHODOLOGY: FORWARD STEPWISE REGRESSION

Excess probabilities of achievement, $\hat{e}_c$ were calculated per center for both outcome and process quality control indexes (see additional figures in Appendix 6, sections 1.2-1.8). We regressed them on selected center-specific characteristics through a weighted least squares linear model. The centers were inversely weighted with their estimated variance on outcome $\hat{e}_c$. This is a first approximation to such explanatory regression. If one believes that there remains residual between center variation in this regression model, then this would contribute to the variance around the line and can be modelled to improve the weights and efficiency in the weighted least squares regression model as in van Houwelingen et al. 2002 [46]. If one furthermore believes that the true expected center effects on the process QCI determine the effect on outcome, one may recognize that the current predictors measure this with imprecision which would lead to an attenuated slope in the regression. With estimated variances of the excess probabilities on the process QCI, one can correct for this measurement error on the predictor. As our outcomes take both positive and negative values (which are constrained by -1 and +1) we start off with simple linear regression. If the data show substantial deviation from the model, we propose to transform outcomes into the 0-1 range and then fit a generalized linear model with logit link on the mean. Such transformation could consist of dividing $\hat{e}_c$ first by two and then adding 0.5 to that result.

We specifically entered the following centre-specific predictors in a forward stepwise fashion, including:

- Excess probabilities on process QCIs (unadjusted, provided sufficient data quality).
- Center size (derived from the BCR database)

For confidentiality reasons we lacked substantial detail on the center characteristics, which could be of interest at this stage and point to reasons for differential achievement.

5.3 EXPLANATORY DESCRIPTIVES FOR EXTREME CENTER PERFORMANCE

The list of predictors considered for explaining the adjusted excess probability of 3 year survival is given in Table 44. For 5 centers with the highest (see Table 45) and lowest (see *Colour code outcome QCI: orange: mean excess prob is not significantly below P25; red: significantly below P25; -Colour code process QCI: dark red: negative mean excess prob; dark green: positive mean excess prob; BCR: Belgian Cancer registry; PRO:PROCARE database*

Table 46) estimated value of the latter, the mean center-specific unadjusted excess probabilities on 12 process indicators (with sufficient data quality) are given in Table 47.

Table 44: List of selected process QCIs with sufficient data quality

QCI	Abbreviations	Description	Domain
1211	%DocDist	Proportion of patients with a documented distance from the anal verge	2
1213	%CEA_Preop	Proportion of patients in whom a CEA was performed before any treatment	2
1214	%Preop_Bowel_Im	Proportion of patients undergoing elective surgery that had preoperative complete large bowel-imaging	2
1217	Time_histo-1ther	Time between first histopathologic diagnosis and first treatment	2
1221	%Preop_RT	Proportion of cStage II-III patients that received a neoadjuvant pelvic RT	3
1232a	%Defin_ostomy	Proportion of APR- Hartmann's procedure or total excision of colon and rectum with definitive ileostomy	4
1271	%Path_Rep_Use	Use of the pathology report sheet	8
1272	%TME_Qual_Rep	Quality of TME assessed according to Quirke and mentioned in the pathology report	8
1273	%Dist_Margin_Pos_Rep	Distal margin involvement mentioned after SSO or Hartmann	8
1273b	%Dist_Margin_Rep	Distal tumor-free margin mentioned in the pathology report	8
1274	#Nodes_Examined	Number of lymph nodes examined	8
1275	pCRM_mm_Rep	(y)pCRM mentioned in mm in the pathology report	8

Table 45: Excess probabilities (%) for process QCIs for centers with lowest adjusted excess probability for 3 year survival

#patients[15]		S(3) QCI1111	1211	1213	1214	1217	1221	1232a	1271	1272	1273	1273b	1274	1275
BCR	PRO													
23	5	-30.3	-1.4	8.1	-5.6	4.5	19.7	-0.3	-5.5	24.8	5.6	0.7	-4.8	1.5
34	13	-29.5	3.3	-1.4	-1.1	-13.7	-26.9	-0.3			-17.9	-31.8	-2.8	-34.2
81	12	-22.6	3.1	4.9	-8.8	-0.1	13.1	1.6	-1.0	-21.9	8.9	4.1	-2.7	7.3
8	9	-22.1	1.9	11.4	-2.8	12.1		24.7					0.9	20.2
36	17	-17.3	-1.4	2.5	-6.1	14.7	-15.3	13.0	-2.7	-59.8	10.1	5.2	6.7	-5.9

Colour code outcome QCI: orange: mean excess prob is not significantly below P25; red: significantly below P25; -Colour code process QCI: dark red: negative mean excess prob; dark green: positive mean excess prob; BCR: Belgian Cancer registry; PRO:PROCARE database

Table 46: Excess probabilities (%) for process QCIs for centers with highest adjusted excess probability for 3 year survival

#patients[14]		S(3) QCI1111	1211	1213	1214	1217	1221	1232a	1271	1272	1273	1273b	1274	1275
BCR	PRO													
14	16	23.2	4.0	-10.0	-0.2	17.8	17.4	13.0	-0.1	24.3	1.4	-3.4	-3.6	16.5
47	40	18.5	3.2	5.4	1.4	-12.1	0.3	8.5	1.3	9.4	-25.7	-3.0	0.7	-24.9
44	14	18.1	-23.1	-13.6	-2.3	18.3	-3.2	-11.6	-1.7	10.4	-13.9	-18.7	-3.0	1.5
36	23	18.0	4.8	6.0	0.5	8.9	1.6	8.8	-0.3	-37.7	-25.4	-23.1	-0.1	-23.5
166	8	17.5	1.4	10.9	-3.5	2.5	-13.6	2.5	-10.8		5.6	-15.9	0.1	-4.7

Colour code outcome QCI: light green if mean excess prob is not significantly higher than P75; dark green: significantly higher than P75; Colour code process QCI: dark red: negative mean excess prob; dark green: positive mean excess prob

15 Overall about one third of patients in the BCR database are represented in PROCARE, the apparent assignment of more patients to the PROCARE database for some centers is due to the different center assignment procedures used by both registers, as explained in section 3.1 and Appendix 7.

Table 47: Average center size for the 5 centers with highest and lowest estimated adjusted excess chance of 3 year survival.

Ranking	BCR Center size	PROCARE Center-size
Top-5	61	20
Bottom-5	36	11

Univariate weighted least squares linear regression with excess survival chance as an outcome yields a significant (at the nominal .05 level) linear association with process QCI 1272 (%TME_Qual_Rep). The univariate regression models are summarized in Table 48. QCI 1221 (%Preop_RT) and center size (the number of patients included in either the BCR or the PROCARE database) are significant at the nominal .1 level.

Table 48: Univariate associations with excess 3-year survival chance as an outcome

Predictor	Estimate	StdErr	Pr > |t|
PROCARE size	0.03	0.02	0.081
BCR size	0.02	0.01	0.059
QCI 1211	0.07	0.11	0.545
QCI 1213	0.13	0.09	0.140
QCI 1214	0.46	0.39	0.238
QCI 1217	0.10	0.09	0.280
QCI 1221	0.10	0.05	0.077
QCI 1232a	-0.17	0.12	0.165
QCI 1271	0.08	0.22	0.708
QCI 1272	0.15	0.05	0.004
QCI 1273	-0.16	0.11	0.141
QCI 1273b	-0.02	0.14	0.863
QCI 1274	-0.15	0.45	0.744
QCI 1275	0.10	0.09	0.305

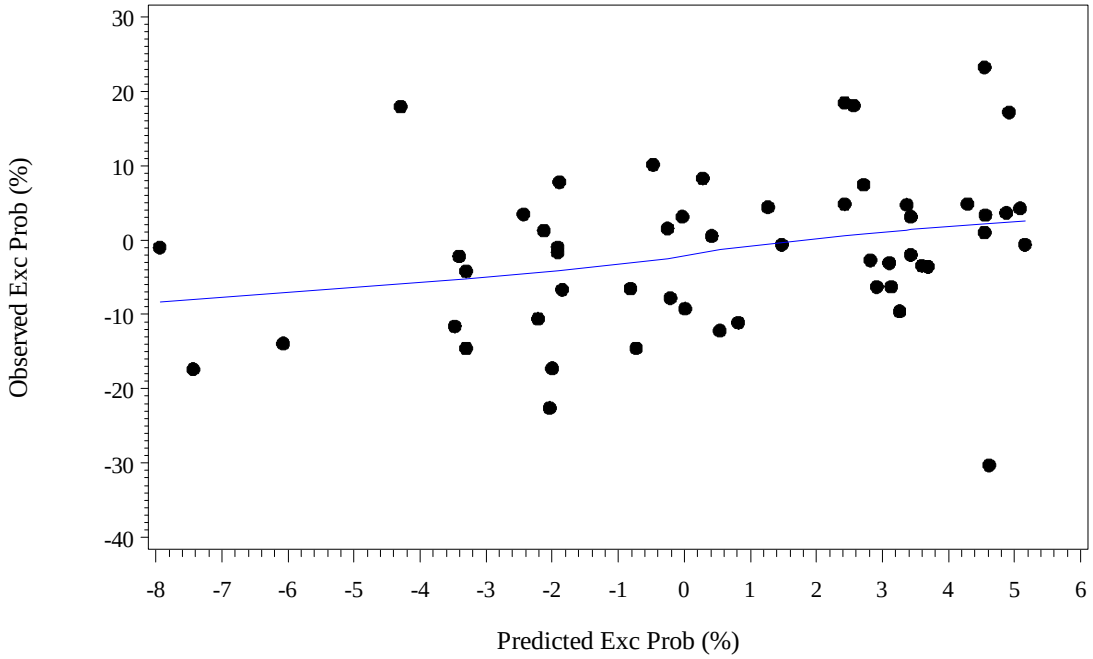
5.4 FORWARD STEPWISE REGRESSION

After stepwise selection of predictors (with and without center size included), a model with 1 significant predictor was selected, i.e. excess probability on process QCI 1272 (see Table 49). To support the linearity assumption of the fitted model, Figure 24 shows observed versus predicted excess outcomes with the locally smoothed regression line.

Table 49: Parameter estimates for the significant (at the .05 level) predictors in the model for excess probability of 3 year survival.

Variable	Parameter Estim.	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.09	1.06	2.55	1.05	0.310
QCI 1272 (%TME_Qual_Rep)	0.14	0.05	20.02	8.26	0.006

Figure 24: Model predictions for the adjusted excess probability of 3 year survival.



We note at this stage that data for this outcome QCI are still immature and the limited information implies relatively little power for the detection of significant effects.

For the purpose of explanation, we also added interaction effects between center and age to the final explanatory model at the patient level that had been reached for individual QCI111 measures. The added terms did not significantly improve the model (P=0.2038). Again, with limited data, this is at present not to be taken as evidence of the fact that no differences could exist between centers and their treatment of different age groups. There is at present however no evidence that such interaction indeed exists.

Summary points

- **Center performance, measured as adjusted excess probability of an outcome QCI or aggregate outcome QI, may be explained in terms of center characteristics using (generalized) linear models. Such analysis is of a hypothesis generating nature.**
- **Explanatory factors of interest include performance on process Q(C)Is, center size, region, treatment facilities, case load per surgeon, … some of these require more detailed knowledge about the center.**
- **Interactions between center and baseline characteristics of the patient, measured at the patient level, may point to different centers caring better for specific subpopulations.**
- **Limited center sizes and errors in measurement both contribute to reduced power to detect explanatory factors.**

6 DISCUSSION

The present report documents results of the study, development and application of methods for the evaluation of quality of care among rectal cancer centers entering information on quality indicators for some of their patients in the voluntary PROCARE registry. With its focus on causal inference, the methodology developed is generic in nature and the approach applicable for the monitoring of care in a whole range of cancers and other areas of disease.

The monitoring and evaluation of care envisaged here primarily serves an educational purpose, realized through confidential feedback on quality indicators to individual care centers on their performance - adjusted for patient mix when necessary. This is accompanied by a flagging of lower performance which is not overly conservative since it provides foremost an early warning signal for individual center use. While this effort is expected to lead to improved care, no specific monitoring of response to this feedback is currently conducted. Whether and how feedback indeed (best) leads to the desired response resulting in an optimization of care is yet to be addressed explicitly and is work for further study.

Feedback takes the form of domain-specific (or more global) presentations of center performance expressed as an excess probability: the expected percentage of patients in the center, beyond that in the average center, for which all relevant quality indicators in the composite index are met. Centers are plotted in order from the highest to the lowest estimated excess probability and presented with the 95% CI on this excess probability. Overlapping confidence intervals emphasize that the plotting order should not be seen as an indication of the intrinsic rank order of the centers. Substantial uncertainty on the relative quality index performance remains, particularly in centers which contributed few patients to the database so far. By choosing this presentation, we place ourselves on the sceptical side in the debate on the value of ranks in evaluating performance of care in line with Lilford et al. [53].

Ideally, confidence intervals on the quality indexes should be compared to absolute bench marks of good quality care. In practice, such objective bench marks are not yet available and hence the intervals are simply related to observed percentiles (quartiles) of performance over the centers. Relative benchmarks have the obvious drawback that 1) stricter demands are made in domains which generally perform better (i.e. with less variability over the centers) and 2) the bench marks are likely changing over time and cannot be used in an international comparison.

Domain-specific aggregate scores allow the comparison of specific services across centers. Their construction is designed to involve fewer QCIs than the original set of QCIs in the database without losing much information. The obtained reduction in QCIs for this evaluation thus suggests more economic ways of gathering data. A word of caution is in place, however, when interpreting current results of the implementation of this strategy. In some instances, the most clinically meaningful and important variables lacked the quality required for entry into the construction of the aggregate quality index. This was the case for example in the domain `diagnosing and staging', where such indexes did not reach the quality threshold criteria. This problem is expected to be resolved once the methods developed here are re-applied to a more mature database.

In general, the question of whether or not to reduce the number of QCIs to be measured is a difficult one, with several dimensions:

- How much information is lost? This is the line we followed and targeted in the report: retaining a sufficient amount of variation in the full set of indicators. One could alternatively focus more directly on the impact of a reduced number of QCIs on the accurate ranking of centers either in relative terms or in terms of absolute bench marks
- Would the recording of fewer QCIs lead to more accurate and complete data?

- Will fewer QCIs make it more difficult to diagnose the reasons for center underperformance and to intervene successfully?
- Are there other ways of retrieving the latter information post hoc in a more cost-efficient way, for instance through linking with administrative databases?

The application of the proposed methods to the PROCARE database demonstrates the feasibility of the approach in this setting, but also indicates the limitations caused by incompleteness and data quality of the database. Results should not uncritically be used to objectively judge the quality of rectal cancer care in Belgium at present. For instance, while the database structure and maturity is evolving, one expects a growing number of variables will hit the quality threshold and be allowed to enter the construction for the aggregated quality index. As a result these indexes and the corresponding models for adjustment are expected to change over a period to come. Once the database structure is stable one should be able to work with a given derived model for quite a while without need for rebuilding the model for each feedback application.

Notwithstanding the database limitations, certain conclusions emerge with some clarity. In Chapter 4 (and appendix 8) we learn, for instance, that the efforts towards standardizing surgical treatment appear to have paid off among the selected surgeons contributing to the PROCARE database. This emerges as the quartiles of the adjusted performance on the all-or-none score range from a mere 6% below to 6% above the center average performance. In addition, even with these narrow margins, just 1 center stays significantly below the P25 (and two above the P75). This is in contrast with results for Domain 4 (diagnosing and staging) and Domain 8 (histopathology) where much more variation is seen with a distance between quartiles of adjusted excess probabilities of 25% and 24%, respectively. The unadjusted ranges are even wider. Notwithstanding the wide margins, we found that 5 and 2 centers, respectively, have indicators that are significantly below the P25. This indicates that there is room for improvement in this domain. An educational system as put in place for surgery could be considered as a fruitful way forward.

We note that - in a continuous effort to protect confidentiality - the report contains marginal presentations of performance on separate QCIs or quality indexes. An inspection of the joint performance on several indexes could point more strongly to generalized drop of quality in some centers. While we have not presented that link in the report, it is present in the feedback to centers since each is presented with its own performance on several indexes. Clearly, centers featuring in the tail of the distribution of several indicators should be especially concerned about improving their performance in those areas.

Lessons have been learned that suggest ways for improvement as one builds on results on this report both for data gathering and analysis.

On the data: issues of completeness and reliability.

Currently, about one third of the targeted patient population is registered in the PROCARE database. Less than a sixth of them contribute to all QCIs. This is not only the consequence of specific eligibility criteria, but also a symptom of a considerable amount of missing data. It comes as no surprise, therefore, that the request for more comprehensive coverage of all relevant patients and centers has been echoed by most stakeholders in the quality improvement process. This will be a continued concern. Even when coverage is more comprehensive the care for data quality remains important. This project was unfortunately not able to work with some important variables because their data quality was not good enough. Ideally both preventive and curative action can be taken.

- the amount of data to enter should be limited to what is necessary and not excessively time-consuming to enter to avoid registration fatigue
- whenever possible a two-stage data entry procedure may be considered whereby more detailed information is requested only once a certain quality threshold is hit,

- a system of peer review should be put in place (on a voluntary basis) that allows for regular random audits of data records.
- (if all else fails) some compulsory data entry could be considered

This work would benefit further in particular from linking to administrative databases to improve the quality of data gathering and reduce the registration efforts for the centers. The availability of further predictors of center choice, such as distance between the patient's residence and the centers, and predictors of outcome, such as socio-economic factors, would enhance the assessment of effects truly due to centers rather than factors confounded with the patient choice. The available resource was rather poor in that sense which is an important reason why concrete results beyond the methodological approach should be considered as quite tentative here and still confounded with patient characteristics to some degree.

On model building quality

The current project was charged to analyze a total of 44 QCIs and aggregated quality indexes. For several of them (16 outcome QCIs) and for the aggregated quality indexes (8 per domain + 3 global ones (+1 modified build up)) full models have been built for the correction of QCIs for baseline covariates, on the currently still immature data. These prediction models have no ambition of showing causal effects of patient characteristics, but do aim to predict expected QCIs in function of these characteristics such that the additional center effects can be interpreted causally if the predictive value of baseline confounders for center effects has been well captured by the model.

While we have sought to provide methods that are readily adaptable in the routine practice of an institution like the Belgian Cancer Register, and we believe we have succeeded, the models and the database are complex, and the model building cannot yet be automated in standard statistical software.

On a more technical note:

The working logistic and Cox regression models involved a Firth Bias correction which cannot be used in combination with an automated stepwise regression procedure in SAS. For the construction of confidence intervals on the resulting center-specific predictions a delta method was developed and implemented in R for the production of caterpillar plots. This involves an extra 'manual' step in the production of the analysis. The combination of sample size and number of patients registered, means that several methods have hit their limit of applicability which could lead to extra convergence and coverage issues. The same is true for the missing data that contaminate this database, and every other one like it (voluntary based). The model building procedure applied here could be subject to more detailed scrutiny and more sophisticated adaptation if it did not have to be developed for so many different outcomes with the limited resources of time and staff. We encourage future users to expand on this once the means are available. In particular we advocate the selection of a single (or a few) key measures of evaluation which can be modeled in much more detail than here possible. Cross validation can then also be applied for instance for evaluation of the stability of the models.

Another area for further research concerns the burden of evidence to put on small centers. Due to their typically large confidence intervals small centers have relatively little chance of producing `evidence of significant underperformance' even though there are indications of a volume outcome relationship in the literature. One might therefore consider demanding varying levels of confidence for crossing relative (and ideally absolute) bench marks in this regard. In particular, approaches as developed in Uno et al. [54] would lead to an average confidence level of 95% whereby smaller centers be presented with lower confidence levels than larger centers.

In conclusion, we emphasize that the present report presents first and foremost a methodological approach to the problem of assessing quality of care. Secondly, the application to the current PROCARE database demonstrates practical feasibility of the selected methods in this context. On both accounts methodology as well as application is not 'the final answer', but several important steps were achieved and a feasible approach was presented for work that will continue to progress. There are important open questions about some of the candidate methods, questions that will benefit from further targeted statistical research. The application to and results on the PROCARE database inherit all the limitations of the current database. Selectiveness of the recorded patients, missing data, and uneven data quality all contribute to results that must be interpreted with caution.

More generally, we have presented work that will evolve with increased experience with the data and the methods. Our current developments have similarly built on the previous PROCARE reports which resulted in several recommendations for adapted or refined definitions of previously validated quality of care indicators. For instance, upon analyzing QCIs and correspondingly studying them in more detail, it emerged that some of the previously validated QCIs needed revised definitions. Some QCIs have also changed because they turned out harder to measure than anticipated.

We finally conclude that registration, analysis and feedback on quality of care indicators form an important and irreplaceable instrument for quality of care improvement. It is essential for evidence-based assessments of the performance of health care centers and for the design of intervention strategies to improve care. Imperfect data are better than no data when approached with appropriate care.

7 BIBLIOGRAPHY

1. Penninckx F, Roels S, Leonard D, Laurent S, Decaestecker J, Vleeshouwer CD, et al. Quality of care for rectal cancer - Phase 1: a clinical practice guideline for rectal cancer. Brussels: Belgian Health Care Knowledge Centre (KCE); 2007. Good Clinical Practice (GCP) KCE reports 69

2. Vlayen J, Verstreken M, Mertens C, Eycken EV, Penninckx F. Quality insurance for rectal cancer - Phase 2: Development and testing of a set of quality indicators. Brussels: Belgian Health Care Knowlegde Centre (KCE); 2008. Good Clinical Practice (GCP KCE reports 81

3. Olsson LI, Granstrom F, Pahlman L. Sphincter preservation in rectal cancer is associated with patients' socioeconomic status. British Journal of Surgery. 2010;97(10):1572-81.

4. Lejeune C, Sassi F, Ellis L, Godward S, Mak V, Day M, et al. Socio-economic disparities in access to treatment and their impact on colorectal cancer survival. International Journal of Epidemiology. 2010;39(3):710-7.

5. Morris E, Quirke P, Thomas JD, Fairley L, Cottier B, Forman D. Unacceptable variation in abdominoperineal excision rates for rectal cancer: time to intervene? Gut. 2008;57(12):1690-7.

6. Basu A, Heckman JJ, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: An application to treatments of breast cancer patients. Health Economics. 2007;16:1133-57.

7. Brooks JM, Chrischilles EA. Heterogeneity and the interpretation of treatment effect estimates from risk adjustment and instrumental variable methods. Medical Care. 2007;45(10):S123-S30.

8. Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: Instrumental variable and propensity analysis. Journal of Clinical Oncology. 2001;19(4):1064-70.

9. Hadley J, Polsky D, Mandelblatt JS, Mitchell JM, Weeks JC, Wang Q, et al. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. Health Economics. 2003;12(3):171-86.

10. Sussman JB, Hayward RA. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. British Medical Journal. 2010;340.

11. Wisnivesky JP, Halm E, Bonomi M, Powell C, Bagiella E. Effectiveness of Radiation Therapy for Elderly Patients with Unresected Stage I and II Non-Small Cell Lung Cancer. American Journal of Respiratory and Critical Care Medicine. 2010;181(3):264-9.

12. Zeliadt SB, Potosky AL, Penson DF, Etzioni R. Survival benefit associated with adjuvant androgen deprivation therapy combined with radiotherapy for high- and low-risk patients with nonmetastatic prostate cancer. International Journal of Radiation Oncology Biology Physics. 2006;66(2):395-402.

13. Firth D. Bias reduction of maximum-likelihood-estimates. Biometrika. 1993;80(1):27-38.

14. Othus M, Li Y, Tiwari RC. A Class of Semiparametric Mixture Cure Survival Models With Dependent Censoring. Journal of the American Statistical Association. 2009;104(487):1241-50.

15. Normand SLT, Shahian DM. Statistical and clinical aspects of hospital outcomes profilling. Statistical Science. 2007;22:206-26.

16. Munro AJ, Bentley AH, Ackland C, Boyle PJ. Smoking compromises cause-specific survival in patients with operable colorectal cancer. Clin Oncol (R Coll Radiol). 2006;18(6):436-40.

17. Anwar S, Fraser S, Hill J. Surgical specialization and training - its relation to clinical outcome for colorectal cancer surgery. J Eval Clin Pract. 2010.

18. Nugent E, Neary P. Rectal cancer surgery: volume-outcome analysis. Int J Colorectal Dis. 2010;25(12):1389-96.

19. Kressner U, Glimelius B, Bergstrom R, Pahlman L, Larsson A, Lindmark G. Increased serum p53 antibody levels indicate poor prognosis in patients with colorectal cancer. Br J Cancer. 1998;77(11):1848-51.

20. Borowski DW, Bradburn DM, Mills SJ, Bharathan B, Wilson RG, Ratcliffe AA, et al. Volume-outcome analysis of colorectal cancer-related outcomes. Br J Surg. 2010;97(9):1416-30.

21. van Gijn W, Gooiker GA, Wouters MW, Post PN, Tollenaar RA, van de Velde CJ. Volume and outcome in colorectal cancer surgery. Eur J Surg Oncol. 2010;36 Suppl 1:S55-63.

22. Faerden AE, Naimy N, Wiik P, Reiertsen O, Weyessa S, Tronnes S, et al. Total mesorectal excision for rectal cancer: difference in outcome for low and high rectal cancer. Dis Colon Rectum. 2005;48(12):2224-31.

23. Kusters M, Holman FA, Martijn H, Nieuwenhuijzen GA, Creemers GJ, Daniels-Gooszen AW, et al. Patterns of local recurrence in locally advanced rectal cancer after intra-operative radiotherapy containing multimodality treatment. Radiother Oncol. 2009;92(2):221-5.

24. Pinsk I, Phang PT. Total mesorectal excision and management of rectal cancer. Expert Rev Anticancer Ther. 2007;7(10):1395-403.

25. den Dulk M, Marijnen CA, Collette L, Putter H, Pahlman L, Folkesson J, et al. Multicentre analysis of oncological and survival outcomes following anastomotic leakage after rectal cancer surgery. Br J Surg. 2009;96(9):1066-75.

26. Bernstein TE, Endreseth BH, Romundstad P, Wibe A. Circumferential resection margin as a prognostic factor in rectal cancer. Br J Surg. 2009;96(11):1348-57.

27. Quirke P, Steele R, Monson J, Grieve R, Khanna S, Couture J, et al. Effect of the plane of surgery achieved on local recurrence in patients with operable rectal cancer: a prospective study using data from the MRC CR07 and NCIC-CTG CO16 randomised clinical trial. Lancet. 2009;373(9666):821-8.

28. Eberhardt JM, Kiran RP, Lavery IC. The impact of anastomotic leak and intra-abdominal abscess on cancer-related outcomes after resection for colorectal cancer: a case control study. Dis Colon Rectum. 2009;52(3):380-6.

29. Law WL, Choi HK, Lee YM, Ho JW, Seto CL. Anastomotic leakage is associated with poor long-term outcome in patients after curative colorectal resection for malignancy. J Gastrointest Surg. 2007;11(1):8-15.

30. Jorgren F, Johansson R, Damber L, Lindmark G. Anastomotic leakage after surgery for rectal cancer: a risk factor for local recurrence, distant metastasis and reduced cancer-specific survival? Colorectal Dis. 2011;13(3):272-83.

31. Bertelsen CA, Andreasen AH, Jorgensen T, Harling H. Anastomotic leakage after curative anterior resection for rectal cancer: short and long-term outcome. Colorectal Dis. 2010;12(7 Online):e76-81.

32. Lee WS, Yun SH, Roh YN, Yun HR, Lee WY, Cho YB, et al. Risk factors and clinical outcome for anastomotic leakage after total mesorectal excision for rectal cancer. World J Surg. 2008;32(6):1124-9.

33. Eriksen MT, Wibe A, Norstein J, Haffner J, Wiig JN. Anastomotic leakage following routine mesorectal excision for rectal cancer in a national cohort of patients. Colorectal Dis. 2005;7(1):51-7.

34. O'Brien SM, Shahian DM, DeLong ER, Normand SLT, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery: Part 2 - Statistical considerations in composite measure scoring and provider rating. Annals of Thoracic Surgery. 2007;83(4):S13-S26.

35. Smith P. Developing composite indicators for assessing health system efficiency. Measuring Up: Improving Health System Performance in Oecd Countries. 2002:295-316.

36. Normand SLT, Wolf RE, McNeil BJ. Discriminating quality of hospital care in the United States. Medical Decision Making. 2008;28(3):308-22.

37. Normand SLT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: Issues and applications. Journal of the American Statistical Association. 1997;92(439):803-14.

38. Normand SLT, Wolf RE, Ayanian JZ, McNeil BJ. Assessing the accuracy of hospital clinical performance measures. Medical Decision Making. 2007;27(1):9-20.

39. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, et al. Cardiac surgery report cards: Comprehensive review and statistical critique. Annals of Thoracic Surgery. 2001;72(6):2155-68.

40. Shahian DM, Edwards FH, Ferraris VA, Haan CK, Rich JB, Normand SLT, et al. Quality measurement in adult cardiac surgery: Part 1 - Conceptual framework and measure selection. Annals of Thoracic Surgery. 2007;83(4):S3-S12.

41. Ferris TG, Torchiana DF. Public Release of Clinical Outcomes Data — Online CABG Report Cards. The New England Journal of Medicine. 2010(10.1056/NEJMp1009423).

42. Wouters MWJM, Brand R, Eddes EH, Gooiker GA, Kolfschoten NE, Tollenaar RAEM. Uitkomst van zorg registratie; transparantie, keuzes en kwaliteit van zorg. 2009. Dutch Surgical Colorectal Audit

43. van Gijn W, Van de Velde CJH. Improving quality of cancer care through surgical audit. European Journal of Surgical Oncology. 2010;36:S23-S6.

44. Epstein A. SOUNDING BOARD - PERFORMANCE REPORTS ON QUALITY - PROTOTYPES, PROBLEMS, AND PROSPECTS. New England Journal of Medicine. 1995;333(1):57-61.

45. Wouters MWJM, Brand R, Eddes EH, Gooiker GA, Kolfschoten NE, Tollenaar RAEM. Uitkomst van zorg registratie; transparantie, keuzes en kwaliteit van zorg. 2009. Dutch Surgical Colorectal Audit

46. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine. 2002;21(4):589-624.

47. Efron B. REGRESSION AND ANOVA WITH ZERO-ONE DATA - MEASURES OF RESIDUAL VARIATION. Journal of the American Statistical Association. 1978;73(361):113-21.

48. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Statistics in Medicine. 2005;24(8):1185-202.

49. Goldstein H, Spiegelhalter DJ. League tables and their limitations: Statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society Series a-Statistics in Society. 1996;159:385-409.

50. Marshall EG, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. British Medical Journal. 1998;316(7146):1701-4.

51. Higashizawa T, Koga Y. Modified ASA Physical Status (7 grades) May Be More Practical In Recent Use For Preoperative Risk Assessment. The Internet Journal of Anesthesiology. 2007;15(1).

52. Goetghebeur E, Ryan L. Analysis of competing risks survival data when some failure types are missing. Biometrika. 1995;82(4):821-33.

53. Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. Lancet. 2004;363(9415):1147-54.

54. Uno H, Tian L, Wei LJ. The optimal confidence region for a random parameter. Biometrika. 2005;92(4):957-64.

This page is left intentionally blank.

Legal depot : D/2010/10.273/40

KCE reports

33 Effects and costs of pneumococcal conjugate vaccination of Belgian children. D/2006/10.273/54.
34 Trastuzumab in Early Stage Breast Cancer. D/2006/10.273/25.
36 Pharmacological and surgical treatment of obesity. Residential care for severely obese children in Belgium. D/2006/10.273/30.
37 Magnetic Resonance Imaging. D/2006/10.273/34.
38 Cervical Cancer Screening and Human Papillomavirus (HPV) Testing D/2006/10.273/37.
40 Functional status of the patient: a potential tool for the reimbursement of physiotherapy in Belgium? D/2006/10.273/53.
47 Medication use in rest and nursing homes in Belgium. D/2006/10.273/70.
48 Chronic low back pain. D/2006/10.273.71.
49 Antiviral agents in seasonal and pandemic influenza. Literature study and development of practice guidelines. D/2006/10.273/67.
54 Cost-effectiveness analysis of rotavirus vaccination of Belgian infants D/2007/10.273/11.
59 Laboratory tests in general practice D/2007/10.273/26.
60 Pulmonary Function Tests in Adults D/2007/10.273/29.
64 HPV Vaccination for the Prevention of Cervical Cancer in Belgium: Health Technology Assessment. D/2007/10.273/43.
65 Organisation and financing of genetic testing in Belgium. D/2007/10.273/46.
66. Health Technology Assessment: Drug-Eluting Stents in Belgium. D/2007/10.273/49.
70. Comparative study of hospital accreditation programs in Europe. D/2008/10.273/03
71. Guidance for the use of ophthalmic tests in clinical practice. D/200810.273/06.
72. Physician workforce supply in Belgium. Current situation and challenges. D/2008/10.273/09.
74 Hyperbaric Oxygen Therapy: a Rapid Assessment. D/2008/10.273/15.
76. Quality improvement in general practice in Belgium: status quo or quo vadis? D/2008/10.273/20
82. 64-Slice computed tomography imaging of coronary arteries in patients suspected for coronary artery disease. D/2008/10.273/42
83. International comparison of reimbursement principles and legal aspects of plastic surgery. D/200810.273/45
87. Consumption of physiotherapy and physical and rehabilitation medicine in Belgium. D/2008/10.273/56
90. Making general practice attractive: encouraging GP attraction and retention D/2008/10.273/66.
91 Hearing aids in Belgium: health technology assessment. D/2008/10.273/69.
92. Nosocomial Infections in Belgium, part I: national prevalence study. D/2008/10.273/72.
93. Detection of adverse events in administrative databases. D/2008/10.273/75.
95. Percutaneous heart valve implantation in congenital and degenerative valve disease. A rapid Health Technology Assessment. D/2008/10.273/81
100. Threshold values for cost-effectiveness in health care. D/2008/10.273/96
102. Nosocomial Infections in Belgium: Part II, Impact on Mortality and Costs. D/2009/10.273/03
103 Mental health care reforms: evaluation research of 'therapeutic projects' - first intermediate report. D/2009/10.273/06.
104. Robot-assisted surgery: health technology assessment. D/2009/10.273/09
108. Tiotropium in the Treatment of Chronic Obstructive Pulmonary Disease: Health Technology Assessment. D/2009/10.273/20
109. The value of EEG and evoked potentials in clinical practice. D/2009/10.273/23
111. Pharmaceutical and non-pharmaceutical interventions for Alzheimer's Disease, a rapid assessment. D/2009/10.273/29
112. Policies for Orphan Diseases and Orphan Drugs. D/2009/10.273/32.
113. The volume of surgical interventions and its impact on the outcome: feasibility study based on Belgian data
114. Endobronchial valves in the treatment of severe pulmonary emphysema. A rapid Health Technology Assessment. D/2009/10.273/39
115. Organisation of palliative care in Belgium. D/2009/10.273/42
116. Interspinous implants and pedicle screws for dynamic stabilization of lumbar spine: Rapid assessment. D/2009/10.273/46

117. Use of point-of care devices in patients with oral anticoagulation: a Health Technology Assessment. D/2009/10.273/49.
118. Advantages, disadvantages and feasibility of the introduction of 'Pay for Quality' programmes in Belgium. D/2009/10.273/52.
119. Non-specific neck pain: diagnosis and treatment. D/2009/10.273/56.
121. Feasibility study of the introduction of an all-inclusive case-based hospital financing system in Belgium. D/2010/10.273/03
122. Financing of home nursing in Belgium. D/2010/10.273/07
123. Mental health care reforms: evaluation research of 'therapeutic projects' - second intermediate report. D/2010/10.273/10
124. Organisation and financing of chronic dialysis in Belgium. D/2010/10.273/13
125. Impact of academic detailing on primary care physicians. D/2010/10.273/16
126. The reference price system and socioeconomic differences in the use of low cost drugs. D/2010/10.273/20.
127. Cost-effectiveness of antiviral treatment of chronic hepatitis B in Belgium. Part 1: Literature review and results of a national study. D/2010/10.273/24.
128. A first step towards measuring the performance of the Belgian healthcare system. D/2010/10.273/27.
129. Breast cancer screening with mammography for women in the agegroup of 40-49 years. D/2010/10.273/30.
130. Quality criteria for training settings in postgraduate medical education. D/2010/10.273/35.
131. Seamless care with regard to medications between hospital and home. D/2010/10.273/39.
132. Is neonatal screening for cystic fibrosis recommended in Belgium? D/2010/10.273/43.
133. Optimisation of the operational processes of the Special Solidarity Fund. D/2010/10.273/46.
135. Emergency psychiatric care for children and adolescents. D/2010/10.273/51.
136. Remote monitoring for patients with implanted defibrillator. Technology evaluation and broader regulatory framework. D/2010/10.273/55.
137. Pacemaker therapy for bradycardia in Belgium. D/2010/10.273/58.
138. The Belgian health system in 2010. D/2010/10.273/61.
139. Guideline relative to low risk birth. D/2010/10.273/64.
140. Cardiac rehabilitation: clinical effectiveness and utilisation in Belgium. d/2010/10.273/67.
141. Statins in Belgium: utilization trends and impact of reimbursement policies. D/2010/10.273/71.
142. Quality of care in oncology: Testicular cancer guidelines. D/2010/10.273/74
143. Quality of care in oncology: Breast cancer guidelines. D/2010/10.273/77.
144. Organization of mental health care for persons with severe and persistent mental illness. What is the evidence? D/2010/10.273/80.
145. Cardiac resynchronisation therapy. A Health technology Assessment. D/2010/10.273/84.
146. Mental health care reforms: evaluation research of 'therapeutic projects'. D/2010/10.273/87
147. Drug reimbursement systems: international comparison and policy recommendations. D/2010/10.273/90
149. Quality indicators in oncology: testis cancer. D/2010/10.273/98.
150. Quality indicators in oncology: breast cancer. D/2010/10.273/101.
153. Acupuncture: State of affairs in Belgium. D/2011/10.273/06.
154. Homeopathy: State of affairs in Belgium. D/2011/10.273/14.
155. Cost-effectiveness of 10- and 13-valent pneumococcal conjugate vaccines in childhood. D/2011/10.273/21.
156. Home Oxygen Therapy. D/2011/10.273/25.
158. The pre-market clinical evaluation of innovative high-risk medical devices. D/2011/10.273/31
159. Pharmacological prevention of fragility fractures in Belgium. D/2011/10.273/34.
160. Dementia: which non-pharmacological interventions? D/2011/10.273/37
161. Quality Insurance of rectal cancer – phase 3: statistical methods to benchmark centers on a set of quality indicators. D/2011/10.273/40.

This list only includes those KCE reports for which a full English version is available. However, all KCE reports are available with a French or Dutch executive summary and often contain a scientific summary in English.