Optimizing Quality of Cancer Care Using Outcome Information



E.M. Verheul

Propositions

1. There is a lack of consensus on the evaluation and reporting of case-mix models, which reduces the validity of benchmarking approaches – *this thesis.*

2. For public reporting purposes, it is essential to report quality indicator scores of individual hospitals over multiple years or at another aggregate level to achieve higher rankability - *this thesis.*

3. Transparent reporting of the percentage of missing outcome data and the strategies used to address it is essential to ensure the credibility and comparability of hospital performance - *this thesis.*

4. Rather than prioritizing the continuous development of new prediction models, greater emphasis should be placed on externally validating those prediction tools that are already in clinical use or have shown promise in prior studies - *this thesis.*

5. As PROMs are increasingly integrated into routine clinical practice, HRQoL models provide a new opportunity to improve personalized care - *this thesis.*

6. The ethics of data use will only reach its full potential when "not sharing" is recognized as a moral problem - *Bull et al., J Empir Res Hum Res Ethics, 2015.*

7. Interdisciplinary collaboration becomes truly effective only when methodological expertise is structurally embedded in working with outcome data - *van Baalen S, Boon M. BMC Med Educ. 2024.*

8. The *Not Invented Here* syndrome hampers the widespread adoption of successful quality initiatives – *Inspired by Zorginstituut Magazine, December 2024.*

9. True quality of care also lies in acknowledging the person beyond the data - *Inspired by Zorginstituut Magazine, October 2024.*

10. Scientific integrity requires that quality outweigh quantity, and that research is conducted for the right reasons – *D. G. Altman, 1994*

11. Teamwork makes the dream work - *John C. Maxwell*

Optimizing Quality of Cancer Care Using Outcome Information

E.M. Verheul

Colophon

The publication of this thesis was financially supported by the Department of Public Health, Erasmus MC.

Lay-out and printing: ProefschriftMaken | proefschriftmaken.nl
ISBN: 978-94-6534-370-9

© Copyright 2026: Elfi Malou Verheul, Rotterdam, the Netherlands
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without permission of the author.

Optimizing Quality of Cancer Care Using Outcome Information

Het verbeteren van de kwaliteit van kankerzorg
met behulp van uitkomstinformatie

Thesis

to obtain the degree of Doctor from
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. ir. A.J. Schuit

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Tuesday 23 June 2026 at 15.30 hrs
by

Elfi Malou Verheul
born in Nieuwegein, the Netherlands.

Erasmus University Rotterdam

DOCTORAL COMMITTEE:

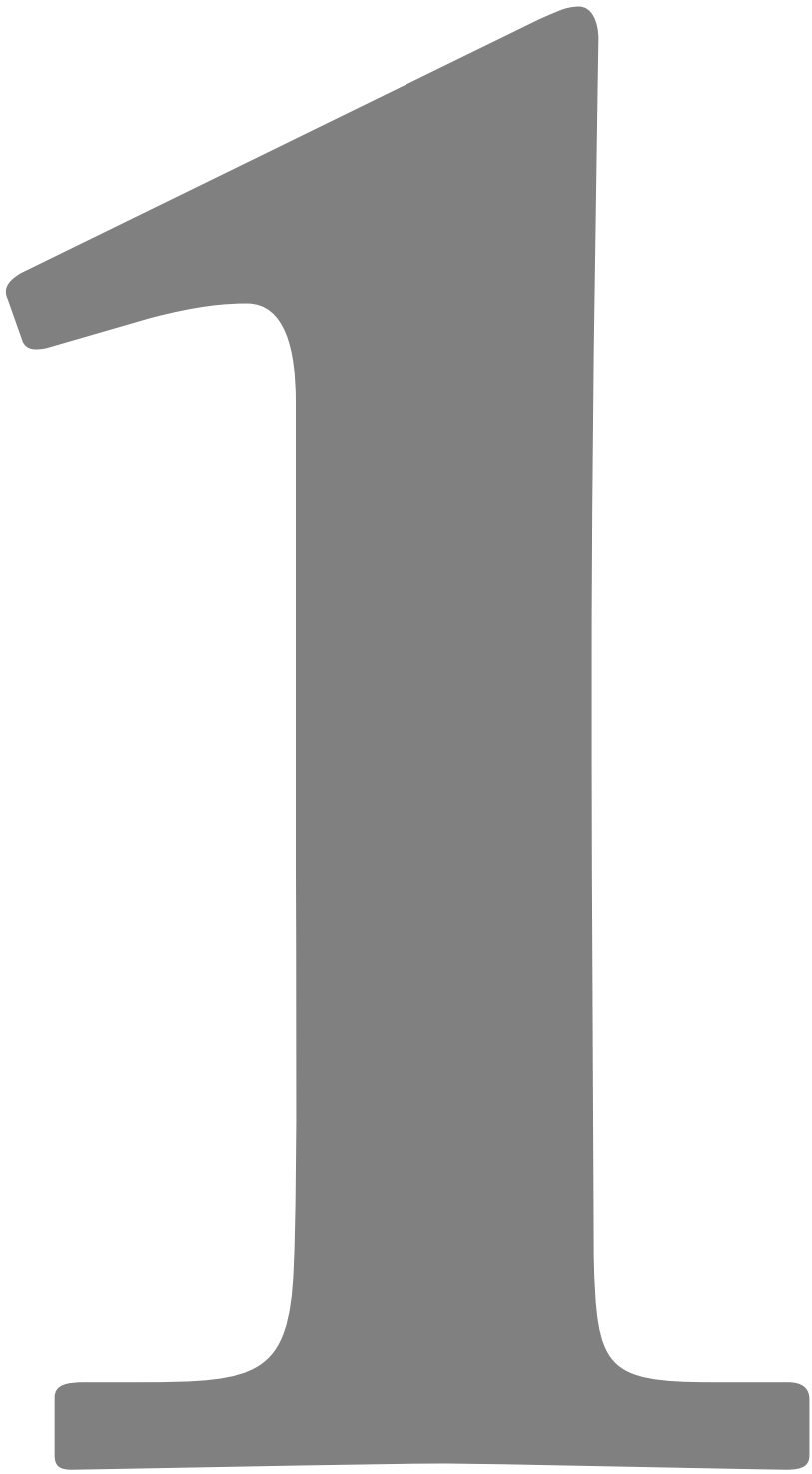
Promotors:	Prof. dr. H.F. Lingsma Prof. dr. S. Siesling Dr. L.B. Koppert
Co-promotor:	Dr. D. van Klaveren
Assessment committee:	Prof. dr. M.T.F.D Vrancken Peeters Dr. E.W. van Zwet Prof. dr. D.J. Grünhagen

TABLE OF CONTENT

Chapter 1	General introduction	9
Part I	**Using outcome information to stimulate continuous quality improvement**	25
Chapter 2	Benchmarking cancer outcomes in Europe: a scoping review of methodologies and case-mix adjustments	27
Chapter 3	High-impact complications after breast cancer surgery in the Dutch national quality registry: evaluating case-mix adjustment for hospital comparisons	49
Chapter 4	Comprehensive evaluation of quality indicators: analysing the Dutch breast cancer audit	71
Chapter 5	Public performance reporting of oral cancer surgery: detecting meaningful between-hospital variation	99
Part II	**Using outcome predictions to support personalised care**	117
Chapter 6	External validation of PREDICT breast v3.1 for overall survival in international cohorts, including young and invasive lobular subgroups	119
Chapter 7	Comparative validation of PREDICT versions 3.1 and 2.2 for overall survival in the Dutch breast cancer population	141
Chapter 8	Development and validation of prediction models for health-related quality of life outcomes after breast cancer surgery and reconstruction	163
Chapter 9	General discussion	185
Addendum	Summary (English)	203
	Samenvatting (Dutch)	209
	List of publications	213
	PhD portfolio	215
	About the author	218
	Dankwoord	219

LIST OF ABBREVIATIONS

4D PICTURE	**D**esign-based **D**ata-**D**riven **D**ecision-support tools: **P**roducing **I**mproved **C**ancer outcomes **T**hrough **U**ser-**C**entred **R**esearch (International research project)
DICA	Dutch Institute for Clinical Auditing
ECIBC	European Commission Initiative on Breast Cancer
HRQoL	Health-Related Quality of Life
IKNL	Netherlands Comprehensive Cancer Organisation *(Integraal Kankercentrum Nederland)*
ICHOM	International Consortium for Health Outcomes Measurement
IOM	Institute of Medicine
MRI	Magnetic Resonance Imaging
NABON	National Breast Cancer Organisation Netherlands *(Nationaal Borstkanker Overleg Nederland)*
NBCA	NABON Breast Cancer Audit
NCR	Netherlands Cancer Registry
NST	Neoadjuvant Systemic Therapy
PALGA	Dutch Nationwide Pathology Archive *(Pathologisch-Anatomisch Landelijk Geautomatiseerd Archief)*
PRO	Patient-Reported Outcome
PROM	Patient-Reported Outcome Measure
QI	Quality Indicator
SDM	Shared Decision-Making
SONCOS	Dutch Oncological Collaboration Foundation *(Stichting Oncologische Samenwerking)*
VBHC	Value-Based Healthcare



Chapter 1

General Introduction

QUALITY OF CARE

The concept of quality of care is fundamental yet inherently complex. It encompasses multiple dimensions such as effectiveness, safety, patient-centeredness, timeliness, efficiency, and equity[1], which are not always directly measurable and are valued differently by diverse stakeholders. Over time, various definitions have emerged, each reflecting different perspectives and priorities. A widely accepted and influential definition of quality of care by the Institute of Medicine (IOM) states[2]:

"The degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge."

Alongside evolving definitions, conceptual frameworks have been developed to guide the assessment and improvement of quality of care. According to the widely adopted framework by Donabedian[3], quality can be conceptualized across three core domains:

- Structure: referring to the context in which care is delivered;
- Process: referring to the interactions and procedures involved in care delivery;
- Outcomes: referring to the effects of care on patients' health status.

Structures, processes, and outcomes are the three domains used to develop quality indicators (QIs). While all three domains are important, modern health systems increasingly prioritize outcomes. This is in line with the principles of the so-called principle of value-based healthcare (VBHC). Within the VBHC framework, value is defined as the measured improvement in a person's health outcomes relative to the costs of achieving that improvement[4], with the systematic measurement of patient-level outcomes being a core principle. These outcomes include both clinical endpoints, such as survival or recurrence, and patient-reported outcomes (PROs) as measured by patient-reported outcome measures (PROMs) that capture quality of life and functional recovery, enabling healthcare systems to assess whether care truly delivers what matters most to patients. At the same time, outcomes are the most challenging domain of quality of care to measure and interpret, and their use requires robust methodology. The challenge of using outcomes to improve the quality of care, while ensuring they are measured and applied appropriately, forms the starting point of this thesis.

Part I of this thesis examines how outcome information can be used for valid and reliable hospital comparisons, thereby stimulating continuous quality improvement at the hospital level. **Part II** shifts the focus to the individual level, exploring how outcome predictions can support personalised care and shared decision-making (SDM), and improve the quality of care for individual patients.

PART I: USING OUTCOME INFORMATION TO STIMULATE CONTINUOUS QUALITY IMPROVEMENT

Clinical audits and QIs are widely used to compare hospital performance and to identify and stimulate opportunities for improvement[5 6]. By systematically collecting and analysing outcome data, healthcare institutions can benchmark their results, support professional reflections, and guide evidence informed quality improvement. For such comparisons to be meaningful, QIs must be relevant, valid, and reliable, especially when they are made publicly available[7].

For the interpretation of differences in hospital outcomes it is essential to recognise that observed differences are not solely attributable to variation in quality of care. Other factors such as statistical uncertainty and case-mix also contribute to the variation in outcomes across hospitals (Figure 1). Statistical uncertainty relates to the statistical reliability of QIs and is often assessed using rankability, which quantifies the proportion of observed variation that reflects true differences rather than random noise or chance [8 9]. Case-mix, referring to differences in patient and tumour characteristics between hospital populations, can influence QI scores. When its impact is large, scores may reflect variation in underlying populations rather than true differences in quality of care. Case-mix is therefore closely linked to the validity of QIs[9]. Depending on the QI, adjustment for case-mix may be necessary to ensure fair comparisons between hospitals.

However, balancing the need for fair comparisons with the burden of extensive data collection presents a persistent dilemma. On the one hand, adjusting for case-mix variables can improve the validity of quality indicators by accounting for patient and tumour characteristics that influence outcomes. On the other hand, each additional variable increases the registration workload and complexity of data collection. Moreover, not all case-mix factors are equally relevant: for example, adjustment is known to be more crucial for mortality indicators, while its added value for complication rates is often more limited[10-12]. **Internationally, there is little transparency about which case-mix variables are routinely collected and used for case-mix adjustment, and which methodologies are used for this purpose.** This underscores the need for a better understanding of current practices and their implications for the validity and interpretability of outcome comparisons.

Even after collecting and, where necessary, adjusting for case-mix variables, residual confounding (unmeasured or unknown patient factors influencing outcomes) and registration bias (systematic errors or inconsistencies in data recording) may still exist and are difficult to quantify (Figure 1). Being aware of these limitations is essential when interpreting QIs and evaluating hospital performance.

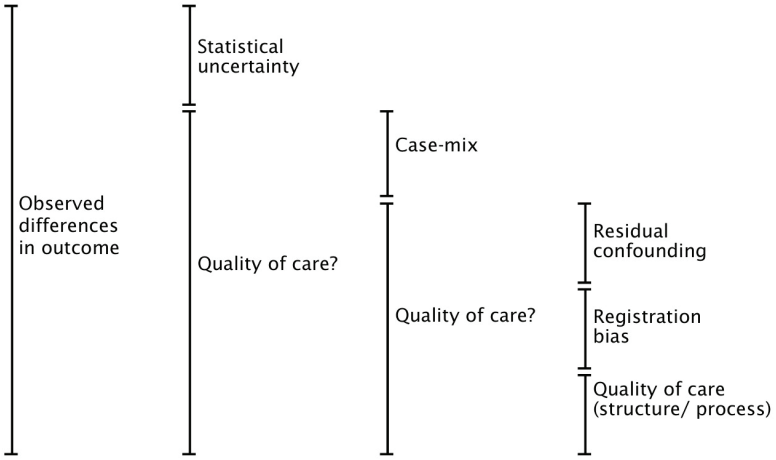


Figure 1. Sources contributing to observed differences in outcomes across hospitals[13].

NABON Breast Cancer Audit (NBCA)

NBCA, a nationwide clinical audit of Dutch breast cancer care, illustrates how outcome-based quality monitoring is implemented in practice (Box 1). NBCA integrates QI, benchmarking, and continuous evaluation. The data from NBCA is used in chapters 3 and 4 of this thesis.

Box 1: The Dutch perspective of breast cancer quality programmes

This part was previously published as part of Chapter 8: *Quality Programmes Around the Globe – Defining Quality Breast Care in the World* in the book *Breast Cancer: Global Quality Care* by D. Verhoeven, and is included here in full and unchanged [14].

The Dutch perspective on defining quality of breast cancer care: quality programmes

In the Netherlands, the National Breast Cancer Organisation Netherlands (NABON) is a nationwide multidisciplinary working group. Members are all regional breast cancer care networks and all healthcare professionals working in the breast cancer care. The goal of the NABON is to improve quality in breast cancer care with different initiatives, including the distribution and adaptation of national guidelines, standardization of data registration in the patient files, patient education, implementation of Patient Reported Outcome Measurements (PROMs) and development of quality indicators (QIs) and performing audits.

Quality monitoring by the NABON Breast Cancer Audit (NBCA)

In 2011, the multidisciplinary NABON Breast Cancer Audit (NBCA) was developed to measure the quality of care in surgically treated Dutch breast cancer patients[15]. The primary goal is nationwide monitoring of quality of breast cancer care, and the provision of feedback to individual hospitals on their results in relation to 'real-time' national benchmark information.

For the gathering and reporting of data, a collaboration with the Netherlands Comprehensive Cancer Organisation (IKNL; www.iknl.nl) and the Dutch Institute for Clinical Auditing (DICA; www.dica.nl) was established. Hospitals have the option to either self-register data through DICA's web-based system or have IKNL registrars perform the registration as part of the Netherlands Cancer Registry (NCR). A data dictionary is available to secure uniform data acquisition. When data are registered by IKNL, which is in practice in about 75% of the 72 hospitals, hospitals can check the data at patient level for inconsistencies before the data are transferred to the (national) DICA system, the indicators are calculated and partly become transparent. Patient data is anonymized before data transfer.

Participation in the NBCA is obligatory for all 72 hospitals delivering breast cancer care. Other conditions that must be achieved for breast centres are defined in the SONCOS (Oncological Collaboration Foundation) standardisation report[16]. The scientific committee of the NBCA consists of mandated members of all medical associations involved in breast cancer care (radiology, pathology, surgery, plastic surgery, radiotherapy, medical oncology, and clinical genetics), the breast cancer patient's association and a representative of the Dutch health care insurances. They define and evaluate the QIs based on their validity annually. They discuss if indicators enhance transparency, and which indicators may be adapted or removed when considered appropriate. DICA ensures the calculation of the indicators. A selection of indicators is publicly available on https://zorginzicht.nl/kwaliteitsinstrumenten/mammacarcinoom.

Individual hospitals have continuous insight into their own performance on the QIs along with other baseline information such as patient-, tumour-, and treatment characteristics; these are regularly updated on a secured website. The QIs are nationally benchmarked against the other (anonymously presented) hospitals using funnel plots showing the 95% confidence interval in conjunction with the benchmark results. The NABON sets up audit visits to the hospitals together with the scientific societies of the specialists. Moreover, regional breast cancer working groups, supported by IKNL, discuss the breast cancer care including the NBCA indicators on a regular basis.

In the last ten years, improvements were seen in nearly all QI. As examples, pre- and post-operative patients nowadays are almost all discussed in multidisciplinary meetings, and more MRIs (Magnetic Resonance Imaging) are performed prior to Neoadjuvant Systemic Therapy (NST) and for patients with lobular breast cancer (figure 2).

In the context of the NBCA, there is an annual assessment based on expert opinion of the QI set by the scientific committee. Internal and external Indicators will be critically assessed based on the criteria of relevance, validity, reliability, and feasibility, which can lead to changes in the QI set[15]. For example, the external indicator 'if a patient is discussed in multidisciplinary team' is removed from the QI set, as the QI consistently scored 100% in the last years, leaving no room for improvement. Internal indicators are evaluated to determine if they can be transitioned to external indicators (publicly available). Furthermore, new QIs can be introduced as internal indicators based on new insights and/or guideline revisions.

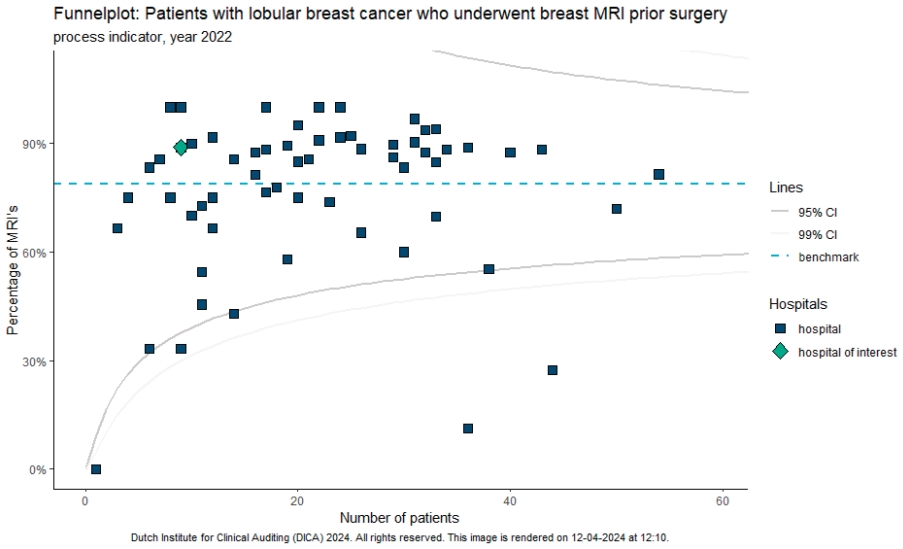


Figure 2: Quality indicators are displayed in funnel plots, providing short-term feedback. This funnel plot shows the percentage of MRI scans conducted before breast cancer surgery for patients with lobular breast cancer. Each square presents a hospital. The hospital of interest is presented as a diamond (now randomly selected), allowing to assess their performance relative to others. The dashed line represents the Dutch benchmark (mean of all hospitals), while the grey lines indicate the 95% and 99% confidence intervals.

The **Dutch Head and Neck Audit (DHNA)** is another quality registry hosted by DICA. It collects comprehensive data on patient- and tumour characteristics, treatment, complications, recurrence, and survival of patients with head and neck cancer in the Netherlands[17 18]. One of the included tumour sites is oral cavity cancer (OCC), a malignancy arising from the mucosal epithelium of the lips and oral cavity. It is often associated with tobacco and alcohol use and significantly impacts speech, swallowing, and overall quality of life[19-21]. A specific challenge in measuring quality of care for OCC, compared with breast cancer, is its low case volume, which complicates performance evaluation. Therefore, Chapter 5 of this thesis uses DHNA data to examine outcome-based quality indicators for low-incidence diseases.

Concerns about the Validity and Reliability of QIs

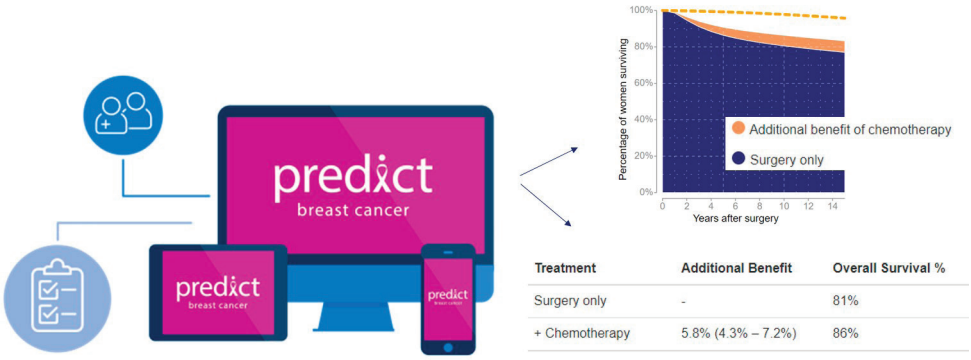
Although considerable efforts have been made to monitor and compare hospital performance using QIs, for example in the two audits introduced above, important knowledge gaps remain**.** Despite the advanced development of Dutch quality registries, **many QIs still lack robust and validated case-mix adjustment models, which may undermine their validity [9 13 22].** Furthermore, **concerns persist regarding the reliability of QIs, and it remains unclear whether this affects both low-volume cancers (e.g., head and neck cancer) and high-volume cancers (e.g., breast cancer).**

PART II: USING OUTCOME PREDICTIONS TO SUPPORT PERSONALISED CARE

Beyond their role in evaluating healthcare quality at an aggregated level, as described above, outcomes can also be used to support personalised care and shared decision-making (SDM), thereby improving the quality of care for individual patients. The focus of healthcare continues to shift towards patient-centred care, where providing individualized outcome information is essential to facilitate SDM. Prediction models address this need by predicting patient-specific outcomes based on individual characteristics, enabling more realistic expectation setting and facilitating informed treatment choices.

An example of a prediction model for breast cancer is *BREAST Predict* [23]. This tool provides individualized estimates of 5-, 10-, and 15-year breast cancer–specific survival, as well as overall survival, based on routinely available patient and tumour characteristics (Figure 3). The model also incorporates the expected benefit of various adjuvant treatment options, including chemotherapy, endocrine therapy, and trastuzumab. Models need to be regularly updated and externally validated [23-27]. The aim of such external validation is to assess whether a model remains accurate when applied outside the development cohort, for example in other geographical regions or patient settings. This process includes comparing predicted outcomes with observed outcomes (calibration) and assessing how well

the model distinguishes between outcomes, for example alive or dead (discrimination). **While earlier versions have been validated, it remains uncertain whether the most recent version of *BREAST predict* (3.1) is more accurate than the previous version (2.2), and whether the tool performs well in unselected national cohorts and in specific subgroups, such as young patients and those with lobular breast cancer.**



Treatment	Additional Benefit	Overall Survival %
Surgery only	-	81%
+ Chemotherapy	5.8% (4.3% – 7.2%)	86%

Figure 3: Concept of the BREAST Predict tool: by entering patient and tumour characteristics, the algorithm predicts the additional benefit of treatment (in this case, chemotherapy) in relation to the predicted overall survival.

Patient reported outcomes

In the context of patient-centred care and Value Based Health Care (VBHC), outcomes encompass not only traditional clinical outcomes such as complication rates or readmissions, but also patient-reported outcomes (PROs). PROs capture the patient's perspective on Health-Related Quality of Life (HRQoL), reflecting the broader impact of disease and treatment on physical, emotional, and social functioning[28]. In oncology, HRQoL is commonly assessed using generic cancer questionnaires, which provide a broad overview of the patient experience, as well as disease-specific questionnaires that focus on symptoms and functional concerns relevant to particular cancer types[29]. As such, the ability to systematically measure, compare, and predict both traditional clinical outcomes and PROs have become central to efforts to improve quality of care at individual patient level. **Despite increasing efforts to measure PROs to discuss HRQoL with patients, there remains a gap in predicting individual HRQoL trajectories over time. Developing models that use baseline HRQoL scores and patient characteristics to predict outcomes represents a novel approach with potential to strengthen personalised care.**

4D PICTURE project

The integration of prediction models into clinical practice is a key objective of the international 4D PICTURE project, which aims to improve the cancer journey for patients and ensure that personal preferences are fully respected[30]. Using MetroMapping, a

co-creation approach that visualizes care pathways, the project helps patients, families, and healthcare providers better understand treatment options and navigate complex decisions (Figure 4). By embedding personalised predictions from validated models into these care paths at decision, 4D PICTURE transforms outcome-driven care from a data-oriented concept into a patient-centered tool that actively supports SDM in daily clinical practice. Part II of this thesis was conducted within the context of the 4D PICTURE project, specifically focusing on the development and validation of outcome prediction models that can be integrated into the Metro Map for breast cancer patients.

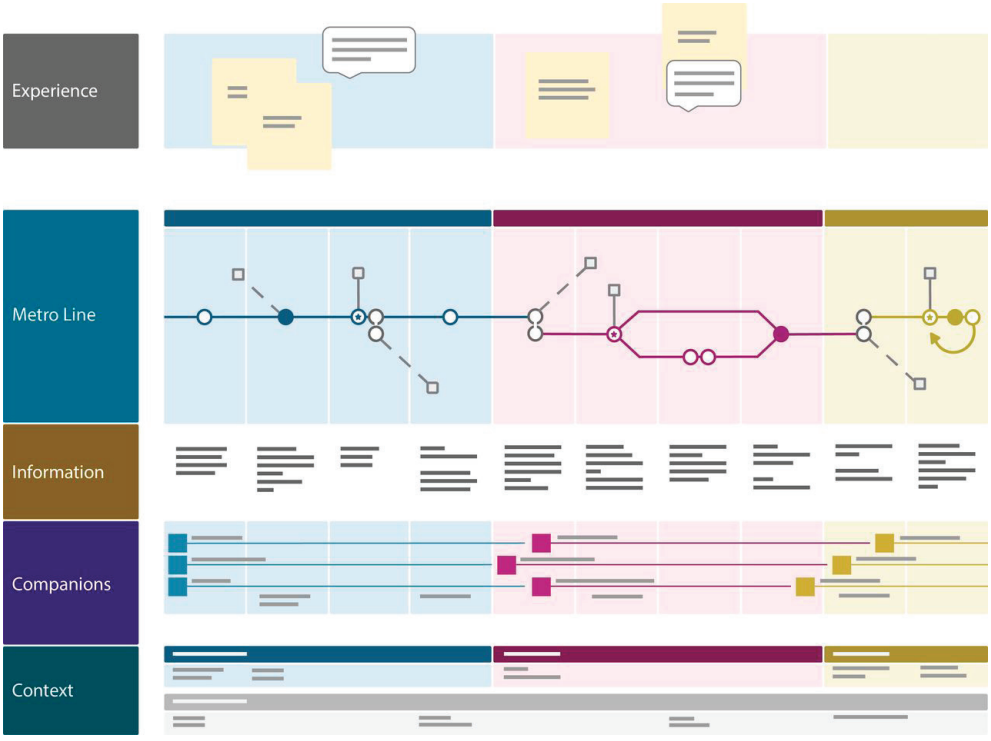


Figure 4: The Metro Map of an oncological treatment trajectory, showing the five layers of the Metro Mapping methodology: patient and clinician experiences (Layer 1), the redesigned treatment line with decision points (Layer 2), information provision (Layer 3), companions involved in care (Layer 4), and the physical context (Layer 5)[30]. Prediction tools can be integrated within Layer 2 (at decision points) and Layer 3 (as personalised information) to inform patients and clinicians about expected outcomes and thereby strengthen personalised care.

AIMS AND OUTLINE OF THIS THESIS

This thesis focuses on the central role of healthcare outcomes in optimizing the quality of cancer care. **The overall aim is to advance both the understanding and the practical application of outcome-based quality assessment in oncology.**

Part I explores how outcome information can be used for valid and reliable hospital comparisons, thereby stimulating continuous quality improvement, while **Part II** examines how outcome predictions can support personalised care and improve SDM (Table 1). Together, these approaches are essential for achieving outcome-driven, value-based cancer care.

In **Chapter 2**, we review the current landscape of cancer outcome benchmarking in Europe, focusing on quality indicators and the methodology of case-mix adjustment models. In **Chapter 3**, we develop a case-mix adjustment model for an important outcome indicator in the NBCA: complications after surgery. Since case-mix adjustment alone does not ensure the validity and reliability of an indicator, we introduce a structured framework to evaluate quality indicators on feasibility, discriminative ability, validity, and reliability (**Chapter 4**), and apply it to both breast cancer (a high-incidence cancer) and oral cavity cancer (a low-incidence cancer; **Chapter 5**).

In **Chapter 6**, we externally validate the updated PREDICT tool (version 3.1) for supporting clinical decision in Dutch and Swedish breast cancer patients, with a focus on lobular breast cancer and younger patients, two groups for whom outcomes are harder to predict and more uncertain. In **Chapter 7**, we compare version 3.1 of the PREDICT tool with the previous version 2.2, using data from the Dutch population and 36 clinically relevant subgroups. Finally, in **Chapter 8**, we develop models to predict trends of health-related quality of life (HRQoL) after breast cancer surgery and reconstruction, using 15 different HRQoL outcomes, aiming to better support patients and clinicians in managing expectations during SDM.

Table 1: Overview of Research Gaps and Corresponding Research Questions.

Research Gap	Research Question	Addressed in Chapter
Part I: Using outcome information to stimulate continuous quality improvement		
Limited insight into which cancer outcomes are nationally benchmarked across hospitals in European countries, how extensively case-mix adjustment is applied, and which methodologies are used for this purpose.	What is currently known about cancer outcome performance comparison and the use of case-mix adjustment models in European countries?	Chapter 2
Uncertainty about the validity and reliability of widely used quality indicators.	How can we assess if quality indicators are valid and reliable? Are there differences in benchmarking hospital performance between high- and low-incidence cancers?	Chapters 3, 4 & 5
Part II: Using outcome predictions to support personalised care		
Limited evidence on whether existing prediction models can be reliably used for specific patient subgroups in the context of SDM.	Is the most widely used prediction model PREDICT in breast cancer care valid for guiding personalised, evidence-based treatment decisions across different geographical regions and specific patient groups?	Chapter 6, 7
Limited evidence on the ability to predict PROs and their potential use in expectation management.	Can we develop models to predict individual trajectories of HRQoL?	Chapter 8

SDM = shared decision making; PROs = Patient Reported Outcomes; HRQoL = Health-Related Quality of Life

REFERENCES

1. World Health Organization. Quality of care. Retrieved August 6, from https://www.who.int/health-topics/quality-of-care#tab=tab_1.
2. Institute of Medicine Committee to Design a Strategy for Quality Review and Assurance in Medicare; Lohr KN, editor. Medicare: A Strategy for Quality Assurance: VOLUME II Sources and Methods. Washington (DC): National Academies Press (US); 1990. 5, Defining Quality of Care. Available from: https://www.ncbi.nlm.nih.gov/books/NBK235476/.
3. Donabedian A. The quality of care. How can it be assessed? JAMA. 1988;260(12):1743-8.
4. Porter ME, Teisberg EO. Redefining Health Care: Creating Value-Based Competition on Results. 2006Boston, MA: Harvard Business School Press.
5. Maes-Carballo M, Gomez-Fandino Y, Reinoso-Hermida A, Estrada-Lopez CR, Martin-Diaz M, Khan KS, Bueno-Cavanillas A. Quality indicators for breast cancer care: A systematic review. Breast. 2021;59:221-31.
6. Beck N, van Bommel AC, Eddes EH, van Leersum NJ, Tollenaar RA, Wouters MW, Dutch Clinical Auditing Group. The Dutch Institute for Clinical Auditing: Achieving Codman's Dream on a Nationwide Basis. Ann Surg. 2020;271(4):627-31.
7. Fischer, C. (2015). Quality indicators for hospital care: Reliability and validity [Doctoral dissertation, Erasmus University Rotterdam]. Erasmus University Repository. http://hdl.handle.net/1765/79066.
8. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. BMJ Qual Saf. 2011;20(10):869-74.
9. Vos EL, Lingsma HF, Jager A, Schreuder K, Spronk P, Vrancken Peeters M, et al. Effect of Case-Mix and Random Variation on Breast Cancer Care Quality Indicators and Their Rankability. Value Health. 2020;23(9):1191-9.
10. Lingsma HF, Bottle A, Middleton S, Kievit J, Steyerberg EW, Marang-van de Mheen PJ. Evaluation of hospital outcomes: the relation between length-of-stay, readmission, and mortality in a large international administrative database. BMC Health Serv Res. 2018;18(1):116.
11. Talsma AK, Lingsma HF, Steyerberg EW, Wijnhoven BP, Van Lanschot JJ. The 30-day versus in-hospital and 90-day mortality after esophagectomy as indicators for quality of care. Ann Surg. 2014;260(2):267-73.
12. Algera MD, Baldewpersad Tewarie NMS, Driel WJV, van Ham M, Slangen BFM, Kruitwagen R, et al. Case-mix adjustment to compare hospital performances regarding complications after cytoreductive surgery for ovarian cancer: a nationwide population-based study. Int J Gynecol Cancer. 2023;33(4):534-42.
13. Lingsma, H. (2010, November 26). Measuring quality of care: Methods and applications to acute neurological diseases. Retrieved from http://hdl.handle.net/1765/21742.
14. . Siesling, S, Verheul, EM, et all (2026) Chapter 07 Defining quality breast cancer care around the globe In D Verhoeven et al, Breast cancer: Global quality care (2e ed) Oxford University Press
15. van Bommel AC, Spronk PE, Vrancken Peeters MT, Jager A, Lobbes M, Maduro JH, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: The national NABON Breast Cancer Audit. J Surg Oncol. 2017;115(3):243-9.
16. Platform Oncologie - SONCOS fms. SONCOS normeringsrapport versie 11, 2023: Multidisciplinaire normering oncologische zorg in Nederland. https://demedischspecialist.nl/normeringsrapport-van-soncos 16-02-2023.
17. van Oorschot HD, Hardillo JA, Baatenburg de Jong RJ, Dutch H, Neck Audit- Oral Cavity Collaborator G. Cohort profile: the Dutch oral cavity cancer cohort. BMJ Open. 2025;15(5):e092467.

18. van Overveld LFJ TR, Smeele LE, et al. . The Dutch Head and Neck Audit: The First Steps. J Head Neck Surg 2018;1(1):1-8.
19. Chamoli A, Gosavi AS, Shirwadkar UP, Wangdale KV, Behera SK, Kurrey NK, et al. Overview of oral cavity squamous cell carcinoma: Risk factors, mechanisms, and diagnostics. Oral Oncol. 2021;121:105451.
20. Xu H, Gao Z, Liu H, An L, Yang T, Zhang B, et al. Associations of lifestyle factors with oral cancer risk: An umbrella review. J Stomatol Oral Maxillofac Surg. 2025;126(3S):102234.
21. Dzebo S, Mahmutovic J, Erkocevic H. Quality of Life of Patients with Oral Cavity Cancer. Mater Sociomed. 2017;29(1):30-4.
22. Fischer C, Lingsma HF, van Leersum N, Tollenaar RA, Wouters MW, Steyerberg EW. Comparing colon cancer outcomes: The impact of low hospital case volume and case-mix adjustment. Eur J Surg Oncol. 2015;41(8):1045-53.
23. Grootes I, Wishart GC, Pharoah PDP. An updated PREDICT breast cancer prognostic model including the benefits and harms of radiotherapy. NPJ Breast Cancer. 2024;10(1):6.
24. Hsiao YW, Wishart GC, Pharoah PDP, Peng PC. Assessing the Performance of the PREDICT Breast Version 3.0 Prognostic Tool in Patients With Breast Cancer in the United States. J Natl Compr Canc Netw. 2025;23(6):227-33.
25. van Maaren MC, van Steenbeek CD, Pharoah PDP, Witteveen A, Sonke GS, Strobbe LJA, et al. Validation of the online prediction tool PREDICT v. 2.0 in the Dutch breast cancer population. Eur J Cancer. 2017;86:364-72.
26. Zaguirre K, Kai M, Kubo M, Yamada M, Kurata K, Kawaji H, et al. Validity of the prognostication tool PREDICT version 2.2 in Japanese breast cancer patients. Cancer Med. 2021;10(5):1605-13.
27. Nita S. Nair BK, Sudeep Gupta, Sadhana Kanann, Vaibhav Vanmali, Rohini Hawaldar, Ashutosh Tondare, Shabina Siddique, Vani Parmar, Shalaka Joshi, and RA Badwe. Validation of PREDICT Version 2.2 in a Retrospective Cohort of Indian Women With Operable Breast Cancer. JCO Global Oncology. 2023;9.
28. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. J Nurs Scholarsh. 2005;37(4):336-42.
29. Churruca K, Pomare C, Ellis LA, Long JC, Henderson SB, Murphy LED, et al. Patient-reported outcome measures (PROMs): A review of generic and condition-specific measures and a discussion of trends and issues. Health Expect. 2021;24(4):1015-24.
30. Rietjens JAC, Griffioen I, Sierra-Perez J, Sroczynski G, Siebert U, Buyx A, et al. Improving shared decision-making about cancer treatment through design-based data-driven decision-support tools and redesigning care paths: an overview of the 4D PICTURE project. Palliat Care Soc Pract. 2024;18:26323524231225249.

PART I

USING OUTCOME INFORMATION TO STIMULATE CONTINUOUS QUALITY IMPROVEMENT

2

Chapter 2

Benchmarking Cancer Outcomes in Europe: A Scoping Review of Methodologies and Case-Mix Adjustments

Jacob Thurell, Lea Doppelbauer, Elfi M. Verheul, Iurii Petrov, Maria Margarete Karsten, Linetta B. Koppert, Jonas Bergh, Irma Fredriksson, Peter Lindgren, Narsis Kiani, Elham Hedayati.

Published in: ESMO Real-World Data and Digital Oncology, 2025.

ABSTRACT

Background: Benchmarking hospital outcomes is crucial for identifying inequities and improving cancer care. Meaningful comparisons require selecting relevant outcomes and adjusting for case-mix factors such as age, comorbidity, and stage. Without case-mix adjustment, hospitals may be unfairly assessed based on patient mix rather than care quality. No prior review has examined benchmarking practices in European cancer care. This scoping review addresses: (i) Which health outcomes are frequently benchmarked? (ii) What case-mix factors are commonly used for adjustment? (iii) Which statistical approaches are utilized? (iv) How are case-mix models developed and evaluated?

Methods: We conducted a systematic scoping review searching OVID MEDLINE, Web of Science, and EMBASE. Eligible studies focused on benchmarking populations with a cancer diagnosis, involved European hospitals, and evaluated health outcomes like survival. Abstract screening and full-text appraisal were done independently by two authors. Data were extracted into a pre-specified matrix, and results synthesized by research question.

Findings: After screening 4953 abstracts, 65 studies were included. Key gaps include a lack of validated case-mix models, under-representation of long-term outcomes, and a tendency to 'over-adjust' by including treatment factors in case-mix models, potentially obscuring true differences in performance. Regression modelling remains the gold standard for adjustment. A consensus is needed on reporting and evaluating case-mix models, akin to TRIPOD guidelines.

Interpretation: A shift toward standardized, validated benchmarking practices is essential to drive health care improvements. Only through rigorous methodologies, standardized reporting, and international collaboration can hospital benchmarking become a transformative tool for improving cancer care quality and patient outcomes.

HIGHLIGHTS

- Cancer benchmarking lacks clear standards on methods for case-mix adjustment.
- Benchmarking is concentrated in two countries, limiting learning across Europe.
- Lack of long-term survival and patient-reported outcomes in benchmarking studies.
- Case-mix models lack validation and clear reporting, undermining fair comparisons.
- We propose practical guidance on case-mix modeling to support fair benchmarking.

INTRODUCTION

In many countries there is a growing interest in assessing health care quality and equity. Over the past few decades, benchmarking has emerged as a vital tool to address these demands. By systematically comparing performance among health care institutions, benchmarking identifies best practices and drives improvements in underperforming hospitals.[1] In the rapidly evolving field of cancer care, continuously monitoring real-world outcomes is essential to ensure all patients receive optimal treatment, regardless of the hospital they attend. However, benchmarking must be done carefully. Inappropriate practices, especially when results are publicly reported, risk drawing misleading conclusions that can unfairly harm hospitals and erode public trust in health care systems.

A key challenge in benchmarking is determining which quality metrics to measure. Although process indicators assess specific care steps and guideline adherence, outcomes provide a more comprehensive, patient-centred view of health care performance by capturing the effectiveness of the entire care process.[2] However, benchmarking outcomes require careful consideration of confounding factors. Hospitals serve diverse patient populations with varying characteristics (i.e. case-mix) related to age, disease severity, and socioeconomic factors influencing outcomes. Without robust statistical adjustments for these factors—over which hospitals have no control—benchmarking may yield unfair comparisons that distort the accurate picture of performance.[3]

Despite the growing use of benchmarking, no review has systematically examined its application in European cancer care. This scoping review aims to describe how benchmarking of cancer outcomes has been applied in European countries, specifically addressing:

- Which health outcomes are most frequently benchmarked?
- What case-mix factors are commonly used for adjustment?
- Which statistical approaches are utilized for case-mix adjustment?
- How are case-mix models developed, and quality assessed?

Effective benchmarking is essential for ensuring high-quality and equitable cancer care. If poorly implemented, it risks compromising the clinical pathways and care processes it is designed to enhance. This review aims to provide a foundation for more accurate and transparent benchmarking practices that can transition successfully to real-world applications.

METHODS

Search strategy and selection criteria

This scoping review was conducted in accordance with the PRISMA guidelines.[4] Studies were eligible if they focused on benchmarking populations with a cancer diagnosis, involved European hospitals or regions, and evaluated health outcomes (e.g. survival, patient-reported outcomes, or post-operative complications). This review defined Europe as the 27 member states of the European Union (EU), along with Norway, the UK and Switzerland, given their comparable health care systems, and applying case-mix adjustments (e.g. regression modelling and advanced machine learning). Studies comparing variations between hospitals or cancer networks/regions within countries were included. To ensure the inclusion of rigorous research, we included peer-reviewed studies only, published in English, Swedish, Dutch, or German before 11 November 2024. Studies were ineligible if their models aimed solely at individual prediction as this review focused on case-mix adjusted models designed to compare the quality of care across health care providers rather than to predict individual patient outcomes. Studies focusing exclusively on process indicators or conducted outside Europe were excluded, to limit the scope of the review. Cross-sectional studies were also excluded as we wanted to track outcomes longitudinally. Finally, we excluded studies without case-mix adjustment, as assessing case-mix adjustment methods is a key aim of this review. We searched OVID MEDLINE, Web of Science, and EMBASE using structured search terms related to 'cancer,' 'case-mix adjustment,' and 'benchmarking.' The full search strings are presented in Supplementary Appendix 1, available at https://doi.org/10.1016/j.esmorw.2025.100176.

Data analysis

Two reviewers independently screened titles and abstracts using Rayyan.ai, resolving conflicts through discussion or consultation with a third reviewer. Full-text screening followed the same procedure. Studies excluded after full-text screening are described in Supplementary Appendix 2. For quality appraisal, we applied the Newcastle-Ottawa Scale[5] for cohort studies and incorporated elements from the Joanna Briggs Institute (JBI) scale[6] to evaluate the robustness of statistical methods. No studies were excluded based on the quality appraisal but instead contributed to synthesizing the results. The full quality appraisal is described in Supplementary Appendix 3. Data were extracted using a predefined matrix based on the JBI framework for scoping reviews,[7] focusing on population, intervention, and context (Supplementary Appendix 4. Results were synthesized according to predefined research questions.

Role of funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

RESULTS

We screened 2571 articles based on title and abstract. Eighty-three articles were retrieved in full text, and 57 met the eligibility criteria. We also screened all included articles' reference lists and hand searched for additional articles. This resulted in 8 further added articles, summing to a total of 65 articles[8-72] included in the review (Figure 1), and a total of 1 304 433 cases from eight European countries (Supplementary Appendix 4). Fifty-two studies (80%) were conducted in either the Netherlands or the UK, whereas the remaining 13 studies (20%) represented other European countries. Colorectal cancer was the most commonly benchmarked cancer type, accounting for 31% of included studies, followed by head and neck cancers and lung cancer (Table 1).

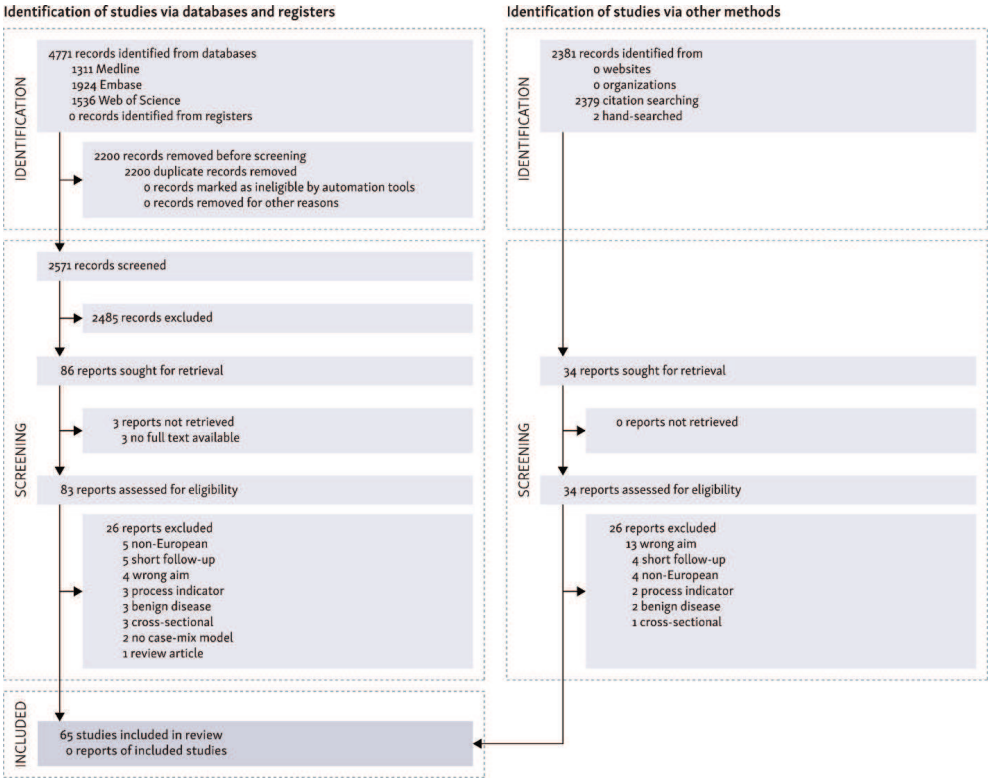


Figure 1: Study selection.

Table 1: Summary of the included studies.

Category	Subcategory	Studies, *n* (%)
Countries	The Netherlands	29 (44.5)
	United Kingdom	23 (35)
	France	3 (5)
	Germany	3 (5)
	Norway	2 (3)
	Sweden	2 (3)
	Denmark	1 (1.5)
	Austria	1 (1.5)
	European collaborations	1 (1.5)
Cancer types	Colorectal	20 (31)
	Head and neck	9 (14)
	Lung	9 (14)
	Gastroesophageal	6 (9)
	Breast	6 (9)
	Liver metastasis	5 (8)
	Gynaecological	5 (8)
	Skin	1 (1.5)
	Brain	1 (1.5)
	Leukaemia	1 (1.5)
	Prostate	1 (1.5)
	Several cancer types	1 (1.5)
Case-mix factors	Basic demographics	65 (100)
	Tumour factors	54 (83)
	Patient factors	51 (78)
	Socioeconomic factors	14 (22)
	Treatment factors	39 (60)

Seventy-seven outcomes were benchmarked across the 65 included studies. Fifty-three studies (82%) benchmarked short-term outcomes occurring within 90 days of surgery and/or oncological treatment. The most commonly benchmarked short-term outcomes were 30-day post-operative mortality and 30-day post-operative morbidity. Seven studies benchmarked so-called 'textbook outcomes' (TOs). TO is a composite measure of process indicators and outcomes specific to each cancer type that reflect high-quality care. It is an all or none score for each patient. These typically include no death or severe complications within 30 days post-surgery, alongside other criteria such as no readmission and sufficient surgical margins. Failure to rescue (FTR) was used in three studies. FTR assesses mortality rates among patients who develop severe complications after major surgery (Figure 2).

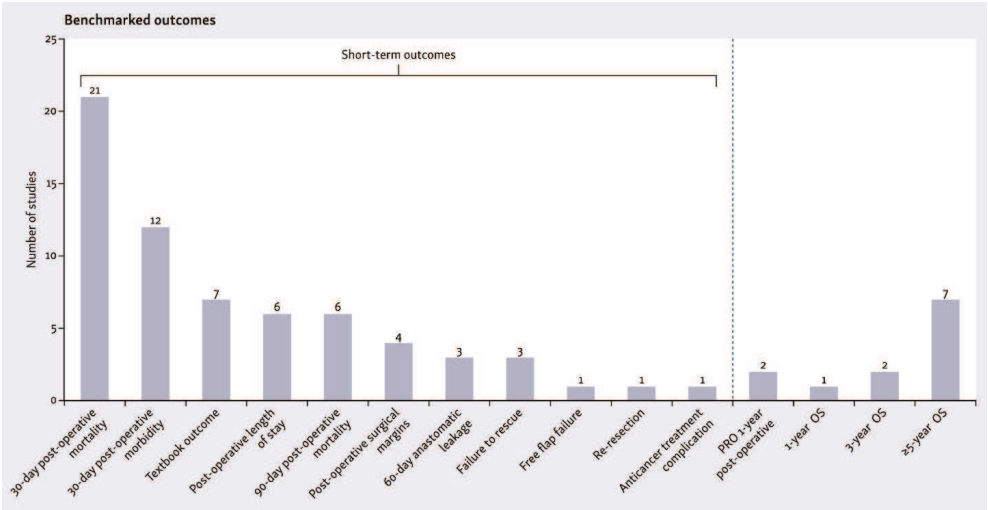


Figure 2: Outcomes that were benchmarked in the included studies[a]. [a]Some studies benchmarked more than one outcome.

Long-term outcomes (here defined as follow-up of at least 1 year) were benchmarked in 12 studies (18%) (Table 1). Of these, 10 studies benchmarked survival outcomes, whereas the remaining two studies benchmarked patient-reported outcomes (PROs) (Figure 2).

All studies accounted for either age or sex in their adjustments. Most studies also adjusted for tumour characteristics (e.g. stage) and patient characteristics (e.g. comorbidity burden, Eastern Cooperative Oncology Group performance status). Fourteen studies (22%) incorporated socioeconomic factors, primarily using country-specific social status indices based on postal codes.[10,15,27,29,30,32,36,39,41,42,46-48,70] All but two of these studies were from the UK. Sixty percent of studies (n = 39) adjusted for treatment factors within hospital control.[8,10-13,16-23,25,26,31,33-38,47-50,54-60,62-64,66,67,72] The most common treatment factors adjusted for were surgical complexity, surgical approach, and preoperative systemic treatment and/or radiotherapy (Table 1).

Of the 65 included studies, 18 (28%) failed to describe the content of their case-mix models. Furthermore, 61 studies (94%) developed new case-mix models, whereas the remaining 4 focused on externally validating previously established models. None of the four external validation studies[18,21,58,60] succeeded in their validation as either discrimination or calibration differed significantly between the development and validation cohorts (Supplementary Appendix 4).

Classical regression models were the predominant approach for case-mix adjustment, serving as the primary method in 56 studies (86%). In studies applying regression models for short-term binary outcomes (up to 90 days), logistic regression was uniformly ap-

plied. In contrast, long-term survival analyses (spanning 1-10 years) exhibited greater methodological heterogeneity. Although Cox proportional hazards modelling was the most common approach—employed in five studies—other methods were also used, including parametric regression ($n = 2$),[32,48] logistic regression for discrete survival benchmarks ($n = 2$),[36,40] and multilevel latent class modelling (LCM) ($n = 1$).[30] LCM, which groups patients into unobserved subgroups based on data patterns, yielded hospital rankings comparable with those obtained from traditional multilevel regression, but provided no in-depth comparative analysis between the two methods. Furthermore, linear regression was applied in two studies to benchmark PROs[39,47] (Figure 3).

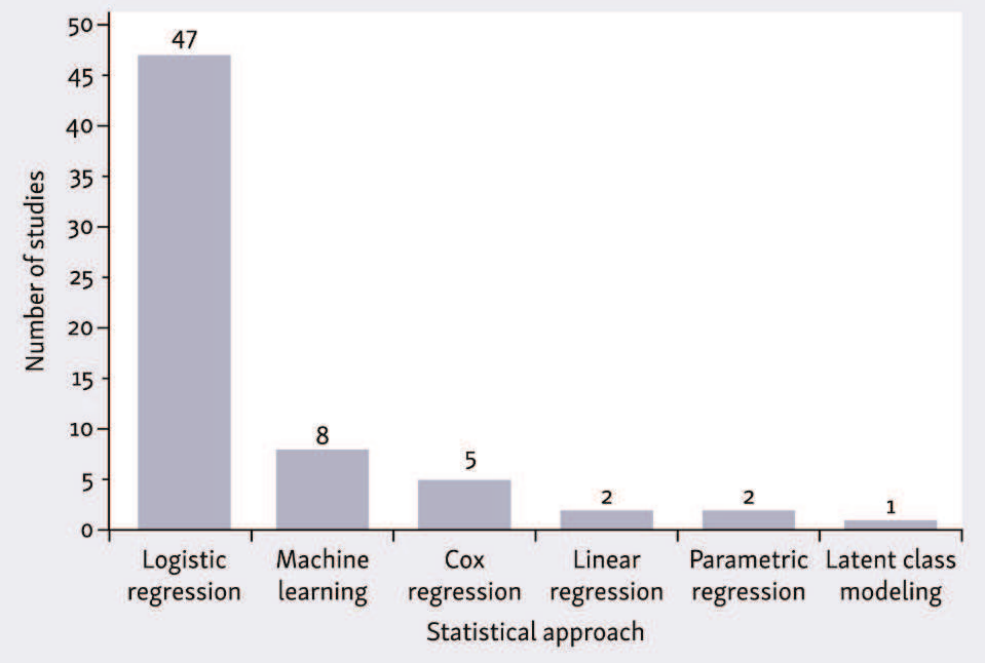


Figure 3: Distribution of the main statistical approach applied for case-mix adjustment in the included studies.

Machine learning (ML) methods, including techniques beyond traditional regression such as decision trees, random forests, and neural networks were applied in eight studies (13%), seven of them by the same research group (Tighe et al.). One of these compared logistic regression with three ML methods for adjusting post-operative complication rates after head and neck cancer surgery[54] and found that a Naïve Bayes model (a probabilistic classifier) slightly outperformed traditional logistic regression in terms of discrimination [area under the curve (AUC) 0.72 versus 0.68]. The only study outside this research group benchmarked 30-day post-operative mortality rates in colorectal cancer, comparing logistic regression and three advanced ML techniques. Elastic net regression—a method that improves predictions by combining two regularization techniques (Least Absolute

Shrinkage and Selection Operator, and ridge regression)—outperformed logistic regression by a small margin (AUC 0.82 versus 0.81) (Supplementary Appendix 4).

Two primary strategies were used to assess the performance of case-mix models. The first involved visualizing the comparison between crude and adjusted outcomes. The second focused on statistical performance metrics, primarily discrimination (i.e. the model's ability to distinguish between patients who experienced the outcome and those who did not) and calibration (i.e. the alignment between predicted and observed outcomes across different risk groups). Of the 65 studies, 18 neither visualized crude to adjusted outcomes, nor provided any statistical performance metrics. Fourteen studies relied solely on visualizing the crude to adjusted outcomes, whereas 16 studies used only statistical metrics, such as discrimination and calibration. Finally, 17 studies provided both visualizations and statistical performance metrics (Supplementary Appendix 4).

Twenty-five studies explicitly aimed to develop case-mix adjustment models for benchmarking between hospitals. However, seven of these studies did not describe the details of their models, and six failed to report visual or statistical performance metrics. Internal validation was conducted in 12 of the 25 studies. Seven[13,14,27,40,59,64,71] of the 25 model development studies provided sufficient detail regarding their models, reported performance metrics, and carried out internal validation, ensuring transparency, validity, and reproducibility (Supplementary Appendix 4).

DISCUSSION

This scoping review evaluated hospital-based benchmarking of cancer outcomes across Europe. Of 4950 screened articles, 65 met the inclusion criteria, encompassing studies from eight countries, primarily the Netherlands and the United Kingdom. Most studies benchmarked short-term post-operative outcomes, such as 30-day mortality and morbidity. Surprisingly, only 10 studies focused on long-term survival, and just 2 addressed PROs. Case-mix adjustments commonly included demographic variables, comorbidity indices, functional status assessments, and tumour stage, whereas socioeconomic factors were notably underrepresented. More than half of the studies inadequately adjusted for treatment variables within the control of participating hospitals, potentially obscuring true performance differences. Traditional regression models remain the gold standard for case-mix adjustment, with advanced ML methods offering limited performance gains at the expense of transparency. The optimal approach for evaluating case-mix models remains unclear. Visualizing crude to adjusted outcomes enhances clinical interpretability, whereas discrimination and calibration metrics provide statistical validation. Ideally, both approaches should be combined to ensure both clinical relevance and methodological robustness. Furthermore, there is no consensus on how to report case-mix models. Of the

25 studies specifically aimed at developing case-mix models, only 7 provided sufficient detail for reproducibility, carried out internal validation, and presented performance metrics.

The strengths of this scoping review include a rigorous method with two blind reviewers screening and assessing the full text of all the papers. The second main strength is that the team that conducted this review is multinational with extensive experience in benchmarking cancer outcomes in Europe. The main limitation of this review is its geographical focus on Europe. A global scope would have been valuable but was deemed too extensive. Even within Europe, variations in health care data pose challenges, and expanding internationally would introduce greater heterogeneity, potentially compromising validity. However, given Europe's advanced role in benchmarking research, we expect this focus to yield meaningful insights. A second limitation is the exclusion of gray literature, which could have increased comprehensiveness but also increased the risk of low-quality data leading to misleading conclusions.

Most studies benchmarked short-term outcomes within 30 days post-surgery, likely due to their ease of measurement and potential to drive rapid quality improvements. However, such focus risks misaligning hospital priorities by emphasizing immediate postoperative results over long-term outcomes. Short-term outcomes often reflect isolated care steps, such as surgery, whereas long-term outcomes capture the success of the entire care continuum, promoting cross-disciplinary collaboration. A balanced benchmarking framework should incorporate short-term actionable metrics and long-term outcomes to provide a comprehensive assessment of care quality, ensuring health care systems prioritize outcomes that truly matter to patients. PROs reflect patients' perspectives on well-being, symptom burden, and functional status, yet they are underutilized in cancer care benchmarking. Only two studies[39,47] benchmarked PROs, exposing a significant gap in practice. A reason for this could be that PROs are not yet routinely collected in clinical practice. PROs are inherently influenced by individual, social, and psychological factors, and, as a recent review highlighted,[73] there is a need for a standardized approach to case-mix adjustment for PROs. Further research is needed to determine whether, how, and which PROs can be used for valid benchmarking. TOs present a promising approach to capture different aspects of quality of care into a single metric. Achieving TO has also been associated with improved long-term survival in several studies,[74,75] potentially bridging the gap between short- and long-term outcomes. However, a TO is usually binary in nature, where the outcome is achieved if all included subcomponents are met. This all-or-nothing approach can obscure actionable insights, as underperformance offers no clarity on whether issues stem from surgical technique, perioperative care, or follow-up. Assessing individual TO components may help hospitals identify targeted areas for improvement. Moreover, none of the studies using TOs reported case-mix model validation metrics, such as discrimination or calibration, which are critical for assessing

reliability. Standardizing cancer outcome measures across countries is essential for effective international benchmarking. This can be achieved, for example, by aligning with the International Consortium of Health Outcome Measurement sets.[76]

Sixty percent of the studies adjusted their results for treatment-related variables, such as the choice of oncological interventions or surgical techniques. The choice of treatment clearly affects the risk of complications and mortality. It may therefore seem intuitive to adjust for these factors. However, this approach poses challenges in benchmarking. Treatment-related adjustments are often within hospitals' direct control and adjusting for them may inadvertently mask hospital-level variations that could highlight areas for improvement. Moreover, because clinical guidelines define standard care practices, adjusting for treatment variables may conceal deviations from these recommendations that signal opportunities for enhancing care. A more effective strategy is to adjust for tumour stage or other indicators of disease complexity, reflecting the expected difficulty of treatment without obscuring hospital performance in delivering the most appropriate care.

One significant concern arising from our findings is the low representation of socioeconomic factors in case-mix adjustments. Although socioeconomic status is a well-established determinant of health outcomes, it was often not considered in the included studies. This omission may lead to unfair benchmarking by obscuring the true performance of hospitals serving more disadvantaged populations. Among the 14 studies that included socioeconomic adjustors, 6 did not report sufficient model details to assess their impact. The findings were mixed among the remaining eight studies. Future research should clarify whether including socioeconomic factors enhances the accuracy of case-mix models, thereby supporting more equitable benchmarking.

The choice of case-mix variables is crucial, as these directly impact the interpretation of outcomes across hospitals. We argue that research teams developing these models should adopt an interdisciplinary approach—engaging experienced clinicians to identify clinically relevant factors alongside biostatisticians and epidemiologists with expertise in statistical modelling. This collaboration ensures that the selected variables are both clinically pertinent and methodologically robust. Moreover, transparency in the selection and application of case-mix factors is vital for reproducibility and for fostering trust in benchmarking results. Another key consideration is that outcome indicators are differentially influenced by case-mix. A recent study[77]—published after our review's cut-off date—reported that the type of breast cancer surgery is the predominant predictor of post-operative complication rates. After stratifying by type of surgery, additional case-mix adjustment did not alter hospital rankings, suggesting a negligible impact of such adjustments in that context. Given that collecting case-mix data is resource intensive, it is imperative to critically evaluate its impact to ensure that the effort yields meaningful improvements in benchmarking.

Our analysis highlights the predominance of traditional regression modelling for case-mix adjustment, reflecting their methodological reliability and widespread acceptance. Classical regression generally performs well when relationships between predictors and outcomes are fairly linear, and the key variables are well understood. However, these models can struggle to capture complex nonlinear relationships and interactions between case-mix factors, which in turn could favour more complex modelling.

LCM offers a novel approach by identifying unobserved patient subgroups with distinct risk profiles. This represents a promising intermediate method with characteristics of both classical regression and advanced ML. LCM could potentially uncover hidden subgroups of risk that classical regression fails to capture, but it is yet to be explored whether this theoretical advantage can translate into real-world improvements in cancer outcome benchmarking, and in which contexts these advantages outweigh the increased model complexity and limited transparency.

Only a small number of studies in our review employed more advanced ML techniques, such as decision trees or gradient boosting. This limited exploration may represent a missed opportunity for innovation, but it must be weighed against important considerations. Advanced ML models generally require large, high-quality datasets to avoid overfitting and achieve reliable performance—a condition not always met in real-world oncology data. The added complexity and 'black box' nature of many ML models have rightly raised concerns about interpretability and clinical trust. However, advances in explainable artificial intelligence (AI) techniques[78] offer promising avenues to enhance transparency and support clinical adoption.

In our review, advanced ML methods showed only modest performance gains compared with classical regression. Notably, most of the included studies developed models on relatively small datasets, which likely limited the potential of ML approaches. One exception was a study using a dataset of 62 000 patients and 103 preoperative variables.[64] Even in this large, rich dataset, advanced ML techniques—including random forests, gradient boosting, and elastic net regression—provided none to minimal improvements in discrimination over classical logistic regression. This aligns with previous evidence showing that advanced ML techniques often do not outperform logistic regression in clinical prediction modeling.[79]

Ultimately, the choice of statistical approach for case-mix adjustment should be guided by the clinical context. Simple, transparent regression models may suffice—and even excel—when predictor–outcome relationships are well understood or when sample sizes are limited. A pragmatic approach is to consider advanced ML methods when classical regression performs inadequately—such as in cases of poor calibration, low discrimination, or limited clinical plausibility. Scenarios involving complex nonlinear relationships,

high-order interactions, or a large number of predictors could theoretically favour ML. However, without clear evidence of added value, simpler and more interpretable models remain the more practical and defensible choice in most benchmarking contexts. A systematic framework for evaluating when ML techniques offer meaningful advantages would help advance methodological rigor in cancer outcome benchmarking.

There is currently no consensus on how to best evaluate case-mix models. In our review, two main strategies emerged: some studies relied on visualizations of hospital outcomes before and after adjustment, whereas others used statistical performance metrics like discrimination and calibration, adapted from prediction modelling. However, it is important to recognize that case-mix modelling serves a different purpose than prediction modelling. While prediction models aim to estimate an individual patient's risk, case-mix models are used to adjust hospital-level outcomes to account for differences in patient populations. This means the priorities for model assessment are also different. In prediction modelling, discrimination—the ability to distinguish between patients who experience the outcome and those who do not—is often emphasized. In contrast, case-mix models prioritize calibration, ensuring that the model accurately predicts outcomes across different risk groups so that hospitals treating sicker or more complex patients are not unfairly penalized.

Although metrics like discrimination and calibration provide an objective view of model performance, they do not show how case-mix adjustment affects hospital rankings or performance comparisons. Visualizations—such as plots of hospital outcomes before and after adjustment—offer a clearer and more intuitive view of how the model works in practice. A combined approach, using both performance metrics and visualizations, offers the most balanced assessment. It captures both statistical validity and practical impact and helps ensure the adjusted outcomes are both fair and clinically meaningful.

Although the TRIPOD guidelines[80] have set a benchmark for the transparent reporting of individual prediction models, no analogous guidelines exist for case-mix modelling. In our review only 7 of 25 studies that explicitly aimed to develop case-mix models provided sufficient detail regarding their models for reproducibility, reported performance metrics, and carried out internal validation. In Table 2 we summarize our suggestions for developing case-mix models in a cancer outcome benchmarking context. These suggestions are intended to support more transparent, reproducible, and clinically meaningful benchmarking efforts, particularly in the absence of formal consensus guidelines. The development of dedicated reporting standards for case-mix modelling would facilitate standardized model description, validation, and assessment across studies. This, in turn, would enable more meaningful cross-study comparisons, helping to identify which disease areas and outcome domains are well served by current benchmarking practices and where methodological improvements are still needed. Moreover, standardized practices

are essential for advancing international benchmarking, as it remains an open question whether case-mix models can be generalized across health care systems and countries. Furthermore, standardized guidelines would facilitate external validation and model reuse, improve peer review practices, and enable systematic evidence synthesis—critical steps toward establishing benchmarking as a rigorous, generalizable, and collaborative field.

To our knowledge, this is the first scoping review on benchmarking practices in cancer care. Hospital benchmarking of cancer outcomes is essential to improving both quality and equity in cancer care across Europe. However, significant methodological challenges remain. A shift toward standardized, validated benchmarking practices is essential to

Table 2: Recommendations for conducting benchmarking studies.

Category	Suggestion
Case-mix factors	Select case-mix factors based on clinical expertise and prior knowledge (i.e. hypothesis-driven selection) Case-mix variables should be available before treatment (i.e. at the time of diagnosis or treatment planning) Avoid including treatment factors in the case-mix model Consider adding socioeconomic factors if available and deemed relevant
Statistical approach	Transparent regression can be applied in most cancer benchmarking settings Consider advanced machine learning methods when traditional regression shows poor performance, especially in settings with complex, high-dimensional, or nonlinear data
Model assessment	Clarify that the model is intended for benchmarking purposes—not individual outcome prediction—as this guides how performance metrics like calibration and discrimination should be interpreted, and how the model can be used Statistical performance should, at minimum, be assessed by model discrimination and model calibration Assess clinical validity using visualizations that show how hospital outcomes shift from crude to adjusted, making the impact of case-mix modelling transparent Perform internal validation to mitigate the risk of overfitting
Reporting	Clearly report which case-mix factors were considered, which were included in the final model, and provide justification for inclusion/exclusion Clearly report model performance and validation results, including any limitations or potential biases

drive meaningful improvements. Future benchmarking frameworks must integrate validated case-mix models, incorporate standardized outcome sets, especially focusing on long-term outcomes and PROs, and refine adjustment strategies to ensure fair and meaningful comparisons. A consensus on case-mix model assessment and dedicated reporting guidelines, akin to the TRIPOD statement, are urgently needed to ensure transparency and comparability. Addressing these challenges will enable robust international

comparisons, promote equity in cancer care, and ensure benchmarking leads to tangible improvements in patient outcomes across Europe. Only through rigorous methodologies, standardized reporting, and international collaboration can hospital benchmarking become a transformative tool for cancer care improvement.

Supplementary Information

Supplementary material is available at: https://www.esmorwd.org/article/S2949-8201(25)00065-7/fulltext

Acknowledgements

We thank Lovisa Liljegren, librarian at the Karolinska Institutet Library, for her invaluable assistance in constructing search strings across various databases, cleaning duplicate results, and providing comprehensive documentation of the search process.

REFERENCES

1. Ettorchi-Tardy A, Levif M, Michel P. Benchmarking: a method for continuous quality improvement in health. Healthc Policy. 2012;7(4):e101–19.
2. Agency for Healthcare Research and Quality. Types of healthcare quality measures. Available from: https://www.ahrq.gov/talkingquality/measures/types.html. Accessed 2025 Feb 16.
3. Iezzoni LI. Risk adjustment for performance measurement. In: Mossialos E, Papanicolas I, Smith PC, Leatherman S, editors. Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects. Cambridge: Cambridge University Press; 2010. p. 251–85.
4. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73.
5. Wells G, Shea B, O'Connell D, et al. Newcastle-Ottawa quality assessment scale cohort studies. Available from: https://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf. Published 2014. Accessed 2025 Feb 16.
6. Barker TH, Hasanoff S, Aromataris E, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for cohort studies. JBI Evid Synth. 2025;23(3):441–53.
7. JBI Institute. JBI manual for evidence synthesis – scoping reviews. Available from: https://jbi-global-wiki.refined.site/space/MANUAL/355863071/10.3.7.3+Data+extraction. Published 2024. Accessed 2025 Feb 16.
8. Algera MD, Baldewpersad Tewarie NMS, van Driel WJ, et al. Case-mix adjustment to compare hospital performances regarding complications after cytoreductive surgery for ovarian cancer: a nationwide population-based study. Int J Gynecol Cancer. 2023;33(3):534–42.
9. Algera MD, Slangen BFM, van Driel WJ, et al. Textbook outcome as a composite outcome measure to compare hospital performances regarding cytoreductive surgery for ovarian cancer: a nationwide population-based study. Gynecol Oncol. 2023;174:89–97.
10. Aravani A, Samy EF, Thomas JD, et al. A retrospective observational study of length of stay in hospital after colorectal cancer surgery in England (1998–2010). Medicine (Baltimore). 2016;95(36):e5064.
11. Beck N, Hoeijmakers F, van der Willik EM, et al. National comparison of hospital performances in lung cancer surgery: the role of case mix adjustment. Ann Thorac Surg. 2018;106(2):412–20.
12. Bernard A, Cottenet J, Pagès PB, et al. Mortality and failure-to-rescue major complication trends after lung cancer surgery between 2005 and 2020: a nationwide population-based study. BMJ Open. 2023;13(9):e075463.
13. Bernard A, Rivera C, Pages PB, et al. Risk model of in-hospital mortality after pulmonary resection for cancer: a national database of the French Society of Thoracic and Cardiovascular Surgery (Epithor). J Thorac Cardiovasc Surg. 2011;141(2):449–58.
14. Blake HA, Sharples LD, Boyle JM, et al. Improving risk models for patients having emergency bowel cancer surgery using linked electronic health records: a national cohort study. Int J Surg. 2024;110:1564–76.
15. Boyle JM, van der Meulen J, Kuryba A, et al. Measuring variation in the quality of systemic anti-cancer therapy delivery across hospitals: a national population-based evaluation. Eur J Cancer. 2023;178:191–204.
16. Burnell M, Iyer R, Gentry-Maharaj A, et al. Benchmarking of surgical complications in gynaecological oncology: prospective multicentre study. *BJOG.* 2016;123(12):2171–80.
17. Busweiler LAD, Schouwenburg MG, van Berge Henegouwen MI, et al. Textbook outcome as a composite measure in oesophagogastric cancer surgery. *Br J Surg.* 2017;104(6):742–50.

18. D'Journo XB, Berbis J, Jougon J, et al. External validation of a risk score in the prediction of the mortality after esophagectomy for cancer. *Dis Esophagus.* 2017;30(3):1–8.
19. Damhuis R, Coonar A, Plaisier P, et al. A case-mix model for monitoring of postoperative mortality after surgery for lung cancer. *Lung Cancer.* 2006;51(2):123–9.
20. Damhuis RA, Maat AP, Plaisier PW. Performance indicators for lung cancer surgery in the Netherlands. *Eur J Cardiothorac Surg.* 2015;47(5):897–903.
21. Das N, Talaat AS, Naik R, et al. Risk adjusted surgical audit in gynaecological oncology: P-POSSUM does not predict outcome. *Eur J Surg Oncol.* 2006;32(10):1135–8.
22. de Graaff MR, Elfrink AKE, Buis CI, et al. Defining Textbook Outcome in liver surgery and assessment of hospital variation: a nationwide population-based study. *Eur J Surg Oncol.* 2022;48(10):2414–23.
23. de Graaff MR, Klaase JM, van Dam RM, et al. Survival of patients with colorectal liver metastases treated with and without preoperative chemotherapy: nationwide propensity score-matched study. *Eur J Surg Oncol.* 2023;49:106932.
24. Elfrink AKE, Kok NFM, Swijnenburg RJ, et al. Nationwide oncological networks for resection of colorectal liver metastases in the Netherlands: differences and postoperative outcomes. *Eur J Surg Oncol.* 2022;48(3):435–48.
25. Elfrink AKE, Olthof PB, Swijnenburg RJ, et al. Factors associated with failure to rescue after liver resection and impact on hospital variation: a nationwide population-based study. *HPB (Oxford).* 2021;23(12):1837–48.
26. Elfrink AKE, van Zwet EW, Swijnenburg RJ, et al. Case-mix adjustment to compare nationwide hospital performances after resection of colorectal liver metastases. *Eur J Surg Oncol.* 2021;47(3):649–59.
27. Fischer C, Lingsma H, Hardwick R, et al. Risk adjustment models for short-term outcomes after surgical resection for oesophagogastric cancer. *Br J Surg.* 2016;103(1):105–16.
28. Fischer C, Lingsma HF, van Leersum N, et al. Comparing colon cancer outcomes: the impact of low hospital case volume and case-mix adjustment. *Eur J Surg Oncol.* 2015;41(8):1045–53.
29. Gildea C, Nordin A, Hirschowitz L, et al. Thirty-day postoperative mortality for endometrial carcinoma in England: a population-based study. *BJOG.* 2016;123(11):1853–61.
30. Gilthorpe MS, Harrison WJ, Downing A, et al. Multilevel latent class casemix modelling: a novel approach to accommodate patient casemix. *BMC Health Serv Res.* 2011;11:53.
31. Govaert JA, van Dijk WA, Fiocco M, et al. Nationwide outcomes measurement in colorectal cancer surgery: improving quality and reducing costs. *J Am Coll Surg.* 2016;222(1):19–29.e12.
32. Gray E, Figueroa JD, Oikonomidou O, et al. Variation in chemotherapy prescribing rates and mortality in early breast cancer over two decades: a national data linkage study. *ESMO Open.* 2021;6:100331.
33. Greijdanus NG, van Erning FN, van Workum F, et al. Variation in hospital performances after colorectal cancer surgery: a case-mix adjusted Dutch population-based study. *Eur J Surg Oncol.* 2024;50:107296.
34. Henneman D, Snijders HS, Fiocco M, et al. Hospital variation in failure to rescue after colorectal cancer surgery: results of the Dutch Surgical Colorectal Audit. *Ann Surg Oncol.* 2013;20(7):2117–23.
35. Henneman D, van Bommel AC, Snijders A, et al. Ranking and rankability of hospital postoperative mortality rates in colorectal cancer surgery. *Ann Surg.* 2014;259(5):844–9.
36. Jack RH, Gulliford MC, Ferguson J, et al. Geographical inequalities in lung cancer management and survival in South East England: evidence of variation in access to oncology services? *Br J Cancer.* 2003;88(7):1025–31.

37. Kolfschoten NE, Kievit J, Gooiker GA, et al. Focusing on desired outcomes of care after colon cancer resections; hospital variations in 'textbook outcome'. *Eur J Surg Oncol.* 2013;39(2):156–63.
38. Kolfschoten NE, Marang van de Mheen PJ, Gooiker GA, et al. Variation in case-mix between hospitals treating colorectal cancer patients in the Netherlands. *Eur J Surg Oncol.* 2011;37(11):956–63.
39. Kowalski C, Sibert NT, Breidenbach C, et al. Outcome quality after colorectal cancer resection in Certified Colorectal Cancer Centers: patient-reported and short-term clinical outcomes. *Dtsch Arztebl Int.* 2022;119(48):821–8.
40. Matthes-Martin S, Potschger U, Bergmann K, et al. Risk-adjusted outcome measurement in pediatric allogeneic stem cell transplantation. *Biol Blood Marrow Transplant.* 2008;14(3):335–43.
41. McArdle CS, Hole DJ. Outcome following surgery for colorectal cancer: analysis by hospital after adjustment for case-mix and deprivation. *Br J Cancer.* 2002;86(3):331–5.
42. Morris EJA, Taylor EF, Thomas JD, et al. Thirty-day postoperative mortality after colorectal cancer surgery in England. *Gut.* 2011;60(6):806–13.
43. Myrdal G, Lamberg K, Lambe M, et al. Regional differences in treatment and outcome in non-small cell lung cancer: a population-based study (Sweden). *Lung Cancer.* 2009;63(1):16–22.
44. Nickelsen TN, Jørgensen T, Kronborg O. Thirty-day mortality after surgery for colorectal cancer in Denmark. *Colorectal Dis.* 2005;7(5):500–6.
45. Nilssen Y, Brustugun OT, Fjellbirkeland L, et al. Small cell lung cancer in Norway: patterns of care by health region and survival trends. *Clin Lung Cancer.* 2024;25(4):e221–8.
46. Salet N, Stangenberger VA, Bremmer RH, et al. Between-hospital and between-physician variation in outcomes and costs in high- and low-complex surgery: a nationwide multilevel analysis. *Value Health.* 2023;26(4):536–46.
47. Sibert NT, Pfaff H, Breidenbach C, et al. Variation across operating sites in urinary and sexual outcomes after radical prostatectomy in localized and locally advanced prostate cancer. *World J Urol.* 2022;40(6):1437–46.
48. Skyrud KD, Bray F, Eriksen MT, et al. Regional variations in cancer survival: impact of tumour stage, socioeconomic status, comorbidity and type of treatment in Norway. *Int J Cancer.* 2016;138(9):2190–200.
49. Snijders HS, Henneman D, van Leersum NL, et al. Anastomotic leakage as an outcome measure for quality of colorectal cancer surgery. *BMJ Qual Saf.* 2013;22(9):759–67.
50. Talsma AK, Lingsma HF, Steyerberg EW, et al. The 30-day versus in-hospital and 90-day mortality after esophagectomy as indicators for quality of care. *Ann Surg.* 2014;260(2):267–73.
51. Talsma AK, Reedijk AM, Damhuis RA, et al. Re-resection rates after breast-conserving surgery as a performance indicator: introduction of a case-mix model to allow comparison between Dutch hospitals. *Eur J Surg Oncol.* 2011;37(4):357–63.
52. Ten Berge MG, Beck N, Steup WH, et al. Textbook outcome as a composite outcome measure in non-small-cell lung cancer surgery. *Eur J Cardiothorac Surg.* 2021;59(1):92–9.
53. Thurell J, Manouchehri N, Fredriksson I, et al. Risk-adjusted benchmarking of long-term overall survival in patients with HER2-positive early-stage breast cancer: a Swedish retrospective cohort study. *Breast (Edinb).* 2023;70:18–24.
54. Tighe D, Fabris F, Freitas A, et al. Machine learning methods applied to audit of surgical margins after curative surgery for head and neck cancer. *Br J Oral Maxillofac Surg.* 2021;59(2):209–16.
55. Tighe D, Kwok A, Putcha V, et al. Identification of appropriate outcome indices in head and neck cancer and factors influencing them. *Int J Oral Maxillofac Surg.* 2014;43(9):1047–53.

56. Tighe D, Lewis-Morris T, Freitas A, et al. Machine learning methods applied to audit of surgical outcomes after treatment for cancer of the head and neck. *Br J Oral Maxillofac Surg.* 2019;57(8):771–7.
57. Tighe D, McMahon J, Schilling C, et al. Machine learning methods applied to risk adjustment of cumulative sum chart methodology to audit free flap outcomes after head and neck surgery. *Br J Oral Maxillofac Surg.* 2022;60(12):1353–61.
58. Tighe D, Sassoon I, Hills A, et al. Case-mix adjustment in audit of length of hospital stay in patients operated on for cancer of the head and neck. *Br J Oral Maxillofac Surg.* 2019;57(9):866–72.
59. Tighe D, Sassoon I, Kwok A, et al. Is benchmarking possible in audit of early outcomes after operations for head and neck cancer? *Br J Oral Maxillofac Surg.* 2014;52(10):913–21.
60. Tighe D, Sassoon I, McGurk M, et al. Validating a benchmarking tool for audit of early outcomes after operations for head and neck cancer. *Ann R Coll Surg Engl.* 2017;99(4):299–306.
61. Tighe D, Tekeli K, Gouk T, et al. Machine learning methods applied to audit of surgical margins after curative surgery for facial (non-melanoma) skin cancer. *Br J Oral Maxillofac Surg.* 2023;61(1):94–100.
62. Tighe D, Thomas AJ, Hills A, et al. Validating a risk stratification tool for audit of early outcome after operations for squamous cell carcinoma of the head and neck. *Br J Oral Maxillofac Surg.* 2019;57(9):873–9.
63. Tighe DF, Thomas AJ, Sassoon I, et al. Developing a risk stratification tool for audit of outcome after surgery for head and neck squamous cell carcinoma. *Head Neck.* 2017;39(7):1357–63.
64. van den Bosch T, Warps AK, de Nerée Tot Babberich MPM, et al. Predictors of 30-day mortality among Dutch patients undergoing colorectal cancer surgery, 2011–2016. *JAMA Netw Open.* 2021;4(3):e217737.
65. van der Heiden-van der Loo M, de Munck L, Visser O, et al. Variation between hospitals in surgical margins after first breast-conserving surgery in the Netherlands. *Breast Cancer Res Treat.* 2012;131(3):691–8.
66. Voeten DM, van der Werf LR, van Sandick JW, et al. Length of hospital stay after uncomplicated esophagectomy: hospital variation shows room for nationwide improvement. *Surg Endosc.* 2021;35(11):6344–57.
67. Voeten DM, van der Werf LR, Wijnhoven BPL, et al. Failure to cure in patients undergoing surgery for esophageal carcinoma: hospital of surgery influences prospects for cure: a nationwide cohort study. *Ann Surg.* 2020;272(5):744–50.
68. Vos EL, Koppert LB, Jager A, et al. From multiple quality indicators of breast cancer care toward hospital variation of a summary measure. *Value Health.* 2020;23(9):1200–9.
69. Vos EL, Lingsma HF, Jager A, et al. Effect of case-mix and random variation on breast cancer care quality indicators and their rankability. *Value Health.* 2020;23(9):1191–9.
70. Wahba AJ, Phillips N, Mathew RK, et al. Benchmarking short-term postoperative mortality across neurosurgery units: is hospital administrative data good enough for risk-adjustment? *Acta Neurochir (Wien).* 2023;165(9):1695–706.
71. Walker K, Finan PJ, van der Meulen JH, et al. Model for risk adjustment of postoperative mortality in patients with colorectal cancer. *Br J Surg.* 2015;102(3):269–80.
72. Warps AK, Detering R, Tollenaar R, et al. Textbook outcome after rectal cancer surgery as a composite measure for quality of care: a population-based study. *Eur J Surg Oncol.* 2021;47(11):2821–9.
73. Sibert NT, Pfaff H, Breidenbach C, et al. Different approaches for case-mix adjustment of patient-reported outcomes to compare health care providers: methodological results of a systematic review. *Cancers (Basel).* 2021;13(16):3964.

74. Dal Cero M, Roman M, Grande L, et al. Textbook outcome and survival after gastric cancer resection with curative intent: a population-based analysis. *Eur J Surg Oncol.* 2022;48(4):768–75.
75. Xu SJ, Lin LQ, Chen C, et al. Textbook outcome after minimally invasive esophagectomy is an important prognostic indicator for predicting long-term oncological outcomes with locally advanced esophageal squamous cell carcinoma. *Ann Transl Med.* 2022;10(2):161.
76. ICHOM. Patient-centered outcome measures. Available from: https://www.ichom.org/patient-centered-outcome-measures/. Accessed 2025 Feb 16.
77. Verheul EM, van Klaveren D, Lingsma HF, et al. High-impact complications after breast cancer surgery in the Dutch national quality registry: evaluating case-mix adjustment for hospital comparisons. *BJS Open.* 2024;9:zrae147.
78. Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* 2019;40(2):44–58.
79. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12–22.
80. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br Med J.* 2015;350:g7594.



Chapter 3

High-impact Complications After Breast Cancer Surgery in the Dutch National Quality Registry: Evaluating Case-mix Adjustment for Hospital Comparisons

Elfi M. Verheul, David van Klaveren, Hester F. Lingsma, Elvira Vos, Marije J. Hoornweg, Sabine Siesling, Linetta B. Koppert; NBCA consortium.

Published in: British Journal of Surgery (BJS) open, 2024.

ABSTRACT

Background: Comparison of quality indicators can improve quality of care. However, case-mix adjustment is deemed essential. The aim of this study was to develop and validate case-mix adjustment models and to evaluate the effect of case-mix adjustment for the quality indicators related to complications after breast cancer surgery.

Methods: Multivariable logistic regression with backward selection ($P < 0.1$) was used to develop case-mix models in patients undergoing breast cancer surgery (all types, breast-conserving surgery, mastectomy with or without immediate reconstruction) in the Netherlands (NABON Breast Cancer Audit). High-impact complications were defined as Clavien Dindo grade ≥3. Model performance was assessed using the area under the receiver operating characteristic curve (AUC), corrected for optimism with bootstrap validation. Observed-to-expected plots were used to visualize the difference between unadjusted and case-mix adjusted hospital performance (hospital shifts).

Results: In total 32 084 patients from 72 hospitals treated in 2021-2022 were included. A between-hospital variation in complication rates was observed for all surgeries (interquartile range 2.4-6.0%), breast-conserving surgery (interquartile range 1.4-3.4%), and mastectomy with (interquartile range 9.4-9.1%) and without reconstruction (interquartile range 3.3-9.7%). Of the considered variables, body mass index, smoking, multifocality and neoadjuvant therapy were weakly associated with complications. However, surgery type was strongly related to complications (AUC 0.70), resulting in noticeable hospital shifts in the quality indicator scores comprising all surgeries. After stratification for surgery type, no evident hospital shifts were observed after case-mix correction.

Conclusion: For valid comparison of complication rates after breast cancer surgery between hospitals, stratification by surgery type is crucial. Subsequently, the evaluated patient and tumour characteristics have a negligible effect on the hospital variation.

INTRODUCTION

In Europe, nearly half of the countries have implemented national programmes aimed at improving breast cancer (BC) care[1]. A key component of these initiatives involves gathering hospital data and presenting them periodically through quality indicators (QIs). Evidence shows that registries, which benchmark hospitals with these QIs, achieve improved quality of care[2–4]. Hospitals have the opportunity to assess their own performance in comparison to others, enabling those with poorer performance to learn from best practices and improve. When reporting these QIs, it is important that QIs measure actual differences in quality of care, especially when QIs become publicly accessible. This necessitates the use of valid and reliable indicators[5].

The Dutch breast cancer quality registry (also called NABON Breast Cancer Audit, NBCA) includes all surgically treated patients with BC in the Netherlands[6]. Since 2011, data have been gathered and QIs are developed on structures, processes, and outcomes. The Dutch Ministry of Health, Welfare and Sport (VWS) invested from 2018 to 2022 in an outcome-based care programme[7]. This programme identified 'complications after surgery' as an important outcome indicator[8]. This indicator is currently measured in the Dutch breast cancer quality registry for internal use only.

Transparency regarding high-impact complications after BC surgery is important for improving quality of care and informing patients. Patient and tumour characteristics, known as case-mix factors, can influence the validity of QIs[5,9–12]. Significant case-mix effects indicate that QIs reflect differences in patient populations rather than in care quality. Vos *et al*. conducted a study on case-mix adjustment of six BC QIs, highlighting the importance of assessing case-mix impact for each indicator independently[5]. The specific patient and tumour characteristics of importance depended on the specific structure, process or outcome being measured. Outcome indicators showed evident case-mix impact. Therefore, the aim of this study was to assess whether case-mix played a role for the outcome indicator 'complications after surgery'.

The aim was to develop and validate case-mix adjustment models and to evaluate the effect of case-mix adjustment on the QI related to complications after BC surgery.

METHODS

Study population

All patients registered in the Dutch breast cancer quality registry in 2021 and 2022 were included as this was the first year the QI 'complications after surgery' was adequately measured. The inclusion time is the moment of biopsy. The database covers all 72 hospi-

tals that provide BC care in the Netherlands and includes all patients with primary BC and ductal carcinoma *in situ* (DCIS) who underwent surgery. Lobular carcinoma *in situ* (LCIS), recurrent BC, phyllodes tumours and non-surgically treated patients were excluded.

Potential case-mix factors

Potential case-mix variables were retrieved from literature and expert opinion (Dutch breast cancer quality board and scientific committee). The potential case-mix variables included patient characteristics (age, body mass index (BMI), sex, smoking), tumour characteristics (T stage, N stage, histologic subtype, differentiation grade, receptor status, multifocality based on pathology, screen detection), and treatment characteristics (type of surgery and systemic therapy).

Quality indicator

The performance indicator 'complications after surgery' is defined as the proportion of patients undergoing primary surgery for DCIS or invasive BC that have a complication during admission or within 30 days after surgery. In this study complications with Clavien–Dindo grades 3, 4, or 5 were included, which comprise high-impact complications involving the need for surgical, endoscopic, or radiological intervention, life-threatening complications requiring admission to a monitored care unit, or mortality[13,14]. The indicator is measured for three groups: all surgeries, breast-conserving surgery (BCS), and mastectomy.

Statistical analysis and model development

Descriptive statistics are presented with frequencies (percentages) or means(s.d.). Complication rates are calculated over 2021 and 2022, and the variation between hospitals is visualized in boxplots. Hospital variation in potential case-mix variables is assessed by medians and interquartile ranges. The i.q.r. represents the range between the first (25th) and third (75th) quartiles of a variable, indicating the spread of the middle 50% of the data.

Multiple imputation with chained equations (MICE) was used to impute missing case-mix values (20 imputations)[15,16]. Besides the case-mix variables presented, the variables DCIS or invasive status, radiotherapy, adjuvant chemotherapy, endocrine therapy, complications after chemotherapy, and complications after surgery were included to obtain more accurate imputations. In practice, imputing outcomes is sensitive, and therefore, all patients with missing values for the variable 'complications after surgery' were removed from the data set after imputation.

For model development, patient data from 2021 and 2022 was used. Multivariable logistic regression was performed for the QI across all different subgroups. Significant case-mix variables were identified with backward selection ($P < 0.1$). For the continuous variables

age and BMI, restricted cubic splines were used to assess if there was a non-linear relationship with the outcome. The impact of individual case-mix variables was evaluated using Wald statistics minus the degrees of freedom (χ^2 – df). This enables comparison of the proportional contribution of each predictor to the outcome and to other predictors. The performance of the case-mix models was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC). A value of 0.5 indicates a non-informative model, meaning that there is no influence of case-mix variables, whereas a value of 1 indicates a perfect model; all variation can be explained by case mix. Bootstrap validation with 200 resamples was used to correct the AUC for optimism.

The influence of case-mix adjustment is assessed by plotting unadjusted and case-mix adjusted observed/expected (O/E) ratios. The O/E ratio was calculated by dividing each hospital's observed complication rate by its expected complication rate. The expected complication rate is determined using the overall average (unadjusted) or predicted probabilities from the case-mix adjustment model (adjusted). An O/E ratio above one indicates that a hospital performed worse than expected and a ratio below one means it performed better than expected.

All analyses were conducted using R, Version 4.2.3.

RESULTS

Patient and tumour characteristics

The study population comprised 32 084 patients, with a mean(s.d.) age of 65(13) years, who either had invasive BC (87%) or DCIS (13%, Table 1). Of these patients, 71% underwent BCS (*N* = 22 661), 20.4% mastectomy without immediate reconstruction and 8.6% mastectomy with immediate reconstruction (*N* = 6 654 and *N* = 2 769; total mastectomy *N* = 9 423). Patients who underwent a mastectomy had a higher T stage (17% stage T3 or T4 *versus* 2.3%), had a multifocal tumour more often (34% *versus* 9.5%), were less likely diagnosed by screening (23% *versus* 48%) and received neoadjuvant chemotherapy more often (28% *versus* 19%). In 2021 patient numbers were higher as compared to 2022, due to lower incident numbers in 2020 as a result of the COVID pandemic (2021 *N* = 16 684, 2022 *N* = 15 400). Within the entire cohort, three variables had notable rates of missing data, namely smoking (25.1%), BMI (13.4%), and multifocality (5.5%).

Table 1: Baseline characteristics of patients diagnosed with BC between 2021 and 2022 in the Netherlands, divided per group that can be measured in the quality indicator 'complications after surgery'.

	Total	Breast-conserving	Mastectomy		
Characteristic	N = 32 084[1]	N = 22 661[1]	**All** N = 9 423[1]	**Without immediate reconstruction** N = 6 654[1]	**With immediate reconstruction** N = 2 769[1]
Age	65 (13)	65 (12)	64 (15)	69 (14)	53 (11)
(missing)	11	6	5	5	0
BMI	26.5 (5.1)	26.7 (5.2)	26 (5)	26.6 (5.2)	24.9 (4.3)
(missing)	3 164	2 146	1 018	797	221
Female	31 858 (99.3%)	22 647 (99.9%)	9 211 (97.8%)	6 442 (96.7%)	2 769 (100%)
Smoking					
Never or previous	20 149 (84%)	14 246 (84%)	5 903 (84%)	4 052 (83%)	1 851 (85%)
Current	3 871 (16%)	2 736 (16%)	1 135 (16%)	812 (17%)	323 (15%)
(missing)	8 064	5 679	2 385	1 790	595
Screen detected	12 633 (40%)	10 553 (48%)	2 080 (23%)	1 401 (22%)	679 (25%)
(missing)	864	604	260	197	63
T stage					
cT1	16 207 (51%)	13 263 (59%)	2 944 (32%)	1 975 (30%)	969 (35%)
cT2	8 958 (28%)	5 580 (25%)	3 378 (36%)	2 546 (39%)	832 (30%)
cT3/cT4	2 135 (6.7%)	526 (2.3%)	1 609 (17%)	1 293 (20%)	316 (12%)
cTis	4 375 (14%)	3 017 (13%)	1 358 (15%)	743 (11%)	615 (23%)
(missing)	409	275	134	97	37
N stage					
cN0	27 547 (86%)	20 306 (90%)	7 241 (77%)	4 899 (74%)	2 342 (85%)
cN+	4 340 (14%)	2 227 (9.9%)	2 113 (23%)	1 701 (26%)	412 (15%)
(missing)	197	128	69	54	15
Histologic subtype					
No special type	24 667 (79%)	17 942 (81%)	6 725 (73%)	4 675 (72%)	2 050 (76%)
Lobular	3 700 (12%)	2 184 (9.9%)	1 516 (16%)	1 154 (18%)	362 (13%)
Other	2 904 (9.3%)	1 952 (8.8%)	952 (10%)	665 (10%)	287 (11%)
(missing)	813	583	230	160	70
Differentiation					
Grade 1	6 870 (22%)	5 604 (26%)	1 266 (14%)	877 (14%)	389 (15%)
Grade 2	15 860 (52%)	11 100 (51%)	4 760 (53%)	3 424 (54%)	1 336 (52%)
Grade 3	7 893 (26%)	4 980 (23%)	2 913 (33%)	2 055 (32%)	858 (33%)
(missing)	1 461	977	484	298	186

Table 1: Baseline characteristics of patients diagnosed with BC between 2021 and 2022 in the Netherlands, divided per group that can be measured in the quality indicator 'complications after surgery'. (*continued***)**

	Total	Breast-conserving	Mastectomy		
Receptor Status[2]					
HR+, HER2-	21 370 (82.3%)	15 459 (84.4%)	5 911 (78%)	4 322 (77%)	1 589 (79.9%)
HER2+	2 151 (8.3%)	1 375 (7.5%)	776 (10%)	567 (10%)	209 (10.5%)
Triple-	2 434 (9.4%)	1 488 (8.1%)	946 (12%)	755 (13%)	191 (9.6%)
Multifocality					
Unifocal	25 300 (83%)	19 408 (90%)	5 892 (66%)	4 290 (68%)	1 602 (62%)
Multifocal	5 017 (17%)	2 042 (9.5%)	2 975 (34%)	2 009 (32%)	966 (38%)
(missing)	1 767	1 211	556	355	201
Neoadjuvant chemotherapy	6 841 (21%)	4 217 (19%)	2 624 (28%)	1 745 (26%)	879 (32%)
(missing)	15	5	10	6	4
Year					
2021	16 684 (52%)	11 568 (51%)	5 116 (54%)	3 538 (53%)	1 578 (57%)
2022	15 400 (48%)	11 093 (49%)	4 307 (46%)	3 116 (47%)	1 191 (43%)

Na means not applicable. [1]Mean (SD); n (%), [2]Patients with Tis disease are excluded, as receptor status is not determined in this group.

Complication rates

The complication rate after surgery in the overall study population was 4.4% (Table S1). Higher complication rates were observed in patients receiving a mastectomy (± immediate reconstruction, 8.9%) than in those undergoing BCS (2.6%). Patients undergoing immediate reconstruction after mastectomy had higher complication rates (13.5%) than those who underwent mastectomy without immediate reconstruction (7%). The majority of patients experienced grade 3 complications (N = 1284; grade 4: N = 13; grade 5: N = 5; Table S2).

Between hospital variation

The interquartile range for occurrence of surgical complications was 2.4–6.0% after all surgeries, 1.4–3.4% after BCS and 5.3–13% after mastectomy (± immediate reconstruction; Fig. S1). In terms of case mix, most variation was seen in the type of surgery (BCS versus mastectomy; Fig. 1). More specifically, within the mastectomy group, there was variability in the uptake of immediate reconstruction (i.q.r. 18–36%). Additionally, considerable variability was observed in the occurrence of screen-detected cancers (i.q.r. 35–43%), differentiation grade (i/q/r/ grade 1: 20–27%; grade 2: 47–55%, grade 3: 22–30%) and smoking status (i.q.r. 14–21%). Minimal variation was seen in sex (i.q.r. 99–100), as few men are diagnosed with BC. Therefore, sex was excluded from model development.

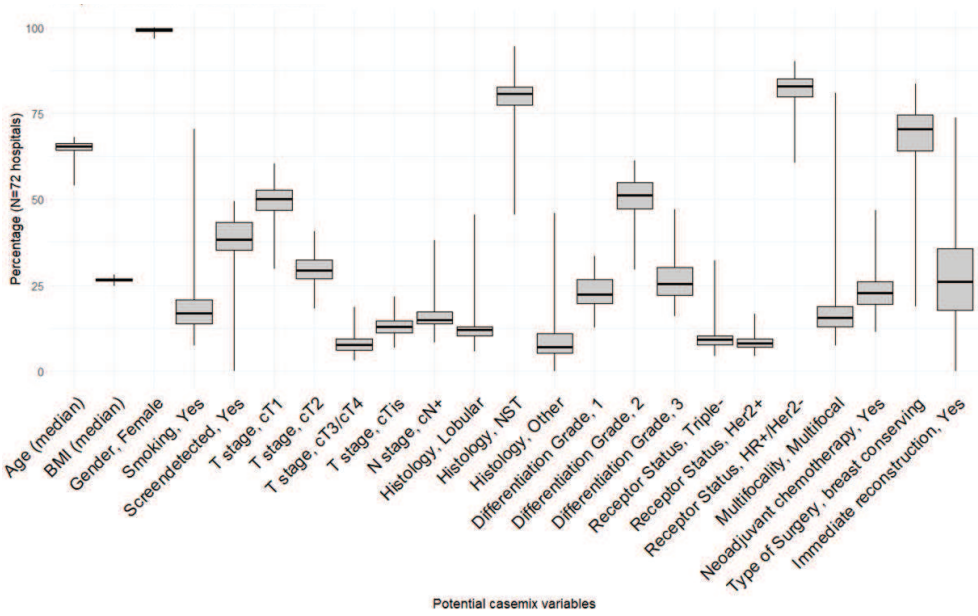


Figure 1: Hospital variation (N=72) in case-mix variables in Dutch patients who underwent breast cancer surgery in 2021 and 2022 (n=32 084). *The plots show the distribution of median age and BMI, along with the mean percentages of other potential case-mix variables. It is presented in a plot showing the 25th, 50th and 75th percentiles, along with the minimal and maximal values (range). For the variable "immediate reconstruction" the plot is presented as percentage of all mastectomy patients instead of all surgeries.*

Model development, performance, and validation

Age and BMI both showed a linear association with complications, indicating that incorporating splines does not lead to significant improvement in model fit.

Smoking, multifocality and neoadjuvant chemotherapy were associated with the occurrence of surgical complications in most analyses (Table 2), followed by BMI, T and N stage. However, the strength of their association with complications was negligible compared to the strength of association between type of surgery and complications and immediate reconstruction (Fig. S2).

The discriminative ability of the model for all surgeries was good (AUC 0.70), and moderate for mastectomy (± reconstruction, AUC 0.63) and BCS (AUC 0.61). At the hospital level, evident shifts in the observed/expected plots were seen in the QI for all surgeries and the QI scores for mastectomy (± reconstruction) (Fig. 2). In contrast, for patients receiving BCS there was almost no effect of case-mix adjustment. For all developed case-mix models there was a small amount of overfitting (optimism for all models <0.036).

Further stratification of mastectomy patients was applied based on whether immediate reconstruction was performed, because explanatory analyses showed that immediate reconstruction did significantly affect between-hospital comparisons. After stratification for immediate reconstruction, the AUC was poor (without reconstruction: 0.56, with reconstruction: 0.53), and consequently the influence of case-mix adjustment on hospital shifts was minimal (Fig. 2).

Table 2: Case-mix adjustment models for complications after surgery, breast-conserving surgery, mastectomy, mastectomy without reconstruction and mastectomy with reconstruction. *Odds ratios with confidence intervals (OR 95% CI) are presented for variables significantly associated with complications after backward selection (p<0.1).*

	Total		Breast-conserving		Mastectomy					
					All		Without immediate reconstruction		With immediate reconstruction	
	N = 32,084[1]		N = 22,661[1]		N = 9,423[1]		N = 6,654[1]		N = 2,769[1]	
	OR (95% CI)	**p**	**OR 95% CI**	**p**	**OR 95% CI**	**p**	**OR 95% CI**	**p**	**OR 95% CI**	**p**
Age (ref: 60:70[1])	X		0.87 (0.82-0.92)	<.0001	1.15 (1.04-1.27)	0.0051	Y		Y	
BMI (ref: 25:30[2])	1.22 (1.15-1.29)	<.0001	1.18 (1.10-1.25)	<.0001	1.12 (1.01-1.25)	0.0371				
Smoking (Ref: no or quitted)	1.27 (1.12- 1.58)	0.0010	1.25 (1.04- 1.53)	0.0205	1.37 (1.01-1.88)	0.0465	1.35 (0.97-1.87)	0.0748	Y	
Type of surgery (ref: breast- conserving surgery)	2.5(2.18-2.57)	<.0001	Na		Na		Na		Na	
Immediate reconstruction (ref: no)	2.53 (2.15-2.95)	<.0001	Na		5.11 (3.82-6.87)	<.0001	Na		Na	
Screen detected (ref: no)	X		0.75 (0.64-0.87)	0.0002	X		X		0.74 (0.42-0.97)	0.0365
T stage				<.0001		0.0004		0.0003	X	
cT1			1		1		1			
cT2			1.23 (1.03- 1.47)		1.62 (1.26-2.092)		1.86 (1.37-2.53)			
cT3/4			1.91 (1.46- 2.49)		1.88 (1.280- 2.77)		2.18 (1.37- 3.45)			
cTis			1.26 (1.02- 1.55)		0.99 (0.70 – 1.41)		1.22 (0.80 – 1.89)			
N stage, positive (ref: negative)	1.30 (1.09-1.55)	0.0033	X		1.38 (1 -1.90)	0.0474	1.42 (0.98-2.06)	0.0643	X	
Receptor status	X		X		X		X		X	

Table 2: Case-mix adjustment models for complications after surgery, breast-conserving surgery, mastectomy, mastectomy without reconstruction and mastectomy with reconstruction. *Odds ratios with confidence intervals (OR 95% CI) are presented for variables significantly associated with complications after backward selection (p<0.1).* **(***continued***)**

	Total		Breast-conserving		Mastectomy					
Differentiation grade 1 grade 2 grade 3	X		1 1.27 (1.05-1.53) 1.24 (0.99– 1.55)	0.0443	X		X		X	
Multifocality (ref: no)	1.30 (1.12- 1.50)	0.0004	1.83 (1.55-2.15)	<.0001	1.82 (1.41-2.34)	<.0001	2.26 (1.70 – 3.00)	<.0001	2.51 (1.68-3.76)	<.0001
Histologic subtype	1.18 (1.01 – 1.40)	0.0737	X		X		X		X	
Neoadjuvant chemotherapy (ref: no)	0.66 (0.56- 0.78)	<.0001	0.52 (0.42- 0.64)	<.0001	0.57 (0.41 – 0.80)	0.0010	0.53 (0.36-0.76)	0.0006		
C-statistic	0.70 (0.69-0.72)		0.61 (0.59 - 0.64) 0.6124		0.63 (0.60 – 0.65)		0.56 (0.53-0.59)		0.53 (0.50-0.56)	
Optimism	0.0043		0.0091		0.01495		0.0253		0.036	
C-statistic after bootstrap validation	0.70 (0.70-0.71)		0.60 (0.58-0.63)		0.61 (0.59-0.63)		0.54 (0.51-0.57)		0.50 (0.47-0.53)	

Na is not applicable. X means not included in the model due to p>0.1. [1]Odds ratio for age is presented as a decade (where one unit increase represents a decade increase). The reference value taken was between 60 and 70 years. [2]For BMI, one unit increase represents 5 points increase in BMI. The reference value taken was between a BMI of 25 and 30.

DISCUSSION

This study aimed to develop practical case-mix models for the QI 'complications after BC surgery'. This QI is applied to all types of surgery, BCS and mastectomy (± immediate reconstruction) in the Dutch breast cancer quality registry. The analyses showed that the type of surgery, and whether immediate reconstruction was performed, were the strongest and most important factors associated with the occurrence of complications. After case-mix adjustment noticeable hospital shifts in outcomes for all types of surgeries and for mastectomy (± reconstruction) were seen. In contrast, the occurrence of complications after BCS was only weakly associated with case-mix variables, meaning negligible influence of case-mix adjustment on QI scores. After further stratifying mastectomy patients based on whether they underwent immediate reconstruction, the former hospital shifts disappeared. This suggests that comparisons can be made without the need for case-mix adjustment by adequate stratification into BCS, mastectomy with immediate reconstruction, and mastectomy without immediate reconstruction.

The importance of type of surgery in relation to complication rates has been highlighted before[17,18], with multiple studies confirming comparable complication rates, indicating the lowest complications for BCS and the highest for mastectomy with immediate reconstruction[19–22].

Only one study has investigated case-mix adjustment regarding complications after BC surgery before. Berlin *et al*. examined the influence of case-mix adjustment specifically for the group undergoing immediate breast reconstruction across 11 American hospitals[23]. Although they reported an impact of case-mix adjustment, they did not specify the strength of the case-mix variables responsible for it, nor did they evaluate the adjustment's impact on estimated hospital performance. This makes it difficult to compare these results to the current study. Nevertheless, Dutch studies on anastomotic leakage after colon surgery and Clavien–Dindo grade >3 complications after cytoreductive surgeries also showed that adjustment for case-mix variables did not significantly affect hospital outcomes[24,25].

Case-mix adjustment for six QIs of the Dutch breast cancer quality registry was previously examined[5]. Vos *et al.* demonstrated differences in the impact of case-mix adjustment for each QI. Four of six QIs showed evident impact of case-mix adjustment, comprising two outcome indicators (irradical resection in BCS for invasive BC and breast contour preserving treatment) and two process indicators (MRI before neoadjuvant chemotherapy and radiotherapy for locally advanced BC). This underlines the importance for individual QI evaluation on validity and reliability, to conclude whether case-mix adjustment has added value. Notably, the selection of potential case-mix variables can differ for each QI.

For example, type of surgery was not included as a case-mix variable in the study of Vos *et al*. due to its lack of immediate relevance to the assessed QIs.

Although adjustment for patient and tumour variables had minimal impact on hospital level, statistically significant predictive variables for individual patients were identified, consistent with developed prediction models[26,27]. The finding that in BCS the likelihood of experiencing a complication decreases with age, whereas in mastectomy it increases with age, is not fully understood. This may be related to the selection of fitter patients for immediate reconstruction post-mastectomy, as immediate reconstruction is not included in the BCS model. Patients receiving neoadjuvant chemotherapy were less likely to have complications; however, the reasons for this observation also remain unclear. There could potentially be a link with the increased use of prophylactic antibiotics in this patient group. Furthermore, an association was found between smoking and the current QI. This effect was not found in the immediate reconstruction group. This is in contrast with a recent published meta-analysis (OR 1.46, 95% c.i. 1.08 to 1.97)[28]. The reason the current results are not in line with this meta-analysis might be attributed to the fact that plastic surgeons already take into account this risk factor by the choice of treatment (selection bias) – that is a reconstruction is discouraged in smokers. Another possibility is that the sample size was too small to show this association.

A proper case-mix model for outcome QIs after a certain treatment should only include variables that are not influenced by the choices of a hospital or healthcare provider in the preoperative decisional pathway. Therefore, it is debatable whether type of surgery can be considered as a case-mix factor and should be included in the adjustment model. In the literature, the potential for including the extent of surgery as a case-mix variable is suggested to add value[9,25], whereas the surgical approach (for example open *versus* laparoscopic procedure) may be more closely linked to the choice of a surgical team with accompanying technical capability and therefore preferably excluded[9,29]. In other terms, the approach of a procedure is more related to quality of care, for which adjustments are undesirable, whereas the extent (for example BCS or mastectomy) could be valuable for case-mix adjustment. Therefore, it was decided to include the type of surgery and, more detailed, immediate reconstruction yes or no in the study. Conversely, it was chosen to incorporate N-stage instead of the type of axillary surgery. As this is a grey area, the decision was further supported by the fact that complications related to axillary surgery are primarily seromas, infections, lymphoedema, and shoulder/arm morbidity, which are generally not classified as Clavien–Dindo grade ≥3[18].

The primary goal of this study is to stimulate quality of care improvements, but only using valid information to avoid misinformation. Based on the current findings, stratification by type of surgery is recommended rather than adjustment. Considerations that support stratification include the opportunity to shorten the time-consuming process of register-

ing variables, particularly if variables with negligible impact on QIs in a quality registry can be removed. Another important reason that supports stratification, especially for mastectomy patients, is the difference in target groups interpreting the QIs. In the Netherlands, mastectomies are typically performed by surgeons, whereas immediate reconstruction procedures are performed in collaboration with plastic surgeons[30]. Consequently, stratifying mastectomy patients with and without immediate reconstruction gives better insight for both professions into their respective results. This provides a better opportunity for improvement, aligning with the central aim of quality registries[6].

Participation in the Dutch breast cancer quality registry is obligatory for all 72 hospitals delivering BC care in the Netherlands. As a result, a considerable strength of this study is the usage of this nationwide registry including all BC patients treated with curative intent. Data quality of the registry is high, which is related to the use of the Netherlands Cancer Registry (NCR), in which data entry is performed by trained registrars from the Netherlands Comprehensive Cancer Organization (IKNL) in about 75% of the hospitals. Still, in this nationwide data set missing values were noted in the potential case-mix variables smoking (25.1%), BMI (13.4%), and multifocality (5.5%). 'Multiple imputation, then deletion' was used, which is the most valid way of reducing bias of missing values and increasing the precision of outcomes when benchmarking hospitals[15,31]. Another strength is the use of the Clavien–Dindo classification, which aligns with the International Consortium for Health Outcomes Measurement core outcome set for BC[32], and therefore provides insight into high-impact complications (Clavien–Dindo grade ≥3). Under-registration in high-impact complications is less common than for minor complications. However, this approach limits the ability to specifically examine the types of complications (for example bleeding, severe infection), which could be valuable for understanding specific case-mix variables, as not all complications share the same risk factors[27]. Additionally, an important limitation is that the current model selectively includes variables collected in the Dutch breast cancer quality registry. This decision was made carefully considering practical implications. Nevertheless, some important risk factors for complications, like co-morbidities[28,33], are not included as they are not measured in the registry.

Based on the current results it is argued that the QI 'complications after BC surgery' including all surgically treated patients is not a meaningful indicator, as differences are mostly related to the particular type of surgery instead of quality of care. The authors recommend that the QI is applied to the suggested stratification into type of surgery. As a result, it provides a more valid insight into the performance in time and/or between hospitals, making it more actionable for (plastic) surgeons and omitting the need for case-mix adjustment.

Nevertheless, the need for case-mix adjustment is only giving information on part of the validity of a QI. To obtain a complete overview of the value of the QIs 'complications after

BCS or mastectomy with or without immediate reconstruction', an additional study conducting a comprehensive evaluation, including aspects such as feasibility and rankability, is needed[34]. In general, future studies should focus on developing meaningful and valid outcome indicators determining all above-mentioned aspects. When necessary, these indicators should include case-mix adjustment.

Stratification by the type of surgery is crucial for a more valid comparison of complication rates between hospitals after breast cancer surgery. After stratification, the evaluated patient and tumour characteristics have a negligible effect on the hospital variation. The QI in the Dutch national breast cancer registry will be refined based on these findings. Consequently, there will be a greater potential for improvement by more accurately measuring the quality of BC care between hospitals.

Supplementary Information

Supplementary material is available at: https://academic.oup.com/bjsopen/article/9/1/zrae147/7998764

Acknowledgements

The authors thank all, registrar and healthcare professionals for data registration and development of the NBCA.

NBCA Consortium

The NBCA Consortium consists of the following participants, displayed with their affiliations: **Alwine A. Hellingman** (Department of Surgery, Tergooi Medical Centre), **A. M. (Yvette) Moorman** (Department of Surgery, Medical Centre Leeuwarden, Leeuwarden), **Anne Brecht Francken** (Department of Surgery, Isala Klinieken, Zwolle), **Bert van der Vegt** (Department of Pathology, University of Groningen, University Medical Centre Groningen, Groningen), **Carla Meeuwis** (Department of Radiology, Rijnstate Hospital, Arnhem), **Carlijn T. I. de Betue** (Department of Surgery, Jeroen Bosch Hospital, Den Bosch), **Carolien H. M. van Deurzen** (Department of Pathology, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Rotterdam), **Claudette E. Loo** (Department of Radiology, Netherlands Cancer Institute–Antoni Van Leeuwenhoek, Amsterdam), **Caroline M. E. Contant** (Department of Surgery, Maasstad Hospital, Rotterdam), **Cristina Guerrero Paez** (Dutch Breast Cancer Society (BVN), Utrecht), **D. M. (Irma) den Hoed** (Department of Surgery, Elizabeth Tweesteden Hospital, Tilburg), **Daniel Henneman** (Department of Surgery, Treant Healthgroup and Chirurgen Coöperatie Noord Nederland), **Dominique J. P. van Uden** (Department of Surgery, Canisius Wilhelmina Hospital, Nijmegen), **Tanja G. Frakking** (Department of Surgery, Rivas Zorggroep Beatrix Hospital, Gorinchem), **Els Van Dessel** (Department of Surgery, ZorgSaam Hospital, Zeeuws-Vlaanderen), **Enja J. Bantema-Joppe** (Department of Radiation Oncology, Radiotherapy Institute Friesland,

Friesland), **Ernst J. P. Schoenmaeckers** (Department of Surgery, Meander Medical Centre, Amersfoort), **Ester Siemerink** (Department of Medical Oncology, Ziekenhuisgroep Twente (ZGT), Hengelo), **Floris P. R. Verbeek** (Department of Surgery, Groene Hart Hospital, Gouda), **Gea A. Gooiker** (Department of Surgery, Northwest Clinics, Alkmaar, The Netherlands), **Henriette Schuttevaer** (Department of Surgery, Martini Hospital, Groningen), **Hinne A. Rakhorst** (Department of Plastic, Reconstructive, and Hand Surgery, Medisch Spectrum Twente and Ziekenhuisgroep Twente), **Ingrid Kappers** (Department of Surgery, Tjongerschans ziekenhuis, Heerenveen), **Ingrid van den Hoven** (Department of Surgery, Anna Hospital, Geldrop), **James Van Bastelaar** (Department of Surgery, Zuyderland Medical Centre, Heerlen), **Janneke Verloop** (Department of Research and Development, Netherlands Comprehensive Cancer Organisation (IKNL)), **José H. Volders** (Department of Surgery, Diakonessenhuis, Utrecht), **Joan B. Heijns** (Department of Medical Oncology, Amphia Hospital, Breda), **Joyce Meijer** (Department of Research and Development, Netherlands Comprehensive Cancer Organization (IKNL)), **Karin. J. Beelen** (Department of Medical Oncology, Rijnstate Hospital, Arnhem), **Klaartje van Engelen** (Department of Human Genetics, Amsterdam UMC and University of Amsterdam, Amsterdam), **Leonienke F. C. Dols** (Department of Surgery, Amphia Hospital, Breda), **Linda de Munck** (Department of Research and Development, Netherlands Comprehensive Cancer Organization (IKNL)), **Marjan van Hezewijk** (Department of Radiation Oncology, Radiotherapiegroep, Arnhem), **M. R. F. (Frederiek) Bosscher** (Department of Surgery, Ommelander Hospital, Groningen), **Marian B. E. Menke-Pluijmers** (Department of Surgery, Albert Schweitzer Hospital, Dordrecht), **Margrethe Schlooz-Vries** (Department of Surgery, Radboud University Medical Centre, Nijmegen), **Marieke E. Straver** (Department of Surgery, Haaglanden Medisch Centrum, Den Haag), **Martinus A. Beek** (Department of Surgery, Isala klinieken, Zwolle), **Maud Bessems** (Department of Surgery, Jeroen Bosch Hospital, Den Bosch), **Marije C. Gordinou de Gouberville** (Department of Surgery, Wilhelmina Hospital, Assen and Chirurgen Coöperatie Noord Nederland), **Milou H. Martens** (Department of Surgery, Laurentius Hospital, Roermond), **Miriam L. Hoven-Gondrie** (Department of Surgery, Gelderse Vallei, Ede), **Marie-Jeanne T. F. D. Vrancken Peeters** (Department of Surgery, Netherlands Cancer Institute–Antoni van Leeuwenhoek Hospital and Amsterdam University Medical Centre, Amsterdam), **Patricia Jansen** (Department of Surgery, Elizabeth Tweesteden Hospital, Tilburg), **Peter A. Neijenhuis** (Department of Surgery, Alrijne Hospital, Leiderdorp), **Rhodé M. Bijlsma** (Department of Medical Oncology, University Medical Centre Utrecht, Cancer Centre, Utrecht), **Robert-Jan Schipper** (Department of Surgery, Catharina Hospital, Eindhoven), **Ramon R. J. P. van Eekeren** (Department of Surgery, Rijnstate Hospital, Arnhem), **Thomas Schok** (Department of Surgery, VieCuri Medical Centre, Venray), **Tim C. van Sprundel** (Department of Surgery, Ommelander Hospital, Groningen), **Tim H. C. Damen** (Department of Plastic, Reconstructive, and Hand Surgery, Ikazia Hospital, Rotterdam; van Weel Bethesda Hospital, Dirksland; Spijkenisse MC, Spijkenisse), **Titia E. Lans** (Department of Surgery, Admiraal de Ruyterziekenhuis, Goes and Vlissingen), **Vivianne C. G. Tjan-Heijnen** (Department of

Medical Oncology, GROW, Maastricht University Medical Centre, Maastricht), and **Yvonne L. J. Vissers** (Department of Surgery, Zuyderland Medical Centre, Heerlen).

REFERENCES

1. Maes-Carballo M, Gomez-Fandino Y, Reinoso-Hermida A, Estrada-Lopez CR, Martin-Diaz M, Khan KS, Bueno-Cavanillas A. Quality indicators for breast cancer care: A systematic review. *Breast* 2021;59: 221-231.
2. Dreyer NA, Garner S. Registries for robust evidence. *JAMA* 2009;302(7): 790-791.
3. Stey AM, Russell MM, Ko CY, Sacks GD, Dawes AJ, Gibbons MM. Clinical registries and quality measurement in surgery: a systematic review. *Surgery* 2015;157(2): 381-395.
4. Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Ann Intern Med* 2008;148(2): 111-123.
5. Vos EL, Lingsma HF, Jager A, Schreuder K, Spronk P, Vrancken Peeters M, Siesling S, Koppert LB. Effect of Case-Mix and Random Variation on Breast Cancer Care Quality Indicators and Their Rankability. *Value Health* 2020;23(9): 1191-1199.
6. van Bommel AC, Spronk PE, Vrancken Peeters MT, Jager A, Lobbes M, Maduro JH, Mureau MA, Schreuder K, Smorenburg CH, Verloop J, Westenend PJ, Wouters MW, Siesling S, Tjan-Heijnen VC, van Dalen T, Audit NBC. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: The national NABON Breast Cancer Audit. *J Surg Oncol* 2017;115(3): 243-249.
7. Stimuleren uitkomstgerichte zorg https://www.rijksoverheid.nl/onderwerpen/kwaliteit-van-de-zorg/stimuleren-uitkomstgerichte-zorg [27-07 2023].
8. Eindrapport aandoeningswerkgroep Mammacarcinoom: Programma Uitkomstgerichte Zorg – Lijn 1 'Meer inzicht in uitkomsten'. www.platformuitkomstgerichtezorg.nl; 2023.
9. Beck N, Hoeijmakers F, van der Willik EM, Heineman DJ, Braun J, Tollenaar R, Schreurs WH, Wouters M. National Comparison of Hospital Performances in Lung Cancer Surgery: The Role of Case Mix Adjustment. *Ann Thorac Surg* 2018;106(2): 412-420.
10. Lijftogt N, Vahl AC, Wilschut ED, Elsman BHP, Amodio S, van Zwet EW, Leijdekkers VJ, Wouters M, Hamming JF, Dutch Society of Vascular Surgery tSCotDSAA, the Dutch Institute for Clinical A. Adjusted Hospital Outcomes of Abdominal Aortic Aneurysm Surgery Reported in the Dutch Surgical Aneurysm Audit. *Eur J Vasc Endovasc Surg* 2017;53(4): 520-532.
11. Fischer C, Lingsma H, Hardwick R, Cromwell DA, Steyerberg E, Groene O. Risk adjustment models for short-term outcomes after surgical resection for oesophagogastric cancer. *Br J Surg* 2016;103(1): 105-116.
12. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf* 2011;20(10): 869-874.
13. Clavien PA, Barkun J, de Oliveira ML, Vauthey JN, Dindo D, Schulick RD, de Santibanes E, Pekolj J, Slankamenac K, Bassi C, Graf R, Vonlanthen R, Padbury R, Cameron JL, Makuuchi M. The Clavien-Dindo classification of surgical complications: five-year experience. *Ann Surg* 2009;250(2): 187-196.
14. Panhofer P, Ferenc V, Schutz M, Gleiss A, Dubsky P, Jakesz R, Gnant M, Fitzal F. Standardization of morbidity assessment in breast cancer surgery using the Clavien Dindo Classification. *Int J Surg* 2014;12(4): 334-339.
15. van Linschoten RCA, Amini M, van Leeuwen N, Eijkenaar F, den Hartog SJ, Nederkoorn PJ, Hofmeijer J, Emmer BJ, Postma AA, van Zwam W, Roozenbeek B, Dippel D, Lingsma HF, Investigators MCR. Handling missing values in the analysis of between-hospital differences in ordinal and dichotomous outcomes: a simulation study. *BMJ Qual Saf* 2023;32(12): 742-749.
16. van Buuren SG-O, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011;45(3): 1–67.

17. Ten Wolde B, Kuiper M, de Wilt JHW, Strobbe LJA. Postoperative Complications After Breast Cancer Surgery are Not Related to Age. *Ann Surg Oncol* 2017;24(7): 1861-1867.
18. Al-Hilli Z, Wilkerson A. Breast Surgery: Management of Postoperative Complications Following Operations for Breast Cancer. *Surg Clin North Am* 2021;101(5): 845-863.
19. Jonczyk MM, Jean J, Graham R, Chatterjee A. Trending Towards Safer Breast Cancer Surgeries? Examining Acute Complication Rates from A 13-Year NSQIP Analysis. *Cancers (Basel)* 2019;11(2).
20. de Boniface J, Szulkin R, Johansson ALV. Major surgical postoperative complications and survival in breast cancer: Swedish population-based register study in 57 152 women. *Br J Surg* 2022;109(10): 977-983.
21. Al-Hilli Z, Thomsen KM, Habermann EB, Jakub JW, Boughey JC. Reoperation for Complications after Lumpectomy and Mastectomy for Breast Cancer from the 2012 National Surgical Quality Improvement Program (ACS-NSQIP). *Ann Surg Oncol* 2015;22 Suppl 3: S459-469.
22. Potter S, Trickey A, Rattay T, O'Connell RL, Dave R, Baker E, Whisker L, Skillman J, Gardiner MD, Macmillan RD, Holcombe C, TeaM, iBra-2 Steering Groups tBRRC, the Mammary Fold A, Research C. Therapeutic mammaplasty is a safe and effective alternative to mastectomy with or without immediate breast reconstruction. *Br J Surg* 2020;107(7): 832-844.
23. Berlin NL, Tandon VJ, Qi J, Kim HM, Hamill JB, Momoh AO, Pusic AL, Wilkins EG. Hospital Variations in Clinical Complications and Patient-reported Outcomes at 2 Years After Immediate Breast Reconstruction. *Ann Surg* 2019;269(5): 959-965.
24. Fischer C, Lingsma HF, van Leersum N, Tollenaar RA, Wouters MW, Steyerberg EW. Comparing colon cancer outcomes: The impact of low hospital case volume and case-mix adjustment. *Eur J Surg Oncol* 2015;41(8): 1045-1053.
25. Algera MD, Baldewpersad Tewarie NMS, Driel WJV, van Ham M, Slangen BFM, Kruitwagen R, Wouters M, Participants of the Dutch Gynecological Oncology Audit Collaborator G. Case-mix adjustment to compare hospital performances regarding complications after cytoreductive surgery for ovarian cancer: a nationwide population-based study. *Int J Gynecol Cancer* 2023;33(4): 534-542.
26. Lemij AA, van der Plas-Krijgsman WG, Bastiaannet E, Merkus JWS, van Dalen T, Vulink AJE, van Gerven L, Guicherit OR, Linthorst-Niers EMH, Lans TE, van der Pol CC, Wyld L, Morgan JL, Portielje JEA, de Glas NA, Liefers GJ. Predicting postoperative complications and their impact on quality of life and functional status in older patients with breast cancer. *Br J Surg* 2022;109(7): 595-602.
27. Jonczyk MM, Fisher CS, Babbitt R, Paulus JK, Freund KM, Czerniecki B, Margenthaler JA, Losken A, Chatterjee A. Surgical Predictive Model for Breast Cancer Patients Assessing Acute Postoperative Complications: The Breast Cancer Surgery Risk Calculator. *Ann Surg Oncol* 2021;28(9): 5121-5131.
28. Mrad MA, Al Qurashi AA, Shah Mardan QNM, Alqarni MD, Alhenaki GA, Alghamdi MS, Fathi AB, Alobaidi HA, Alnamlah AA, Aljehani SK, Daghistani G, Alsharif TH. Predictors of Complications after Breast Reconstruction Surgery: A Systematic Review and Meta-analysis. *Plast Reconstr Surg Glob Open* 2022;10(12): e4693.
29. Elfrink AKE, van Zwet EW, Swijnenburg RJ, den Dulk M, van den Boezem PB, Mieog JSD, Te Riele WW, Patijn GA, Leclercq WKG, Lips DJ, Rijken AM, Verhoef C, Kuhlmann KFD, Buis CI, Bosscha K, Belt EJT, Vermaas M, van Heek NT, Oosterling SJ, Torrenga H, Eker HH, Consten ECJ, Marsman HA, Wouters M, Kok NFM, Grunhagen DJ, Klaase JM, Dutch Hepato Biliary Audit G. Case-mix adjustment to compare nationwide hospital performances after resection of colorectal liver metastases. *Eur J Surg Oncol* 2021;47(3 Pt B): 649-659.

30. Guideline breast reconstruction: 5.Plastic surgeon for oncoplastic surgery. https://richtlijnendatabase.nl/en/richtlijn/breast_reconstruction/plastic_surgeon_for_oncoplastic_surgery.html [23/02/2024 2024].
31. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338: b2393.
32. Ong WL, Schouwenburg MG, van Bommel ACM, Stowell C, Allison KH, Benn KE, Browne JP, Cooter RD, Delaney GP, Duhoux FP, Ganz PA, Hancock P, Jagsi R, Knaul FM, Knip AM, Koppert LB, Kuerer HM, McLaughin S, Mureau MAM, Partridge AH, Reid DP, Sheeran L, Smith TJ, Stoutjesdijk MJ, Vrancken Peeters M, Wengstrom Y, Yip CH, Saunders C. A Standard Set of Value-Based Patient-Centered Outcomes for Breast Cancer: The International Consortium for Health Outcomes Measurement (ICHOM) Initiative. *JAMA Oncol* 2017;3(5): 677-685.
33. Dehal A, Abbas A, Johna S. Comorbidity and outcomes after surgery among women with breast cancer: analysis of nationwide in-patient sample database. *Breast Cancer Res Treat* 2013;139(2): 469-476.
34. Beck N, van Bommel AC, Eddes EH, van Leersum NJ, Tollenaar RA, Wouters MW, Dutch Clinical Auditing Group. The Dutch Institute for Clinical Auditing: Achieving Codman's Dream on a Nationwide Basis. *Ann Surg* 2020;271(4): 627-631.



Chapter 4

Comprehensive Evaluation of Quality Indicators: Analyzing the Dutch Breast Cancer Audit

Elfi M. Verheul*, Margrietha van der Linde*, Hester F. Lingsma, Elvira Vos, Sabine Siesling, Linetta B. Koppert; NBCA consortium.

*Shared first authorship

Published in: International Journal of Health Policy and Management (IJHPM), 2025.

ABSTRACT

Background: Quality indicators (QIs) are widely used to benchmark hospital performance and improve quality of care but are often based on expert opinion rather than data-driven assessment. This study aims to evaluate QIs, using a framework that assesses four criteria: Feasibility, discriminative ability, validity, and reliability.

Methods: We used data from the Dutch breast cancer registry (NABON Breast Cancer Audit, NBCA) and included all surgically treated breast cancer patients in the Netherlands between 2021-2023. Eighteen QIs were evaluated. Feasibility was determined by QI numerator completeness, with >90% data availability considered feasible. Discriminative ability was assessed by between-hospital variation in QI scores, where an interquartile range (IQR) >10% indicated good discrimination. Validity was evaluated by the impact of case-mix adjustment and considered low when (pseudo-) R^2 <0.10. The (pseudo-)R^2 reflects the proportion of variance in QI scores explained by all case-mix variables in a regression model. Reliability was assessed by rankability, the proportion of between-hospital variation not due to chance and, therefore, explainable by quality of care. Rankability >75% was considered high.

Results: After exclusion of one QI with feasibility <1%, feasibility ranged from 80.2%-100%, and 15 QIs were feasible. Overall, the IQR ranged from 1 to 36, with 8 QIs having an IQR higher than 10, indicating good discriminative ability. The (pseudo-)R^2 ranged from 0.01-0.53, with 11 QIs showing low case-mix impact. Rankability ranged from 0-69%, with none of the QIs having a high rankability. None of the QIs met all preset criteria, but six QIs met at least three out of four criteria.

Conclusion: The QIs employed by the Dutch breast cancer registry fulfilled most criteria, but rankability is a concern and requires specific attention, especially for public reporting. Our results show the importance of considering feasibility, discriminative ability, validity, and reliability when evaluating QIs, and these should also be taken into account when developing new QIs.

KEY MESSAGE

Implications for policy makers

- Evaluating quality indicators (QIs) used for provider comparisons is essential to avoid harm caused by steering decisions based on invalid or unreliable data, especially in publicly reporting of data, which has become an increasing concern.
- For public reporting purposes, it is essential to report quality indicator scores of individual hospitals over multiple years or at another aggregate level to achieve higher rankability.
- New QIs should be developed with the criteria feasibility, discriminative ability, validity, and reliability in mind, to avoid creating invalid QIs based solely on expert opinion.
- We provide a practical guide for a data-driven evaluation of quality indicators to assess feasibility, discriminative ability, validity, and reliability, which could be applied in other quality registries

Implications for the public

As quality of care is difficult to measure, quality indicators (QIs) aim to provide signals of quality. Especially when publicly reported, it is important that individual hospitals scores are valid and reliable to prevent unnecessary harm. Valid and reliable QIs guide healthcare providers improving care, assist patients in making informed choices about their care providers' quality and support designing pay-for-performance strategies that incentivise high-quality care. Using a framework of the European Observatory on Health Systems and Policies, we evaluated all QIs for breast cancer care in the Netherlands. We found that while most criteria are met, reliability remains a concern. For more reliable public reporting, we recommend reporting QI scores over multiple years or at an aggregate level to improve reliability. Our findings directly impact the Dutch breast cancer quality registry and could also influence other quality registries if they adopt our approach, as outlined in our practical guide.

INTRODUCTION

The increasing interest in identifying and addressing variation in quality of care is supported by quality registries[1], which aim to assess and monitor healthcare performance using outcome, process, and structure indicators[2]. These Quality Indicators (QIs) have multiple applications, including monitoring and benchmarking. Internally, organizations such as hospitals use the QIs to improve their own practices, while benchmarking involves comparing hospitals' performance with that of others or national standards. Benchmarking sometimes includes public reporting, where QIs are made publicly available to support patient decision-making, contracting of care by health insurers, and evaluations by regulators[3,4].

For benchmarking, QIs must accurately measure quality of care, requiring them to be both valid and reliable[5,6]. If QIs are valid and reliable, observed variation may actually represent differences in quality. Meaningful benchmarking also depends on adequate data quality; however, achieving this quality also creates an administrative burden, potentially undermining the feasibility. Additionally, to clearly identify areas for improvement, indicators must have sufficient discriminative ability.

In 2011, the multidisciplinary NABON Breast Cancer Audit (NBCA) was developed to measure the quality of care in surgically treated Dutch breast cancer patients. Initially the set consisted of thirty-two QIs[7], and it evolved over time. In the context of the NBCA, there is an annual assessment of the QI set based on expert opinion by the scientific committee[7]. New QIs are introduced based on new insights and clinical guideline revisions, while existing indicators are removed from the set when they are presumed to no longer meet the criteria of feasibility, relevance, validity, and reliability[1,7,8].

Some of these criteria have been (partly) assessed empirically in the NBCA. Feasibility has proven to be good, as all hospitals are required to provide data to the NBCA[9]. Validity and reliability of six QIs were evaluated during 2011-2016 by Vos et al.[5], which showed some deficiencies in both criteria. Given that the QI set used in 2024 consists of 19 partly new QIs, the majority have not undergone any assessment. The aim of this study is to perform a comprehensive quantitative evaluation of the current QI set of the NBCA using a framework designed for the assessment of QIs[10].

METHODS

Data

We used data from the NABON Breast Cancer Audit (NBCA). This Dutch national quality registry includes all patients surgically treated for breast cancer in the Netherlands,

including Ductal Carcinoma in Situ (DCIS), invasive carcinoma, Paget's disease and inflammatory carcinoma. Hospitals can either register the data themselves (20% of the hospitals) or make use of the Netherlands Cancer Registry (NCR) hosted by the Netherlands Comprehensive Cancer Organisation (IKNL). We included all patients that underwent surgery during the years 2021-2023.

Quality indicators (QIs)

We considered 19 indicators for evaluation (Table 1). The QI set of the Dutch breast cancer quality registry of 2024 consists of 18 indicators, categorized according to Donabedian's framework: one structure indicator, four outcome indicators, and 14 process indicators[2]. We added the QI "MRI for lobular breast cancer" which was included in the set of 2023, but removed in 2024 after its objective was met[11]. The structure indicator (QI-1) was excluded from the data-driven assessment due to methodological limitations in this format. All other QIs were included. Nine QIs are applied to different subgroups, related to the type of surgery or treatment trajectory.

QI scores were calculated based on numerator and denominator definitions per hospital (Table 1). Binary process indicators (i.e., adherence scores) were calculated by dividing the number of patients who adhered to the QI for each individual hospital (numerator) by the number of patients per hospital to whom the QI was applicable (denominator). Continuous process indicators QI-7 and QI-13 were calculated as the median time in calendar days per year. Outcome indicators were calculated by dividing the number of events (i.e., the occurrence of the specific quality indicator) for each hospital (numerator) by the total number of patients per hospital who were eligible to be included in the indicator (denominator).

Framework

The European Observatory on Health Systems and Policies summarised several frameworks to assess the quality of indicators, to standardise the evaluation (Supplemental figure S1)[13]. The authors emphasise that based on the goal of the QI and the specific setting the relative importance of each of the criteria may differ. For example, when QIs are publicly available, the criteria validity and reliability are more important to ensure fair hospital comparisons[5,14]. For the evaluation of the QIs in this study, we selected the criteria that were both relevant for between-hospital comparisons and were quantitatively assessable. These criteria were feasibility, discriminative ability, validity, and reliability, and were evaluated for all the 18 selected QIs.

Table 1: Description of quality indicators that are selected for evaluation.

QI	Type of indicator	Definition	Development year[1]	Currently publicly reported
QI-1	Structure	**Breast cancer team** QI1a: Number of newly diagnosed patients with invasive BC or DCIS who underwent surgical treatment QI1b: Number of certified medical oncologists treating BC QI1c: Number of certified oncologic surgeons treating BC QI3d. Number of plastic surgeons performing reconstruction surgery QI3e: Number of patients who have their own case manager, according to the SONCOS[2] definition, and whether this is registered in the patient's file QI3f: Is there documented collaboration with, or involvement in, the multidisciplinary team of a clinical genetics department QI3g: Is the collection of family medical history, the assessment of referral criteria, the availability of genetic testing, and the provision of urgent counselling and DNA diagnostics structurally integrated into the care pathway	a,b,c,d: 2019 e: 2020 f,g: 2024	Yes
QI-2	Process	**Breast contour preserving treatment** *Numerators*: Number of patients who underwent breast contour preserving treatment after: QI2a. surgery (b+c+d) QI2b. breast-conserving surgery including re-lumpectomies without neoadjuvant chemotherapy[3] QI2c: breast-conserving surgery following neoadjuvant systemic[3] therapy. QI2d: mastectomy with immediate reconstruction (autologous and/or prothesis) *Denominator*: Number of surgically treated patients with non-metastasized BC[2]	2016	Yes
QI-3	Process	**Immediate breast reconstruction in invasive disease[4]** *Numerator*: Number of patients who underwent immediate reconstruction after mastectomy for invasive non-metastasized BC[2], where reconstruction is performed by a plastic surgeon: QI3a. immediate reconstruction, total (b+c+d+e) QI3b. immediate reconstruction with prostheses QI3c. immediate autologous reconstruction QI3d. immediate reconstruction with prostheses and autologous tissue QI3e. immediate reconstruction, other *Denominator*: Number of patients undergoing mastectomy for invasive non-metastasized BC[4]	2012	Yes

Table 1: Description of quality indicators that are selected for evaluation. (*continued*)

QI	Type of indicator	Definition	Development year[1]	Currently publicly reported
QI-4	Process	**Immediate breast reconstruction in DCIS** *Numerator*: Number of patients who underwent immediate reconstruction after mastectomy for DCIS, where reconstruction is performed by a plastic surgeon: QI4a. immediate reconstruction, total (b+c+d+e) QI4b. immediate reconstruction with prostheses QI4c. immediate autologous reconstruction QI4d. immediate reconstruction with both prostheses and autologous tissue QI4e. immediate reconstruction, other *Denominator*: Number of patients undergoing mastectomy for DCIS	2012	Yes
QI-5	Process	**Seen by radiation oncologist prior to neoadjuvant chemotherapy** *Numerator*: Number of patients seen by the radiotherapist within 28 days of starting treatment *Denominator*: number of patients with invasive BC treated with neoadjuvant chemotherapy[3], surgery, and postoperative radiotherapy	2011	Yes
QI-6	Process	**Neoadjuvant systemic therapy* for patients with triple negative and HER2 positive BC** *Numerator*: Number of patients receiving neoadjuvant systemic therapy[3] *Denominator*: Number of patients under 70 years old with cT2/3/4 any N M0 triple-negative or HER2/Neu positive BC[4]	2020	Yes
QI-7	Process	**Time between diagnosis of invasive breast cancer and primary treatment** Median time in calendar days between date of biopsy where diagnosis is established and the start of: QI7a. any treatment (b+c+d) QI7b. neoadjuvant chemotherapy[3] QI7c. first surgery (excluding direct reconstruction) QI7d. first surgery with direct reconstruction	2011 (2011-2017 in binary form: <5 weeks yes/no)	Yes
QI-8	Process	**PROMs response** *Numerator*: Completing PROM[5] questionnaires at T0 (before start treatment) and at T1 (1 year after start treatment) *Denominator*: Number of patients registered in the NBCA	2019	Yes
QI-9	Process	**MRI-mamma prior to neoadjuvant chemotherapy** *Numerator*: Number of patients who received breast MRI prior to the start of neoadjuvant chemotherapy[3] *Denominator*: Number of patients with primary invasive non-metastasized BC treated with neoadjuvant chemotherapy	2012	Yes

Table 1: Description of quality indicators that are selected for evaluation. (*continued*)

QI	Type of indicator	Definition	Development year[1]	Currently publicly reported
QI-10	Process	**Radiotherapy for locally advanced BC requiring mastectomy** *Numerator*: Number of patients receiving radiotherapy after mastectomy *Denominator*: Number of patients with primary invasive locally advanced non-metastasized BC[4] without distant metastases (Dutch guideline: cT4, pT4, pT3N1 or ≥pN2) who underwent a mastectomy	2011	No
QI-11	Outcome	**Irradical resection in primary breast-conserving surgery for invasive BC** *Numerator*: Number of patients with more than focally positive margins[6] *Denominator*: Number of patients who underwent first breast-conserving surgery (without neoadjuvant therapy*) for primary invasive non-metastasized BC[4]	2011	No
QI-12	Outcome	**Irradical resection in primary breast-conserving surgery for DCIS** *Numerator*: Number of patients with positive margins *Denominator*: Number of patients who underwent a first breast-conserving surgery for DCIS	2011	No
QI-13	Process	**Time from last surgery to start adjuvant therapy** Median time in calendar days between the last surgery and: QI13a. adjuvant treatment QI13b. adjuvant chemotherapy QI13c. adjuvant radiotherapy *Denominator*: Number of surgical treated patients with primary invasive non-metastasized BC and/or DCIS	2011 (2011-2017 in binary form: <5 weeks yes/no)	No
QI-14	Process	**Trial participation** *Numerator*: Number of patients who participated in a registered clinical trial *Denominator*: All surgically treated patients with invasive BC or DCIS	2020	No
QI-15	Outcome	**Complicated course after surgery** *Numerator*: Number of patients with a complicated course[7] *Denominator*: Number of patients with invasive BC or DCIS receiving: QI15a: lumpectomy QI15b: mastectomy without immediate reconstruction QI15c: mastectomy with immediate reconstruction	2020	No
QI-16	Outcome	**Complicated course after chemotherapy** *Numerator*: Number of patients with a complicated[8] course after QI16a: chemotherapy, total (b+c) QI16b: neoadjuvant chemotherapy QI16c: adjuvant chemotherapy *Denominator*: Number of patients with invasive non-metastasized BC treated with chemotherapy and underwent surgery	2020	No

Table 1: Description of quality indicators that are selected for evaluation. (*continued*)

QI	Type of indicator	Definition	Development year[1]	Currently publicly reported
QI-17	Process	**Collaboration of surgeon and plastic surgeon for invasive BC** *Numerator*: Number of patients receiving surgery by a general surgeon and plastic surgeon together *Denominator*: Number of patients diagnosed with invasive non-metastasized BC undergoing primary breast-conserving surgery	2022	No
QI-18	Process	**Collaboration of surgeon and plastic surgeon for DCIS** *Numerator*: Number of patients receiving surgery by a general surgeon and plastic surgeon together *Denominator*: Number of patients diagnosed with DCIS undergoing primary breast-conserving surgery	2022	No
QI-19	Process	**MRI for lobular breast cancer** *Numerator*: Number of patients with lobular BC receiving an MRI QI19a: prior to surgery, total (b+c) QI19b: prior to breast conserving surgery QI19c: prior to mastectomy *Denominator*: Number of patients diagnosed with lobular non-metastasized BC undergoing surgery (a), primary breast-conserving surgery (b) or mastectomy (c)	2017	No[9]

*QI = quality indicator; BC = Breast cancer, DCIS = Ductal carcinoma in situ, NBCA = Nabon Breast Cancer Audit; PROM = Patient Reported Outcome Measurement; MRI = magnetic resonance imaging; NA = not applicable. [1]For indicators developed before 2019, there was no distinction between internal and external use of QIs, this implicates that QIs were made publicly available directly with their development. It is possible that QIs the transparency of some QIs have been reverted. [2]SONCOS is the Dutch multidisciplinary oncology platform aimed at describing the standards of quality of oncologic care that must be met. These standards are established by oncological professionals[34]. [3]Neoadjuvant systemic therapy is defined as neoadjuvant chemotherapy and/or HER2 blockade. [4]Invasive BC can be with or without a DCIS component. [5]PROMs consisting of: EORTC-QLQ-C30, and/or EORTC-QLQ-BR23, and/or BREAST-Q. [6]More than focally positive margins is defined as tumour touching the inked margin over a length of 4 mm or more. [7]A complicated course is defined as grade 3 and higher according to the Clavien-Dindo classification[39]. [8]A complicated course is defined as premature discontinuation or clinical admission due to adverse effects of chemotherapy. [9]This QI was publicly available from 2020- 2023 but removed from the set in 2024. *Neoadjuvant systemic therapy is defined as neoadjuvant chemotherapy and/or HER2 blockade.*

Statistical analysis

We presented descriptive statistics as count and percentages or mean with standard deviation (SD) for calendar year 2023. Missing baseline characteristics were imputed using single imputation, assuming values were missing at random. We fitted the regression imputation model with all outcomes and potential predictors. Missing values in the QI numerator were recoded to zero, as this is common practice in the Netherlands. Depending on the QI definition, this 'numerator to zero' approach can positively or negatively impact the QI score. For example, for the QI complicated course after surgery (QI-15), a lower score indicates better performance. On the contrary, for MRI-mamma prior to neoadjuvant chemotherapy (QI-9), a lower score might be interpreted as worse performance. We used this approach for the evaluation of the criteria discriminative ability, influence of case-mix and rankability. We evaluated all criteria on the most recent calendar year (2023) to align as closely as possible with the way indicators are used in practice, except for the influence of case-mix which was assessed over the years 2021-2023. Additionally, we performed a sensitivity analysis for the QI with the lowest feasibility (QI-5) by imputing outcomes using a logistic regression model for single imputation.

Feasibility

Feasibility encompasses data quality and availability[15]. We quantified feasibility by calculating the completeness of the QI numerator, having already defined the denominator population. To determine feasibility, we used and extended thresholds set by Huijben et al.[16], classifying data completeness of the numerator above 90% as good, between 70-90% as moderate, and below 70% as poor. If feasibility was <25%, the QI was excluded for evaluation on remaining criteria.

Discriminative ability: Between-hospital variation

Discriminative ability, further denoted as between-hospital variation, means the ability of a QI to distinguish in hospital performance. If there is only little hospital variation, there may be small improvement possibilities, while wide hospital variation offers more room for improvement. We quantified between-hospital variation as the median and IQR of the QI score. Discriminative ability was categorised as poor (IQR<5), moderate (IQR 5-10), and good (IQR>10).

Validity: influence of case-mix adjustment

Although validity is a broad concept, in this study we only assessed the impact of adjustment for baseline patient and tumour characteristics (i.e., case-mix). In the context of between-hospital comparisons little impact of case-mix adjustment is favourable, as with high case-mix influence observed between-hospital variation in the QI score is more likely due to differences in underlying patient population rather than quality of care. We combined three years of data (2021-2023) to yield more robust results, and fitted a case-mix model for each QI separately. We identified potential case-mix variables through

literature and expert opinion, which were age, sex, Body Mass Index (BMI), smoking, multifocality, cTNM-stage, histology, differentiation grade, and receptor status[5]. For the continuous variable age, we checked for non-linearity of the association between predictor and outcome. We assessed model performance using Nagelkerke's pseudo-R^2 for the binary QIs and the R^2 for the continuous QIs. The (pseudo-)R^2 quantifies the extent to which the variance in a QI score is explained by the predictors in the model[17]. The closer the (pseudo-)R^2 is to 1, the better the model can explain why patients have different outcomes. We considered a (pseudo-)R^2 ≥0.25 as significant, a (pseudo-)R^2 between 0.10-0.25 as moderate, and a (pseudo-)R^2 lower than 0.10 as minimal case-mix influence.

When patient populations are comparable across hospitals, the influence of adjustment on hospital comparisons is likely to be minimal, even if the case-mix variables have a large impact on the QI. Therefore, we further assessed the effect of case-mix adjustment on hospital comparisons using the observed/expected (O/E)-ratio. This ratio is determined by dividing the observed by the expected number of events per hospital. An O/E-ratio exceeding 1 indicates more events than expected, while below 1 indicates fewer. We estimated both unadjusted and case-mix adjusted O/E-ratios. The observed number of events for each ratio was the individual hospital QI score. For the unadjusted O/E ratio, the expected number of events was calculated as the mean QI score across all hospitals for binary QIs and the mean number of days across all hospitals for continuous QIs. For the case-mix adjusted O/E ratio, the expected number of events was determined as the predicted probability for each hospital for binary QIs and the predicted number of days for each hospital for continuous QIs, with both estimates derived from the case-mix adjustment model. We visualised the O/E-ratios for each hospital in a scatterplot, with the y-axis showing unadjusted and the x-axis the case-mix adjusted O/E-ratio. The total deviation from the diagonal was quantified with the root mean squared error (RMSE), which is the square root of the average squared deviance per hospital. A lower RMSE indicates a smaller effect of case-mix adjustment on hospital comparisons. If the hospitals are exactly on the diagonal, there is no effect of case-mix adjustment (RMSE = 0).

Reliability: rankability

Reliability is related to the ability of QIs to meaningfully distinguish hospitals. We operationalized this with rankability, which reflects the proportion of variation that is not due to chance.[18] For example, a rankability of 60% indicates that 60% of the observed differences are differences not caused by chance, but potentially by quality of care. Rankability is calculated using formula [1], in which ρ denotes rankability, τ^2 denotes the variance of a random effects model where hospital is added as a random intercept (i.e., between-hospital variation), and s_i denotes the standard error of the estimated hospital effect for hospital i. The standard error reflects the precision with which the hospital fixed effects are estimated and is highly impacted by the total number of patients treated in that hospital. Rankability is categorised as low (<50%), moderate (50%-75%), or high (>75%)[5].

$$\rho = \frac{\tau^2}{\tau^2 + median(s_i^2)} \quad [1]$$

For the calculation of the rankability, we excluded hospitals that treated fewer than two patients per year, for technical reasons. Additionally, in some instances we used a simplification of the full case-mix model leaving out variables with no events (Supplementary Table S1.1). We performed sensitivity analyses by calculating rankability over three calendar years (2021-2023) to determine if this approach enhances rankability.

After the analysis, we evaluated which of the QIs met all the criteria to determine their overall performance.

Practical guide

We developed a practical guide, including R codes, which is applicable for QI evaluation of other registries. This guide can be found in Appendix 3. All analyses were performed using RStudio, version 4.2.3.

RESULTS

Descriptive statistics

In 2023, 15,301 patients with invasive breast cancer (IBC) and 1,953 patients with Ductal Carcinoma In Situ (DCIS) were surgically treated across 71 hospitals in the Netherlands (Table 2). Most patients underwent breast conserving surgery (73%) in comparison with mastectomy (27%). Around one-third of the patients who underwent a mastectomy also received an immediate reconstruction (IBC: 26%, DCIS: 40%). A quarter of the IBC patients underwent neoadjuvant chemotherapy (26%). Patients with DCIS less frequently received radiotherapy (45% vs 69%) and were not treated with systemic therapy. The distribution of patient- tumour and treatment characteristics in 2023 was similar to that of the 2021-2023 cohort.

Table 2: Baseline characteristics of included patients (year 2023)

	2023				2021-2023			
	IBC (N= 15,301[1])		DCIS (N= 1,953[1])		IBC (N= 46,187[1])		DCIS (N= 5,787[1])	
	Total N (%)	Missing N	Total N (%)	Missing N	Total N (%)	Missing N	Total N (%)	Missing N
Patient characteristics								
Age, mean (SD)	63 (13)	0	62 (11)	0	62 (13)	10	61 (11)	1
BMI (SD)	27 (5)	2,608	26 (5)	240	27 (5)	5,987	26 (5)	688
Sex		0		0		0		0
Female	15,196 (99%)		1,941 (99%)		45,864 (99%)		5,742 (99%)	
Male	105 (1%)		12 (1%)		323 (1%)		45 (1%)	
Smoking		4,192		568		12,090		1,678
No or quitted	9,170 (83%)		1,199 (87%)		28,363 (83%)		3,534 (86%)	
Yes	1,939 (17%)		186 (13%)		5,734 (17%)		575 (14%)	
Tumour characteristics								
T-stage		164				620		
cT1	8,331 (55%)				25,154 (55%)			
cT2	5,141 (34%)				15,307 (34%)			
cT3/4	1,250 (8%)				3,872 (8%)			
cTis	415 (3%)				1,234 (3%)			
N-stage		107				385		
cN0	12,534 (83%)				37,733 (82%)			
cN+	2,656 (17%)				8,069 (18%)			
Receptor status		0				0		
HR+,Her2-	11,231 (73%)				34,388 (74%)			
Her2+	1,792 (12%)				5,548 (12%)			
Triple-	2,278 (15%)				6,251 (14%)			

Table 2: Baseline characteristics of included patients (year 2023) (*continued*)

	2023				2021-2023			
Differentiation grade		2,202		170		4,276		507
1	3,146 (24%)		274 (15%)		9,711 (23%)		801 (15%)	
2	7,167 (55%)		671 (38%)		22,570(54%)		1,975 (38%)	
3	2,786 (21%)		829 (47%)		9,630 (23%)		2,504 (47%)	
Histology		1,851				3,359		
No Special Type	10,399 (77%)				33,074 (77%)			
Lobular	1,784 (13%)				5,668 (13%)			
other	1,327 (10%)				4,086 (10%)			
Multifocality		2,626		150		5,194		363
Unifocal	10,362 (82%)		1,710 (95%)		33,378 (81%)		5,154 (95%)	
Multifocal	2,313 (18%)		93 (5%)		7,615 (19%)		270 (5%)	
Treatment characteristics								
Type of surgery		1,652*		43		2,608		170
Lumpectomy	9,799 (72%)		1,388 (73%)		30,765 (71%)		4,077 (73%)	
Mastectomy	3,850 (28%)		522 (27%)		12,814 (29%)		1,540 (27%)	
Immediate reconstruction post-mastectomy**	1,056 (26%)	5	223 (40%)	0	3,459 (27%)	6	660 (43%)	0
Neoadjuvant chemotherapy	3,971 (26%)	37			12,320 (27%)	50		
Adjuvant chemotherapy	1,447 (10%)	230			5,571 (12%)			
Radiotherapy	9,449 (69%)	1,568	875 (45%)	23	31,400 (72%)	2,303	2,644 (46%)	60
Hormone therapy	1,495 (49%)	12,279***			16,147 (48%)	12,613		

*IBC = Invasive breast cancer, DCIS = Ductal carcinoma in situ, SD = standard deviation, BMI = Body Mass Index. *1429 patients are coded as "intention to receive surgery". The specific type of surgery is not yet determined, given that they are still receiving neo-adjuvant chemotherapy (the inclusion year is the time of biopsy) **Only patients receiving mastectomy are included in this variable. *** High number of missingness due to exclusion of the data-entry set of the Netherlands Comprehensive Cancer Organisation in 2023. The ratio is comparable to that observed in 2022.*

Feasibility

For nearly all quality indicators (QIs) included the numerators were complete for >90%, and thus were feasible (Table 3). Moderate feasibility was observed for the QIs 'seen by radiation oncologists prior to neoadjuvant chemotherapy' (QI-5), and 'complications after chemotherapy' (QI-16C). QI-8, PROMs response, showed little feasibility and was consequently excluded from further evaluation.

Between-hospital variation

Half of the QIs showed an interquartile range (IQR) >10, suggesting notable differences between hospitals (Table 3). Low between-hospital variation was mostly observed in QIs with either high (QI-6, QI-9) or low median scores (QI-3C/D/E, QI-4C/D/E, QI-11, QI-14, QI-15).

Case-mix

For four QIs, there was significant impact of patient- and tumour characteristics on the QI score, reflected in the pseudo-R^2 >0.25 (Table 3). Three out of these four QIs are related to surgery, namely breast contour preserving surgery (QI-2), and immediate reconstruction after IBC and DCIS (QI-3, QI-4). The fourth QI substantially affected by case-mix was MRI for lobular BC before mastectomy (QI-19C). Most QIs that have a pseudo-R^2 exceeding 0.25 also exhibit high RMSE scores (Supplementary Table S1.2).

Rankability

None of the QIs showed good rankability (>75%). The highest score was observed for QI-9, MRI-mamma prior to neoadjuvant chemotherapy, which achieved a rankability of 69%. Moderate rankability scores were observed in five QIs (QI-5, QI-7A/C, QI-9, QI-13A, QI-17). Furthermore, all QIs with a low number of patients included in the denominator, or a low number of patients in the nominator showed poor rankability (<50%). For QI-3E and QI-3C/D/E, it was not possible to calculate rankability, as the number of included patients in the numerator was too small for individual hospitals (Supplementary Table S1.1). Combining data of 3 years generally resulted in higher rankabilities (Supplementary Table S1.3).

Performance of QIs

None of the QIs met the predefined thresholds on all evaluated criteria. However, six QIs met three out of four criteria. These were time between diagnosis of invasive breast cancer and primary treatment (QI-7A, QI-7C), radiotherapy for locally advanced BC requiring mastectomy (QI-10), median time last surgery to start any adjuvant therapy (QI-13A), complicated course after mastectomy (QI-15C), complicated course after chemotherapy and after neoadjuvant chemotherapy (QI-16A/B) and collaboration of surgeon and plastic surgeon for DCIS (QI-18). Rankability was observed as the least met criteria, with only (subgroups of) 5 out of 18 QIs having moderate rankability, and the remaining showing

poor rankability. All of four outcome indicators showed poor rankability, and three out of four showed moderate to no between-hospital variation. The sensitivity analyses also emphasized the importance of good feasibility, particularly for a reliable benchmark (Supplementary Table S2.1).

Table 3: Overview of quality indicator performance.

Quality indicator[1]	**N patients**[2]	**Patients per hospital (median, IQR)**	**Feasibility** % data available (hospital range)[3]	**Discriminative ability** Between hospital variation[4] Median (IQR)	**Case-mix** (Pseudo) R-squared[5]	**Reliability** Rankability %
Year(s) of data	**2023**	**2023**	**2023**	**2023**	**2021-2023**	**2023**
QI-2						
QI-2A	13502	186 (108-250)	99.7 (61-100)	76 (72-80)	0.26	32
QI-2B	13502	186 (108-250)	99.8 (61-100)	52 (47-56)	0.53	22
QI-2C	13502	186 (108-250)	99.8 (61-100)	16 (13-19)	0.35	16
QI-2D	13502	186 (108-250)	99.9 (96-100)	7 (5-11)	0.28	31
QI-3						
QI-3A	3881	54 (34-70)	99.9 (96-100)	23 (17-34)	0.38	37
QI-3B	3881	54 (34-70)	99.9 (96-100)	20 (13-27)	0.33	33
QI-3C	3881	54 (34-70)	99.9 (96-100)	1 (0-3)	0.10	32
QI-3D	3881	54 (34-70)	99.9 (96-100)	0 (0-1)	0.05	16
QI-3E	3881	54 (34-70)	99.9 (96-100)	0 (0-1)	0.09	NA
QI-4						
QI-4A	533	7 (4-10)	100 (100-100)	39 (28-50)	0.36	8
QI-4B	533	7 (4-10)	100 (100-100)	28 (21-40)	0.27	8
QI-4C	533	7 (4-10)	100 (100-100)	0 (0-7)	0.07	NA
QI-4D	533	7 (4-10)	100 (100-100)	0 (0-0)	0.05	NA
QI-4E	533	7 (4-10)	100 (100-100)	0 (0-0)	0.10	NA
QI-5	1812	23 (14-36)	80.2 (0-100)	74 (52-88)	0.02	61
QI-6	1466	19 (10-27)	99.8 (75-100)	94 (92-97)	0.10	0
QI-7						
QI-7A	13976	190 (110-256)	92.5 (61-100)	29 (23-38)*	0.03	66
QI-7B	3792	47 (28-72)	99.4 (64-100)	28 (22-34)*	0.03	39
QI-7C	9532	134 (78-180)	100 (96-100)	30 (23-39)*	0.04	66
QI-7D	652	9 (4-12)	100 (100-100)	42 (32-55)*	0.02	27

Table 3: Overview of quality indicator performance. (*continued*)

Quality indicator[1]	**N patients**[2]	**Patients per hospital (median, IQR)**	**Feasibility** % data available (hospital range)[3]	**Discriminative ability** Between hospital variation[4] Median (IQR)	**Case-mix** (Pseudo) R-squared[5]	**Reliability** Rankability %
QI-8	15596		<1	-	-	-
QI-9	2773	35 (20-53)	99.5 (64-100)	94 (91-97)	0.07	69
QI-10	535	7 (4-10)	100 (70-100)	75 (58-90)	0.12	0
QI-11	7528	101 (61-141)	99.4 (95-100)	2 (1-3)	0.06	0
QI-12	1384	18 (11-26)	99.9 (96-100)	17 (12-21)	0.03	15
QI-13						
QI13-A	10533	134 (88-198)	94.4 (94-100)	40 (34-48)*	0.02	52
QI13-B	575	6 (3-11)	100 (100-100)	35 (28-44)*	0.02	23
QI13-C	9412	121 (76-180)	100 (100-100)	40 (34-48)*	0.02	56
QI-14	15596	213 (128-285)	92.5 (0-100)	4 (3-7)	0.04	30
QI-15						
QI15-A	15596	139 (91-210)	93.8 (0-100)	2 (1-3)	0.01	0
QI15-B	11187	44 (26-60)	94 (0-100)	5 (3-8)	0.01	13
QI15-C	4372	14 (10-19)	91.8 (0-100)	10 (5-16)	0.02	0
QI-16						
QI16-A	3785	47 (30-72)	90.5 (0-100)	28 (23-33)	0.03	24
QI16-B	2773	35 (20-53)	91.1 (0-100)	29 (22-34)	0.03	25
QI16-C	1406	18 (8-26)	87.8 (0-100)	31 (24-38)	0.03	0
QI-17	9516	122 (78-179)	97.8 (0-100)	8 (4-13)	0.10	67
QI-18	1347	18 (10-26)	97.9 (0-100)	8 (4-14)	0.05	37
QI-19						
QI19-A	1408	17 (11-27)	100 (100-100)	84 (74-98)	0.17	0
QI19-B	889	11 (7-18)	100 (100-100)	87 (76-92)	0.14	53
QI19-C	519	7 (4-10)	100 (100-100)	81 (72-86)	0.25	20

*IQR = interquartile range; QI = quality indicator. *Number of days. [1]See Table 1 for exact definitions of all quality indicators. [2]The number of patients in denominator. [3]Hosital range (minimum-maximum) of percentage data available. [4]This is the QI score expressed as a percentage, except for continuous outcomes QI7 and QI12. For these QIs, the score represents the median value for each hospital, from which the overall median and IQR were calculated. [5]For continuous outcomes QI7 and QI12 the R-squared is presented instead of the Pseudo R-squared. The colours indicate poor (orange), moderate (yellow) and good (green) performance on the selected criteria.*

DISCUSSION

This study aimed to assess all Quality Indicators (QIs) of the Dutch Breast Cancer Quality Registry (NBCA), applying a comprehensive framework that includes criteria feasibility, discriminative ability, validity, and rankability. None of the QIs fulfilled all criteria, but in 6 QIs three out of the four criteria were met. Feasibility was the most met criterion, with 16 out of 18 QIs performing well, while rankability was the least met with only five QIs showing moderate rankability and the remaining demonstrating poor rankability. Across outcome indicators we observed a pattern of both poor to moderate discriminative ability and rankability.

Evaluation on criteria

Data availability was high for nearly all QIs, likely related to the fact that about 80% of the hospitals use data from the Netherlands Cancer Registry (NCR). Furthermore, the collection of QIs over multiple years may further account for the observed high feasibility; for example, QI-4, collected since 2012, currently demonstrates 100% feasibility[19]. However, for six QIs we observed at least one hospital not providing any data (reflected in feasibility range starting at zero). Three potential explanations may account for this finding. First, five of these QIs were developed after 2020 (three in 2020 and two in 2022), which may suggest limited time for implementation. Second, only one of the six QIs is publicly available; hospitals might prioritize transparent QIs, and therefore, these could theoretically have higher feasibility. Third, variation in the method of data registration may influence feasibility, as both batch-delivery and self-registration are possible[7]. Self-registration in combination with administrative burden may explain lower feasibility. Our sensitivity analyses of the QI with the lowest feasibility (QI-5; 80%), showed that feasibility is important, particularly for a reliable benchmark. Although imputing missing outcomes may be scientifically appropriate, it remains highly sensitive in hospital comparisons, as providers may interpret it as guesswork that might not present the real world practice and feel unjustly penalized. Therefore, we followed in our study the strategy used by quality registries in the Netherlands, which code missing values to zero, as this provides an incentive for complete registration by individual hospitals. It would be good to present the rate of missing values also, to get an impression of the data validity, Nevertheless, we argue that, given the complexity of missing outcome data, the optimal approach to handling missing data should always be determined by experts for each quality registry, or in some cases even for individual QIs.

Overall, most QIs showed relatively high between-hospital variation, indicating that there is potential for improvement on these indicators. For QIs with low to no between-hospital variation, we have two potential explanations: QIs may have made themselves 'unemployable' (i.e., health care providers adhere so well that a process indicator does not exhibit variation anymore), or there was initially no meaningful variation between hospi-

tals. The first explanation was seen by van Bommel et al.[7] who assessed between-hospital variation in the QI measuring the proportion of patients with breast cancer discussed by the multidisciplinary team, where all hospitals scored >90%. We argue that such an indicator has served its aim and can be de-implemented again. Another example are the QIs irradicality for invasive disease (QI-11) as DCIS (QI-12); both have been collected since 2011, and currently show poor and moderate discriminative ability, respectively.

It is reassuring that the influence of case-mix on most QIs seems limited, as this influence was high for only four QIs (Pseudo-R^2 ≥0.25). The high Pseudo-R^2 was accompanied by a considerable difference in adjusted and unadjusted scores (high RMSE) for three out of these four (QI-2 - QI-4), emphasising the importance of case-mix adjustment as unadjusted QI scores will be misleading. The influence of case-mix is not surprising here, as these QIs comprise 'breast contour preserving treatment' (QI-2), and 'immediate breast reconstruction' (QI-3/QI-4), and included case-mix factors like age, BMI and tumour stage are associated with the type of surgery that a patient receives[20,21]. It is important to emphasise that QIs can have a high pseudo-R^2 along with little impact of case-mix adjustment on individual hospital scores (low RMSE) if the case-mix variables are evenly distributed across hospitals. Developing specific case-mix adjustment models is essential to avoid unnecessarily complex adjustments. This could also reduce administrative burden by preventing collection of case-mix variables that do not significantly impact hospital comparisons. Nonetheless, case-mix adjusted QI scores should be interpreted cautiously, as there might still be unmeasured confounding: case-mix differences between hospitals that are not measured. Another solution instead of case-mix adjustment could be to stratify patients based on their characteristics[22]; however, this may lead to smaller patient groups and therefore reduce rankability.

We showed that for nearly all QIs used by the NBCA, rankability was lower than 50% meaning that more than 50% of the observed variation is attributable to chance. Using scores of unreliable QIs for benchmarking may lead to misdirected quality improvement efforts and may potentially have harmful consequences for hospitals. We showed that reporting QI scores over multiple years could be a viable solution to increase the rankability, which is in line with previous work.[23] Here we encounter an inherent tension between rankability and actionability (meaning that the indicator provides clear guidance for improving performance in daily practice). Reporting data over multiple years contradicts the objective of short cycle feedback, which is inherent in a quality registry, and decreases actionability because feedback information should be as recent as possible to make sure that it contains information about the care currently being delivered. We propose that distinguishing between internal and external (i.e., publicly available) QIs is desirable. For external reporting QI scores should span multiple years to increase reliability. Internally, healthcare providers may use the one-year QI scores with detailed information for learning and improving, provided the QI shows moderate rankability. We

believe that newly developed indicators should initially be added to the internal QI set before they are publicly reported. Another strategy, currently employed by the NBCA, is to provide QI scores using funnel plots.[7] Funnel plots show QI scores for individual hospitals incorporating sample size, and show statistically defined control limits[24,25]. They are easy to use and helpful in correct interpretation of the data, stimulating the conversation towards learning and improvements initiatives.

There are QIs that still show poor rankability after combining data from three years, especially with those with low number of patients, or with low between-hospital variation. This prompts a discussion on whether information on hospital performance for these QIs should be publicly reported, as this information may be more suitable for internal monitoring over time and as guidance for low-stakes discussions[4]. Other suggested solutions include excluding QIs with very low or very high event rates[26], increase the sample size by clustering hospitals (e.g. regions)[27], and using composite measures[10,23,28], although aggregating data by location or indicators may decrease actionability[29,30].

Relation of QIs to quality of care

Observed differences in QI scores may suggest variations in quality of care, although it is not easy to determine which practice is truly best; high or low frequencies alone do not imply optimal approaches, nor does the average necessarily represent the most effective standard. For some QIs we noted that it is ambiguous how the QI score relates to quality of care. For example, lower irradicality (QI-11/QI-12) does not always imply better quality of care as removing the tumour with large margin may potentially lead to worse patient outcomes in terms of satisfaction and quality of life due to more deformity or increased need for oncoplastic surgery[31]. The first crucial step before any quantitative assessment is to select relevant QIs and define them unambiguously.

Outcome indicators

There is a push to increase transparency of outcome indicators, as these are seen as the most valuable information for patients[32]. Nevertheless, outcomes are generally more influenced by case-mix and less actionable since outcomes do not provide clear guidance for improving. Van Dishoeck et al.[33] showed that outcomes with low event rates in relatively small samples (per hospital) are not suitable for ranking hospitals. This is in line with our results, where three out of four outcome indicators showed moderate to no between-hospital variation and poor rankability (<26%). Additionally, these authors assessed irradicality for invasive disease over one year (2007) and found a rankability of 53%[33]. Later, Vos et al.[5] calculated rankability for this QI over multiple years (2011-2016) and found a rankability of 22%. Since 2007, changes in the organization of breast cancer care, along with growing awareness of this QI, may have contributed to a decrease in the median irradicality score from 7% in 2007 to 2% in 2023. These changes could explain the 0% rankability observed over one year (2023) and the 10% rankability over three years

(2021-2023). This implicates that outcome indicators are not the most suitable candidates for benchmarking. However, it could still be useful for hospitals to monitor their own data and track internal changes over time.

Strengths and limitations

This is the first data-driven assessment of QIs for breast cancer (BC), providing an objective evaluation of all QIs used in the Netherlands. Our study has some limitations. First, we recoded all missing outcomes to zero in our main analysis. This approach does not affect all QIs homogeneously, and this may have impacted our results. However, as this approach is commonly used in quality registries, to create an incentive for providing data, we have chosen to align as closely as possible with practice. In our sensitivity analyses of QI-5, we performed single imputation of missing outcomes. However, it is questionable if the missingness mechanism is missing at random, as there were hospitals with 100% missingness. Unfortunately we did not have access to extra auxiliary variables for our imputation model which might have influenced the validity of the imputation. Furthermore, we assessed feasibility solely on numerators, which may have led to an underestimation of feasibility. However, since providing data is mandatory for all 71 hospitals delivering breast cancer care in the Netherlands[7], we argue that this probably would not have yielded different results. For statistical reasons, rankability could only be calculated for hospitals with more than two patients. Consequently, some hospitals were excluded from certain rankability calculations, which potentially resulted in an overestimation of the rankability. However, as our categorisation was based on values of rankability <50%, 50%-75%, and >75%, we also assume that this limitation did not influence our main results.

Implications for practice and future research

Our results provide a data-driven overview on which QIs are meaningful for assessing the quality of care for breast cancer patients and can inform the Dutch breast cancer registry, an analysis that – to our knowledge – has not been executed before[1,7]. We recommend using this comprehensive evaluation to refine the QI set for the coming years. Our findings also guide other national registries and international initiatives[34,35]. The results can provide insights into proposed QIs in such initiatives, for example by identifying reliability concerns related to indicators that focus on subgroups, while considering the national and/or healthcare organization context. Additionally, our practical guide in combination with expert opinion offers valuable support for both national and international benchmarking initiatives.

If evaluated QI do not meet the criteria, they should serve as fuel for collective learning and improving rather than publicly report them and making providers feel penalized[4]. Future research should focus on methods to enhance the reliability of breast cancer QIs, such as incorporating composite indicators, and establish the most desirable compromise between reliability and actionability[23]. Additionally, this assessment approach for

evaluation of QIs in hospital settings can be applied to other quality registries using our practical guide (appendix 3), and these criteria should be incorporated when developing new QIs.

CONCLUSION

We demonstrate the importance of thoughtful QI analysis, and therefore these criteria should be considered when developing new QIs. We showed that the QIs employed by the Dutch breast cancer registry fulfilled most criteria, but rankability is a concern that requires specific attention, especially for QIs that are publicly reported. Despite not meeting all criteria, QIs can still offer valuable insights for learning and improvement and could be used as fuel for multidisciplinary conversation. Furthermore, we provide a practical guide for performance analysis of QIs.

Supplementary Information

Supplementary material is available at: https://www.ijhpm.com/article_4800.html

Acknowledgements

The authors thank all, registrar and healthcare professionals for data registration and development of the NBCA.

NBCA Consortium

The NBCA Consortium consists of the following participants, displayed with their affiliations: **Alwine A. Hellingman** (Department of Surgery, Tergooi Medical Centre, Hilversum, the Netherlands), **Anne Brecht Francken** (Department of Surgery, Isala klinieken, Zwolle, the Netherlands), **Annemiek Doeksen** (Department of Surgery, Antonius Ziekenhuis Nieuwegein, Nieuwegein, the Netherlands), **Carolien H.M. van Deurzen** (Department of Pathology, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Rotterdam, the Netherlands), **Cristina Guerrero Paez** (Dutch Breast Cancer Society (BVN), Utrecht, the Netherlands), **Danielle M. de Leeuw** (Department of Surgery, Ziekenhuisgroep Twente, Almelo/Hengelo, the Netherlands), **Djaëlla V. Hooiveld** (Department of Surgery, Zaans Medical Centre, Zaandam, the Netherlands), **Ernst. J.P. Schoenmaeckers** (Department of Surgery, Meander Medical Centre, Amersfoort, the Netherlands), **Els Van Dessel** (Department of Surgery, ZorgSaam hospital, Terneuzen, the Netherlands), **Emily L. Postma** (Department of Surgery, St. Antonius Hospital, Nieuwegein, the Netherlands), **Esther D. van den Ende** (Department of Surgery, Saxenburgh Medical Centre, Hardenberg, the Netherlands), **Gabor S.A. Abis** (Department of Surgery, Meander Medical Centre, Amersfoort, the Netherlands), **Grietje Bouma** (Department of Internal Medicine, Nij Smellinghe Hospital, Drachten, the Netherlands), **Henriette A. Schuttevaer** (Department of Surgery, Martini Hospital, Groningen, the Netherlands), **Hinne A. Rakhorst** (1. Department of Plas-

tic, Reconstructive, and Hand Surgery, Medisch Spectrum Twente, Enschede, The Netherlands. 2. Department of Plastic, Reconstructive, and Hand Surgery, Ziekenhuisgroep Twente, Almelo, The Netherlands), **Ilse Jannink** (Department of Surgery, Hagaziekenhuis, The Hague, the Netherlands), **Ingrid Kappers** (Department of Surgery, Tjongerschans ziekenhuis, Heerenveen, the Netherlands), **Janneke Verloop** (Department of Research and Development, Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, the Netherlands), **Joost Nonner** (Breast Cancer Center South Holland South, Ikazia Hospital, Rotterdam), **Klaartje van Engelen** (Department of Human Genetics, Amsterdam UMC and University of Amsterdam, Amsterdam), **Linda de Munck** (Department of Research and Development, Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, the Netherlands), **Loes F.S. Kooreman** (Department of Pathology, Maastricht University Medical Center, GROW – School for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands), **Marianne R.F. Bosscher** (Department of Surgery, Ommelander Hospital Groningen, Scheemda, the Netherlands), **Marian B.E. Menke-Pluijmers** (Department of Surgery, Albert Schweitzer Hospital, Dordrecht, Netherlands), **Margrethe schlooz-vries (**Department of Surgery, Radboud University Medical Centre, Nijmegen, the Netherlands), **Miriam L. Hoven-Gondrie** (Department of Surgery, Gelderse Vallei, Ede, the Netherlands), **Martinus A. Beek** (Department of Surgery, Isala klinieken, Zwolle, the Netherlands), **Marije J. Hoornweg** (Department of Plastic Surgery, Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands), **Menno H. Raber** (Department of Surgery, Saxenburgh Medical Centre, Hardenberg, the Netherlands), **Milou H. Martens** (Department of Surgery, Laurentius Hospital, Roermond, the Netherlands), **Barbara G. Molenkamp** (Department of Surgery, Diakonessenhuis, Utrecht, the Netherlands), **Patricia Jansen** (Department of Surgery, Elizabeth Tweesteden Hospital, Tilburg, the Netherlands), **Ramon R.J.P. van Eekeren** (Department of Surgery, Rijnstate Hospital, Arnhem, the Netherlands), **Thomas Schok** (Department of Surgery, VieCuri Medical Centre, Venray, the Netherlands), **Robert-Jan Schipper** (Department of Surgery, Catharina Hospital, Eindhoven, the Netherlands), **Titia E. Lans** (Department of Surgery, Admiraal de Ruyterziekenhuis, Goes and Vlissingen, the Netherlands), **Vivianne C.G. Tjan-Heijnen** (Department of Medical Oncology, GROW, Maastricht University Medical Centre, Maastricht, the Netherlands), **Yvonne L. J. Vissers** (Department of Surgery, Zuyderland Medical Centre, Heerlen, the Netherlands).

REFERENCES

1. Beck N, van Bommel AC, Eddes EH, van Leersum NJ, Tollenaar RA, Wouters MW. The Dutch Institute for Clinical Auditing: achieving Codman's dream on a nationwide basis. *Ann Surg.* 2020;271(4):627–31. doi:10.1097/sla.0000000000003665
2. Donabedian A. The quality of care. How can it be assessed? *JAMA.* 1988;260(12):1743–8. doi:10.1001/jama.260.12.1743
3. Vuk T. Quality indicators: a tool for quality monitoring and improvement. *ISBT Sci Ser.* 2012;7(1):24–8. doi:10.1111/j.1751-2824.2012.01584.x
4. Altan D, Ahuja V, Kelleher CM, Chang DC. Look in the mirror, not out the window: in favor of internal benchmarking. *Ann Surg Open.* 2022;3(3):e184. doi:10.1097/as9.0000000000000184
5. Vos EL, Lingsma HF, Jager A, et al. Effect of case-mix and random variation on breast cancer care quality indicators and their rankability. *Value Health.* 2020;23(9):1191–9. doi:10.1016/j.jval.2019.12.014
6. Fischer C, Lingsma HF, van Leersum N, Tollenaar RA, Wouters MW, Steyerberg EW. Comparing colon cancer outcomes: the impact of low hospital case volume and case-mix adjustment. *Eur J Surg Oncol.* 2015;41(8):1045–53. doi:10.1016/j.ejso.2015.04.009
7. van Bommel AC, Spronk PE, Vrancken Peeters MT, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: the national NABON Breast Cancer Audit. *J Surg Oncol.* 2017;115(3):243–9. doi:10.1002/jso.24516
8. Gooiker GA, Veerbeek L, van der Geest LG, et al. [The quality indicator 'tumour positive surgical margin following breast-conserving surgery' does not provide transparent insight into care]. *Ned Tijdschr Geneeskd.* 2010;154:A1142. [Dutch].
9. SONCOS. *Standardisation Report 11: Multi-Disciplinary Standardisation of Oncology Care in the Netherlands.* 2023. Available from: https://demedischspecialist.nl/normeringsrapport-van-soncos
10. Vos EL, Koppert LB, Jager A, Vrancken Peeters M, Siesling S, Lingsma HF. From multiple quality indicators of breast cancer care toward hospital variation of a summary measure. *Value Health.* 2020;23(9):1200–9. doi:10.1016/j.jval.2020.05.011
11. Dutch Institute for Clinical Auditing (DICA). *Annual report 2023.* Available from: https://dica.nl/nieuws/een-terugblik-op-het-jaar-2023/
12. Panhofer P, Ferenc V, Schütz M, et al. Standardization of morbidity assessment in breast cancer surgery using the Clavien Dindo Classification. *Int J Surg.* 2014;12(4):334–9. doi:10.1016/j.ijsu.2014.01.012
13. Quentin W, Partanen VM, Brownwood I. Measuring healthcare quality. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. *Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies.* Vol Health Policy Series. Copenhagen: European Observatory on Health Systems and Policies; 2019.
14. Fischer C, Lingsma H, Hardwick R, Cromwell DA, Steyerberg E, Groene O. Risk adjustment models for short-term outcomes after surgical resection for oesophagogastric cancer. *Br J Surg.* 2016;103(1):105–16. doi:10.1002/bjs.9968
15. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc.* 2002;9(6):600–11. doi:10.1197/jamia.m1087
16. Huijben JA, Wiegers EJ, Ercole A, et al. Quality indicators for patients with traumatic brain injury in European intensive care units: a CENTER-TBI study. *Crit Care.* 2020;24(1):78. doi:10.1186/s13054-020-2791-0

17. Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika.* 1991;78(3):691–2. doi:10.1093/biomet/78.3.691
18. Lingsma HF, Steyerberg EW, Eijkemans MJ, Dippel DW, Scholte Op Reimer WJ, van Houwelingen HC. Comparing and ranking hospitals based on outcome: results from The Netherlands Stroke Survey. *QJM.* 2010;103(2):99–108. doi:10.1093/qjmed/hcp169
19. Schreuder K, Bult TJ, Stroop B, et al. European quality indicators developed by the European Commission Initiative on Breast Cancer: a first nationwide assessment for the Dutch setting. *Breast Cancer Res Treat.* 2024;203(3):523–31. doi:10.1007/s10549-023-07158-w
20. Siotos C, Lagiou P, Cheah MA, et al. Determinants of receiving immediate breast reconstruction: an analysis of patient characteristics at a tertiary care center in the US. *Surg Oncol.* 2020;34:1–6. doi:10.1016/j.suronc.2020.02.017
21. van Bommel AC, de Ligt KM, Schreuder K, et al. The added value of immediate breast reconstruction to health-related quality of life of breast cancer patients. *Eur J Surg Oncol.* 2020;46(10 Pt A):1848–53. doi:10.1016/j.ejso.2020.06.009
22. Verheul EM, van Klaveren D, Lingsma HF, et al. High-impact complications after breast cancer surgery in the Dutch national quality registry: evaluating case-mix adjustment for hospital comparisons. *BJS Open.* 2024;9(1):zrae147. doi:10.1093/bjsopen/zrae147
23. Hofstede SN, Ceyisakar IE, Lingsma HF, Kringos DS, Marang-van de Mheen PJ. Ranking hospitals: do we gain reliability by using composite rather than individual indicators? *BMJ Qual Saf.* 2019;28(2):94–102. doi:10.1136/bmjqs-2017-007669
24. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med.* 2005;24(8):1185–202. doi:10.1002/sim.1970
25. Mayer EK, Bottle A, Rao C, Darzi AW, Athanasiou T. Funnel plots and their emerging application in surgery. *Ann Surg.* 2009;249(3):376–83. doi:10.1097/SLA.0b013e31819a47b1
26. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *JAMA.* 2004;292(7):847–51. doi:10.1001/jama.292.7.847
27. Kaplan ZL, van Leeuwen N, van Klaveren D, et al. Regional disparities in the use of intensive chemotherapy for AML in the Netherlands: does it influence survival? *BMJ Oncol.* 2024;3(1):e000264. doi:10.1136/bmjonc-2023-000264
28. Austin PC, Ceyisakar IE, Steyerberg EW, Lingsma HF, Marang-van de Mheen PJ. Ranking hospital performance based on individual indicators: can we increase reliability by creating composite indicators? *BMC Med Res Methodol.* 2019;19(1):131. doi:10.1186/s12874-019-0769-x
29. Barbazza E, Klazinga NS, Kringos DS. Exploring the actionability of healthcare performance indicators for quality of care: a qualitative analysis of the literature, expert opinion and user experience. *BMJ Qual Saf.* 2021;30(12):1010–20. doi:10.1136/bmjqs-2020-011247
30. Kallen MC, Roos-Blom MJ, Dongelmans DA, et al. Development of actionable quality indicators and an action implementation toolbox for appropriate antibiotic use at intensive care units: a modified-RAND Delphi study. *PLoS One.* 2018;13(11):e0207991. doi:10.1371/journal.pone.0207991
31. Toss MS, Pinder SE, Green AR, et al. Breast conservation in ductal carcinoma in situ (DCIS): what defines optimal margins? *Histopathology.* 2017;70(5):681–92. doi:10.1111/his.13116
32. Eindrapport aandoeningswerkgroep Mammacarcinoom. *Programma Uitkomstgerichte Zorg – Lijn 1 'Meer inzicht in uitkomsten'* [Dutch]. 2023. Available from: https://www.platformuitkomstgerichtezorg.nl
33. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf.* 2011;20(10):869–74. doi:10.1136/bmjqs.2010.048058

34. Rubio IT, Marotti L, Biganzoli L, et al. EUSOMA quality indicators for non-metastatic breast cancer: an update. *Eur J Cancer.* 2024;198:113500. doi:10.1016/j.ejca.2023.113500
35. Maes-Carballo M, Gómez-Fandiño Y, Reinoso-Hermida A, et al. Quality indicators for breast cancer care: a systematic review. *Breast.* 2021;59:221–31. doi:10.1016/j.breast.2021.06.013



Chapter 5

Public Performance Reporting of Oral Cancer Surgery: Detecting Meaningful Between-hospital Variation

Hanneke D. van Oorschot, Elfi M. Verheul, Jose A. Hardillo, Robert J.J. van Es, Guido B. van den Broek, Michelle R. de Graaff, Hester F. Lingsma, Robert Jan Baatenburg de Jong; Dutch Head and Neck Audit – Oral cavity collaborator group.

Submitted

ABSTRACT

Introduction: Public performance reporting (PPR) of quality indicators (QIs) is increasingly used to improve quality of care. Selecting valid and reliable QIs for PPR is challenging, especially in low-incidence diseases. Therefore, a comprehensive evaluation of surgical outcome indicators for PPR in oral cavity cancer (OCC) surgery was performed.

Methods: Our population-cased cohort study included 2316 patients form the Dutch Head and Neck Audit database between 2018 and 2021. Five QIs for OCC surgery were evaluated: unplanned reoperation, surgical complications, resection margins, and textbook outcome after local and extensive surgery. The feasibility (data availability), discriminative ability (inter-quartile range [IQR] of hospital results), validity (case-mix impact in pseudo-R2), reliability (rankability), and caseload requirement for outlier detection (power-analysis) were determined.

Results: Feasibility was 98%-100% for all indicators. Discriminative ability was moderate (IQR 6.9) for unplanned reoperations but good for the other QIs when reported over four years (IQR 11-13). Case-mix impact was highest for resection margins (pseudo-R2 0.24) and minimal for textbook outcomes after extensive resections (pseudo-R2 0.05). Rankability was poor over one year but increased when indicators were reported over multiple years. The required difference from the benchmark for outlier detection decreased when the hospital caseload increased and varied per indicator.

Conclusion: Our findings indicate that one-year reporting for low-incidence diseases such as OCC is not reliable for detecting meaningful between-hospital variation. When feasibility, discriminative ability, and case-mix are considered, rankability and caseload calculations can help other countries and cancer networks determine how many years need to be combined for reliable PPR.

INTRODUCTION

There is an increasing demand for public performance reporting (PPR) of quality indicators (QIs) to stimulate hospital quality improvement.[1] PPR requires that QIs measure what they claim to measure and thus are valid and reliable.[1,2] Observed differences between hospitals can be attributed to case-mix characteristics, statistical uncertainty or chance, registration bias, residual confounding, and differences in quality of care.[3] Data misinterpretation can have critical consequences for all involved parties, especially concerning PPR.[3–7] To enhance statistical robustness, indicators selected for PPR may focus on highly prevalent, composite outcomes or be reported over multiple years.[8] Because the detection of outlier hospitals depends on the number of patients per hospital and events, PPR in low-incidence diseases such as head and neck cancer can be challenging.[4,9]

Since 2014, the Dutch Head and Neck Audit (DHNA) has evaluated the quality of care for head and neck cancer.[10] With the initiation of the DHNA, a set of QIs was defined from the perspectives of medical specialists, allied healthcare professionals, and patients.[11,12] Over the past ten years, the DHNA has adapted or removed indicators that were considered redundant, and new indicators have been developed based on new insights.[13] The DHNA indicators are classified as either public or non-public. The results of public indicators are published openly annually for PPR, whereas the results of non-public indicators are provided exclusively to the respective hospitals for internal quality assessment. In 2024, PPR was applied to four DHNA indicators.[14] The current public surgical indicator in the DHNA is the "percentage of patients undergoing an unplanned reoperation due to complications after surgery", with a national benchmark score ranging from 4.8%-6.4% over the past years.[14] The suitability of this indicator for PPR has been questioned, as the reported hospital differences are influenced mainly by chance due to the low event rate. Other non-public DHNA indicators may be more appropriate, and an evaluation could assist in identifying the most suitable indicator for PPR. Given that surgery is the primary treatment for oral cavity cancer (OCC), this subsite is the focus of the current study. With ~1000 new patients per year, OCC is considered a low-incidence malignancy in the Netherlands.[15]

Therefore, this study aims to comprehensively evaluate outcome indicators for PPR in a low-incidence disease. Moreover, the effect of reporting indicators over one or multiple years will be assessed.

METHODS

Study population

Data from the Dutch oral cavity cancer cohort, a population-based cohort within the DHNA database, was used.[16] Head and neck cancer care in the Netherlands is centralized in fourteen hospitals and DHNA participation is mandatory. The DHNA is one of 26 quality registries the Dutch Institute for Clinical Auditing maintains.[17] Annual data verification processes and standardized registration via data dictionaries guarantee data reliability.[18] All patients with first primary OCC who underwent curative resection between January 1, 2018, and December 31, 2021 were included (ICD-O-3 code C00, C02-C04, C05.0, C5.8-9, and C06.0-8).[19] Dutch law allows the use of electronic health record extractions for research purposes under certain conditions. According to this legislation, neither informed consent from patients nor approval by a medical ethics committee is obligatory for observational studies containing no directly identifiable data (Dutch Civil Law, Article 7:458).

Variables and definitions

Definitions of the denominator (number of cases included for the QI) and numerator (number of cases meeting the criteria of the QI) of QIs were previously defined by the DHNA scientific committee. The following five outcome indicators for OCC surgery were evaluated:

- The percentage of patients who underwent an unplanned reoperation due to surgical complications ≤30 days after surgery.[14] This is the current public indicator.
- The percentage of patients with a complicated postoperative course within ≤30 days after surgery.[20]
- The percentage of patients with a definitive tumour-free resection margin <1 mm after surgery.[21]
- The percentage of patients that met textbook outcome criteria after a local resection.[22]
- The percentage of patients that met textbook outcome criteria after an extensive resection: surgery included pedicled or flee flap reconstruction and/or neck dissection.[22]

Included case-mix variables were sex (male and female), age (<67 and ≥67 years), body mass index (BMI, <25 and ≥25 kg/m2), smoking and alcohol history, World Health Organization (WHO) performance status (0-1 and 2-4), American Society of Anesthesiologists score (ASA, I-II and III-IV), squamous cell carcinoma (SCC versus non-SCC), clinical or pathological T-classification (T1-T2 versus T3-T4), clinical or pathological N-classification (N0 versus N1-N3), subsite in the oral cavity (oral tongue, gum, floor-of-mouth, other subsites), spidery growth, perineural invasion, and vaso-invasive growth. Additional cor-

rection for treatment variables neck dissection and free flap reconstruction was applied for the unplanned reoperation and complicated postoperative course indicator. TNM-classification followed the 8th edition.[23] Elaborate QI and case-mix variable definitions are provided in Table 1 and Appendix A.

Statistical analysis

The methodology for the comprehensive evaluation follows the practical guide by Verheul et al.[24] The evaluation metrics included feasibility, discriminative ability, validity, and rankability. Indicator relevance is not included in these criteria but discussed in the papers of the separate QIs.[20–22] Feasibility was defined as the percentage of data completeness for numerator variables and considered good when >90%, moderate when 70-90%, and poor when <70%. Discriminative ability was defined as the ability to identify differences in hospital performance. This between-hospital variation was reported as a median and interquartile range (IQR) on indicator outcome and categorized as good (>10%), moderate (5-10%), or poor (<5%).

Validity was defined as the proportion of between-hospital differences that can be attributed to case-mix variables using the Nagelkerke's pseudo-R^2.[25] The impact of case-mix variables was categorized as significant (>0.25), moderate (0.10-0.25), or poor (<0.10). An additional analysis was conducted to assess the impact at the hospital level. Both unadjusted and case-mix adjusted Observed/Expected (O/E) ratios for each hospital were plotted. The individual hospital's QI score determined the observed number of events for each ratio. The unadjusted O/E ratio used the mean QI score across all hospitals, while the case-mix adjusted O/E ratio utilized predicted probabilities from the case-mix model for each hospital. The deviation from the diagonal was quantified using the root-mean-squared error (RMSE). A higher RMSE indicates greater hospital shifts after adjustment, reflecting more case-mix impact on hospital level. As a result, more differences between hospitals are explained by case-mix rather than by quality of care.

Rankability is the proportion of variation not due to chance and may be related to the quality of care.[4] It is determined by the ratio between the differences between hospitals (using the estimated heterogeneity from a random logistic effect model) and the uncertainty within hospitals (using the estimated uncertainty, the median standard error (SE) of individual hospitals from the fixed effect logistic regression model for differences within-hospitals). It is expressed by the formula: tau2 / (tau2 + median sigma2)), where tau2 represents the random effect variance (between-hospital variation), and sigma2 the SE variance of the estimated hospital effect (within-hospital variation).[26] The SE, and therefore rankability, is influenced by the number of patients treated at each hospital. A higher SE corresponds to hospitals with smaller patient groups, leading to larger statistical uncertainty. As an example, a rankability of 70% indicates that 70% of the observed differences between hospitals are due to factors other than chance, while the remain-

ing 30% may be attributed to natural variation or random variation (chance). Therefore, rankability reflects the reliability of hospital rankings. It is considered low when <50%, moderate between 50-75%, and high (>75%).[26]

The minimum required hospital volume for outlier detection was determined with a sample size calculation based on the benchmark indicator score calculated over four years. The sample size calculation was based on a one-sample proportion test with 80% power and 0.05 two-sided significance level.[9]

RESULTS

Study population

A total of 2,318 OCC patients could be included (Figure 1). Patients were predominantly male (53%), median age of 67, (former) smokers (70%) and/or drinkers (75%) and had cT1 (34%) or cT2 (32%) cancer stage without metastasis (cN0 81%, Table 2). The tongue was the most affected subsite (44%) and neck dissections were performed in 52%. In 36% of the resections a free or pedicled flap reconstruction was performed. The WHO performance score was the only variable with >10% missing data (23%).

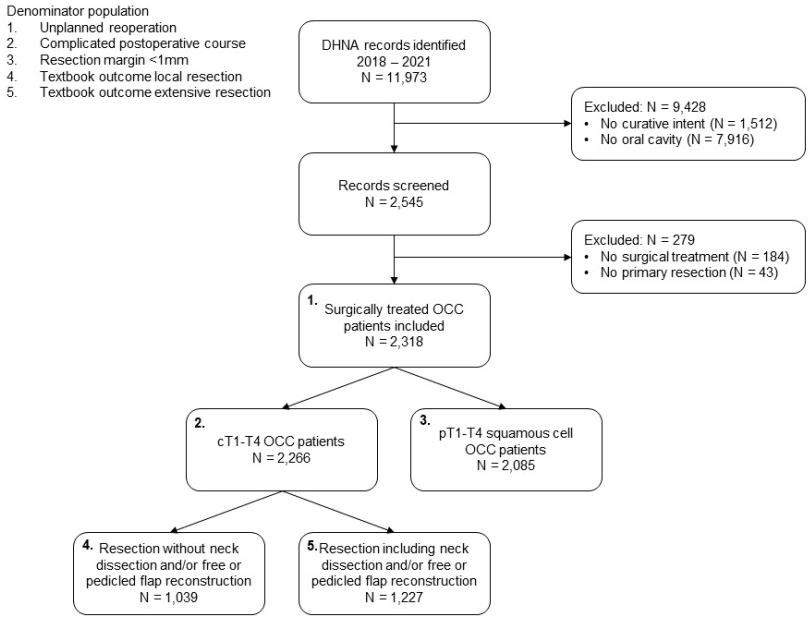


Figure 1: Flow chart for study inclusion. Abbreviations: Dutch Head and Neck Audit (DHNA), OCC = oral cavity cancer.

Table 1: Definitions of denominator and numerator per surgical quality indicator.

1. The percentage of patients who underwent an unplanned reoperation due to surgical complications ≤30 days after surgery.

Denominator	Nominator
Patients surgically treated for oral cavity cancer.	Patients who underwent an unplanned reoperation due to surgical complications ≤30 days after surgery.

2. The percentage of patients with a complicated postoperative course within ≤30 days after surgery. A complicated postoperative course was defined as: 30-day mortality, an unplanned readmission, surgical complications CD grade ≥III or requiring a prolonged hospital stay, or fistula formation.

Denominator	Nominator
Patients surgically treated for cT1-T4 oral cavity cancer.	Patients with a complicated postoperative course, defined as: 30-day mortality, unplanned readmission, surgical complications surgical complications CD grade ≥III or requiring a prolonged hospital stay (>p90), or fistula formation.

3. The percentage of patients with a definitive resection margin <1 mm after surgery for oral squamous cell carcinoma.

Denominator	Nominator
Patients surgically treated for pT1-T4 squamous cell oral cavity cancer	Patients with a resection margin of <1 mm on definitive pathology.

4. The percentage of patients that met textbook outcome criteria after local resection: surgery did not include neck dissection and/or pedicled or flee flap reconstruction. Textbook outcome was defined as the absence of 30-day mortality, hospital readmission, prolonged hospital stay >7 days, surgical complications CD grade ≥III, and a definitive surgical resection margin <1 mm.

Denominator	Nominator
Patients surgically treated for cT1-T4 oral cavity cancer. Surgery did not include uni- or bilateral neck dissection and/or free or pedicled flap reconstruction.	Patients meeting textbook outcome criteria for local resection, defined as: the absence of 30-day mortality, hospital readmission, length of hospital stay >7 days, severe surgical complications (CD grade III or higher), and a definitive surgical resection margin <1 mm.

5. The percentage of patients that met textbook outcome criteria after extensive resection: surgery included neck dissection and/or pedicled or flee flap reconstruction. Textbook outcome was defined as the absence of 30-day mortality, hospital readmission, prolonged hospital stay >21 days, surgical complications CD grade ≥III, a definitive surgical resection margin <1 mm, and a lymph node yield of <18 (only in case of neck dissection).

Denominator	Nominator
Patients surgically treated for cT1-T4 oral cavity cancer. Surgery included uni- or bilateral neck dissection and/or free or pedicled flap reconstruction	Patients meeting textbook outcome criteria for extensive resection, defined as: the absence of 30-day mortality, hospital readmission, length of hospital stay >21 days, severe surgical complications (CD grade III or higher), a definitive surgical resection margin <1 mm, and a lymph node yield of ≥18.

Abbreviations: ASA = American Society of Anesthesiologists, CD = Clavien Dindo, p90 = 90th percentile, WHO = World Health Organization

Table 2: Patient, tumour and treatment characteristics of first primary oral cavity cancer surgery patients diagnosed between 2018 and 2021 in the Netherlands.

Characteristic		*N (%)*
Included patients	Total	2,318 (100%)
Sex	Male	1,236 (53%)
	Female	1,082 (47%)
Age in years	Median (IQR)	67 (59, 75)
	<67 years	1,130 (49%)
	≥67 years	1,188 (51%)
Body mass index kg/m2	Median (IQR)	25.2 (22.6, 28.4)
	<25 kg/m2	1,072 (46%)
	≥25 kg/m2	1,180 (51%)
	Unknown	66 (2.8%)
Smoking history	No history of smoking	598 (26%)
	Former or current smoker	1,618 (70%)
	Unknown	102 (4.4%)
Alcohol history	No history of drinking	382 (16%)
	Former or current drinker	1,728 (75%)
	Unknown	208 (9.0%)
WHO performance status	WHO 0-1	1,606 (69%)
	WHO 2-4	174 (7.5%)
	Unknown	538 (23%)
ASA score	ASA I-II	1,480 (64%)
	ASA III-IV	702 (30%)
	Unknown	136 (5.9%)
Subsite oral cavity	Tongue	1,025 (44%)
	Gum	420 (18%)
	Floor of mouth	397 (17%)
	Other subsites	475 (21%)
Histology type	SCC	2,113 (91%)
	Non-SCC	205 (8.8%)
TNM-classification, cT	cTis/cT0	51 (2.2%)
	cT1	797 (34%)
	cT2	741 (32%)
	cT3	289 (12%)
	cT4	440 (19%)
TNM-classification, cN	cNx	14 (0.6%)
	cN0	1,876 (81%)
	cN1-N2a	182 (7.9%)
	cN2b-N3	246 (11%)

Table 2: Patient, tumour and treatment characteristics of first primary oral cavity cancer surgery patients diagnosed between 2018 and 2021 in the Netherlands. (*continued*)

Characteristic		*N (%)*
Extent of surgery*	Neck dissection	1,205 (52%)
	Free or pedicled flap	826 (36%)
	Local	832 (36%)
	Partial tongue	915 (39%)
	(Sub)total tongue	52 (2.2%)
	Marginal/partial mandibula	310 (13%)
	Mandibula segment	210 (9.1%)
	Partial maxilla	154 (6.6%)
	(Sub)total maxilla	17 (0.7%)
Definitive surgical margin	Margin <1mm	418 (18%)
	Margin 1-5mm	1,002 (43%)
	Margin >5mm	871 (38%)
	Unknown	25 (1.0%)

*IQR = interquartile range, SCC = Squamous cell carcinoma, WHO = World Health Organization / * Sum of percentages can exceed 100% as more than one surgery type can be applicable.*

Data availability and feasibility

The median hospital volume over four years was highest for unplanned reoperations (143 patients) and lowest for textbook outcome after local resection (59 patients). Feasibility was good for all indicators and reporting periods (99%-100%; Table 3).

Discriminative ability

The unplanned reoperation indicator displayed moderate discriminative ability for all reporting periods (IQR 5.3-9.6), except for a good score over 2020 (IQR 11). The discriminative ability for the complicated postoperative course indicator was good for four- and two-year reporting (IQR 11-13) and either good or moderate over the separate years (IQR 8.1-17). For the resection margin <1mm indicator, four-year reporting indicated good discriminative ability (IQR 13) and two- and one-year reporting displayed good or moderate scores (IQR 6.0-16). Both textbook outcome indicators indicated good discriminative ability over all periods (IQR 11-23), except for a moderate score over 2018-2019 for the local resection indicator (IQR 8.4).

Case-mix

For textbook outcome after extensive resection, the proportion of between-hospital differences attributed to case-mix variables was minimal (pseudo-R2 0.05). The case-mix influence was moderate for the other QIs. Ranked from the least to the most influence, these indicators included textbook outcome after local resection (pseudo-R2 0.13), complicated postoperative course (pseudo-R2 0.17), unplanned reoperations (pseudo-R2

0.18), and resection margins <1 mm (pseudo-R2 0.24). At the hospital level, shifts were minor and occurred in both directions (Appendix B), with the most notable impact seen in unplanned reoperations (RMSE 0.104) and resection margins <1mm (RMSE 0.095; Appendix C).

Reliability

Rankability for unplanned reoperations was moderate over four years (62%) and poor or not available (NA) due to limited events for two- and one-year reporting (NA-45). For complicated postoperative course, rankability was good over four years (82%), moderate over two-years (59%-71%), and poor-moderate over one year (21%-54%). Resection margins <1mm indicated moderate rankability for 2018-2021 (64%) and 2020-2021 (52%), but poor rankability for the other periods (7.2%-38%). Rankability for textbook outcome after local resection was poor over all reporting periods (NA-44%). After extensive resections, rankability for textbook outcome was moderate for 2018-2021 (52%) but poor for two- and one-year reporting (NA-39%).

Table 3: Comprehensive evaluation of five outcome indicators for oral cavity cancer surgery.

				Feasibility	Discriminative ability	Case-mix	Reliability
Quality indicator	**Year(s) of data**	**N patients**	**Median N** Hospital range	% data available (hospital range)	Between hospital variation	(Pseudo)	Rankability
			(min – max)		Median (IQR)	R-squared	%
Unplanned reoperation	2018-2021	2318	1423 (61 - 351)	100 (99.1-100)	7.9 (6.9)	0.18	61.9
	2018-2019	1140	73 (28 - 167)	100 (98.1 - 100)	8.1 (6.4)		45.4
	2020-2021	1178	72 (33 - 184)	100 (100 - 100)	6.9 (8.5)		29.5
	2018	564	40 (12 - 90)	100 (100 - 100)	7.6 (9.6)		NA
	2019	576	35 (15 - 78)	100 (97.0 - 100)	6.9 (5.3)		33.4
	2020	585	37 (13 - 91)	100 (100 - 100)	7.4 (11.4)		15.1
	2021	593	36 (18 - 93)	100 (100 - 100)	5.3 (8.1)		17.1
Complicated postoperative course	2018-2021	2266	140 (60-342)	100 (98.2 - 100)	11.8 (11.3)	0.17	81.5
	2018-2019	1114	72 (28 - 164)	100 (96.1 - 100)	13.1 (12.0)		71.3
	2020-2021	1152	71 (32 - 178)	100 (100 - 100)	10.7 (12.9)		58.8
	2018	549	39 (12 - 88)	100 (95.0 - 100)	13.1 (16.4)		44.2
	2019	565	35 (15 - 78)	100 (96.9 - 100)	11.7 (8.1)		53.8
	2020	572	36 (13 - 88)	100 (100 - 100)	9.5 (16.5)		42.5
	2021	580	35 (16 - 90)	100 (100 - 100)	12.9 (8.6)		20.8

Table 3: Comprehensive evaluation of five outcome indicators for oral cavity cancer surgery. (*continued*)

Quality indicator	Year(s) of data	N patients	Feasibility		Discriminative ability	Case-mix	Reliability
			Median N Hospital range	% data available (hospital range)	Between hospital variation	(Pseudo)	Rankability
			(min – max)		Median (IQR)	R-squared	%
Resection margin <1mm	2018-2021	2085	121 (51-324)	100 (100 - 100)	11.9 (12.5)	0.24	64.3
	2018-2019	1026	77 (37 - 177)	100 (100 - 100)	13.4 (6.8)		32.9
	2020-2021	1059	62 (27 - 172)	100 (100 - 100)	10.8 (15.7)		52.2
	2018	511	36 (12 - 83)	100 (100 - 100)	15.9 (6.0)		7.2
	2019	515	33 (12 - 69)	100 (100 - 100)	12.9 (10.8)		38.3
	2020	530	32 (9 - 85)	100 (100 - 100)	11.8 (15.8)		20.3
	2021	529	31 (13 - 87)	100 (100 - 100)	13.2 (11.1)		33.4
Textbook outcome local resection	2018-2021	1039	59 (28-219)	98.9 (89.2 - 100)	78.1 (12.6)	0.13	43.8
	2018-2019	508	30 (12 - 104)	100 (77.8 - 100)	79.5 (8.4)		6.0
	2020-2021	531	30 (16 - 115)	100 (95 - 100)	78.1 (14.4)		41.2
	2018	254	15 (4 - 57)	100 (75 - 100)	78.6 (12.4)		NA
	2019	254	15 (7 - 47)	100 (80 - 100)	82.3 (10.9)		NA
	2020	250	15 (6 - 54)	100 (90 - 100)	81.7 (14.9)		22.6
	2021	281	16.5 (7 - 61)	100 (95 - 100)	78.9 (14.3)		NA
Textbook outcome extensive resection	2018-2021	1227	73 (23-208)	99.1 (94.4 - 100)	47.7 (12.1)	0.05	52.4
	2018-2019	606	37 (13 - 94)	98.6 (87.0 - 100)	44.2 (13.3)		39.3
	2020-2021	621	41 (10 - 114)	100 (90.0 - 100)	49.7 (18.9)		13.4
	2018	295	16 (5 - 49)	100 (88.9 - 100)	49.5 (17.1)		NA
	2019	311	21 (8 - 45)	100 (85.7 - 100)	50.0 (13.9)		26.9
	2020	322	19 (4 - 54)	100 (75.0 - 100)	49.6 (19.4)		NA
	2021	299	19 (3 - 60)	100 (95.6 - 100)	48.6 (22.7)		6.9

Abbreviations: IQR = inter-quartile range, NA = not available due to not enough events

Table 3: legend.

Criterium	Good	Moderate	Poor
Feasibility (%)	>90	70-90	<70
Descriminative ability (IQR)	>10	5-10	<5
Rankability (%)	>75	50-75	<50
Criterium	**Significant**	**Moderate**	**Minimal**
Case-mix (pseudo R2)	≥0.25	0.10-.25	<0.10

Required caseload

The required percentage point difference from the benchmark to identify a hospital as an outlier decreased when the caseload per hospital increased (Table 4). For a caseload of 100 patients, the highest difference was required in textbook outcome after extensive resection (19%), followed by textbook outcome after local resection (18%), complicated postoperative course and resection margins <1 mm (15%), and the lowest in unplanned reoperation (14%).

Table 4: Required percentage point difference from the benchmark to identify a hospital as outlier for a caseload of 10-200 patients for quality indicators in oral cavity cancer surgery.

	Number of patients per hospital				
Quality indicator	**10**	**25**	**50**	**100**	**200**
Unplanned reoperation	35.4	20.6	13.6	9.1	6.2
Complicated postoperative course	38	22.8	15.4	10.4	7.1
Resection margin <1 mm	38.1	22.8	15.4	10.5	7.2
Textbook outcome local resection	41.4	26.1	18.1	12.5	8.7
Textbook outcome extensive resection	39	26.8	19.4	13.9	9.8

DISCUSSION

This nationwide cohort study assessed the eligibility for public performance reporting (PPR) of five surgical oral cavity cancer (OCC) quality indicators (QIs). Our critical evaluation indicated good feasibility and moderate to good discriminative ability and case-mix influence of the five indicators. However, rankability was poor but increased when indicators were reported over multiple years. The sample size calculations indicated that substantial hospital volumes are required to detect outlier hospitals on outcome indicators for OCC. When selecting surgical indicators for PPR in the Dutch Head and Neck Audit (DHNA), reporting over multiple years is necessary to detect meaningful between-hospital variation. Our results underscore the challenges in PPR for low-incidence care and provide guidance on how to decide the number of years that should be combined for a reliable PPR. Moreover, our approach can guide other countries or cancer networks that are implementing PRR for low-incidence diseases.

Each component of this critical evaluation highlights a domain that must be considered when selecting QIs for PPR. Feasibility was good, which is an essential requirement for benchmarking. The DICA registries continually seek processes to maintain and enhance data quality.[27] Discriminative ability was primarily good for 4/5 indicators.. Unplanned reoperations occurred less often and displayed the least variation, complicating outlier detection for this indicator. Before attributing the observed differences to the quality of care, the influence of case-mix variables should be considered. Four out of five indicators

had a moderate pseudo-R^2 based on case-mix assessment. Nevertheless, this does not directly imply evident hospital-level impact. If case-mix variables are distributed evenly across hospitals, no shifts between hospitals are observed when assessing the case-mix at the hospital level. As the RMSE was relatively low for all indicators, it could be considered to report all QIs without case-mix adjustment to relieve the administrative burden of the registry and report data in an understandable form (Appendix B). However, adjusting for case-mix variables facilitates acceptance of the results, enabling constructive discussions on improvement.[7] The rankability increased when reported over multiple years, as was expected from previous studies.[8] The power calculation provided insight into the capacity to identify outliers based on current hospital volumes and outcome prevalence, facilitating the interpretation of observed hospital differences in terms of clinical significance.

Evaluations of rankability often deem indicators unsuitable for ranking hospitals, emphasizing the need for adequate sample sizes. The literature shows that for higher-incidence diseases, rankability can be moderate for one-year reporting in case of sufficient event-rates.[3,28,29] For example, cholecystectomy is a commonly performed surgery, but rankability for bile duct leakage remained poor because this was a rare complication.[3] The Dutch colorectal audit indicated moderate rankability on postoperative complications (rankability 41%-57%, with 27.8% complications), but poor rankability on resection margins (rankability 8.5%, with 2% positive margins), and mortality (0.0%-16.9%, with 2.3% mortality).[28] For the Dutch breast cancer registry, poor rankability was observed for low-incidence outcomes subgroup indicators which reduced the sample size.[26] We have shown that reporting over one year will not achieve moderate rankability for OCC. For multi-year reporting, the balance between sample size and event rates must be considered, and developing subgroup indicators is not advisable.

Our findings provide valuable insights with immediate applicability for PPR in the DHNA. Based on the current event rates, volume requirements, and critical evaluation of surgical OCC QIs, ranking OCC care in Dutch hospitals over one year is not recommended, as this reflects statistical uncertainty instead of differences in quality of care. The caseload calculations were of additional value in evaluating the public unplanned reoperation indicator.[30] Hospitals with the smallest volume over one year require a 21–35% deviation from the benchmark to be classified as outliers, a threshold unlikely to be met given current prevalence rates. The DHNA provides hospitals with dashboard feedback on their indicator outcomes compared to other institutions. Hospital results are presented in funnel plots with confidence intervals, which helps to interpret the statistical uncertainty of this data. While reporting over multiple years enhances robustness, annual reporting offers a shorter feedback cycle. Therefore, it is recommended that public DHNA indicators be reported over multiple years, while non-public indicators remain on a one-year evaluation cycle to enable clinical evaluation. The current PPR of unplanned reoperations over one year might be misleading for stakeholders. Based on the critical evaluation and

sample size calculation, reporting the complicated postoperative course indicator over two or four years provides valid and reliable hospital results for PPR.

Some limitations of the current study should be discussed. The current evaluation only included quantitative measures. Qualitative criteria such as clinical relevance, a precise definition, and registration accuracy are essential but are described in the indicator-specific publications.[11,14,20–22] Missing values in baseline characteristics were present but handled through multiple imputation. This approach is strongly recommended, as conducting a full case analysis would exclude numerous patients in this low-volume population.[31]

Traditionally, case-mix variables are unchangeable characteristics of the hospital population. The applied case-mix adjustment in our study however included treatment characteristics for unplanned reoperations and complicated postoperative course. The textbook outcome QIs were stratified based on treatment variables. The DHNA scientific board motivated this methodology because neck dissections and pedicled or free flaps are individually associated with an increased risk of severe complications.[20] The goal of these indicators was to gain insights into the indicator outcomes irrespective of the type of surgery that was performed. Whether the choice for an extensive resection is protocol-driven or based on the surgeon's preferences remains a pertinent question. All hospitals perform neck dissections when node involvement is suspected clinically. Still, local hospital protocols differ in performing an elective neck dissection or sentinel lymph node biopsy for cN0 patients.[32] Diverse reconstructive strategies are applied in OCC, with no established "gold standard" approach.[33] A DHNA study demonstrated hospital variation in reconstructive surgery in stage I OCC.[34] Operation strategy planning is affected by hospital facility capacities, preference of the surgeon, and patient factors. In the current analysis, identifying outlier hospitals after adjusting for population and treatment factors can initiate initial quality improvement plans. The next step would involve examining differences in surgical strategies to drive further improvements in the DHNA.

PPR is increasingly demanded as evidence supports its positive impact on quality improvement. While the Netherlands has longstanding experience with clinical auditing in national quality registries, international evidence remains limited. This study offers insights for countries or networks still developing such systems. Effect measurements of PPR indicate that hospitals try to improve their outcomes from internal motivation or concern for reputation.[7,35] However, although the high-interest patients have expressed in PPR information, the actual use remains limited.[36] The current reporting format presents challenges for patients, including insufficient awareness, trust, and understanding of the information.[37] It is essential to explore methods for delivering PPR information to patients in a clear, accessible, and meaningful manner in future research.

Conclusion

In conclusion, this study demonstrates that achieving public reporting of is challenging, definitely for low-incidence diseases. Despite the use of a substantial cohort and accepted surgical outcome indicators, no single indicator was identified that was simultaneously feasible, discriminative, valid, and reliable over one year. Meaningful between-hospital variation could only be detected when reporting indicators over multiple years. Low-incidence diseases should focus on more prevalent outcomes to guarantee enough power for outlier detection and identification of best practices. This approach can guide other cancer networks implementing PRR.

Acknowledgements

The authors thank all members of the Dutch head and neck audit (DHNA) oral cavity collaborator group for contributing: R.P. Takes, G.B. Halmos, D.V.C. de Jel, R. Dirven, M. Lacko, L.A.A. Vaassen, J.J. Hendrickx, M.A.E. Oomens, H. Ghaeminia, J.C. Jansen, A. Vesseur, R. Bun, L.Q. Schwandt, C.A. Krabbe, T.J.W. Klein Nulent, A.J.M. van Bemmel, and R. Klijn. The authors thank all registrars, physician assistants and administrative nurses who registered patients in the Dutch Head and Neck Audit. The authors thank the registration team of the Netherlands Comprehensive Cancer Organisation (IKNL) for the collection of data for the Netherlands Cancer Registry and DHNA. This work was supported by the department of Otorhinolaryngology and Head and Neck Surgery of the Erasmus Medical Centre Cancer Institute (Rotterdam, the Netherlands).

REFERENCES

1. Berg M, Meijerink Y, Gras M, et al. Feasibility first: Developing public performance indicators on patient safety and clinical effectiveness for Dutch hospitals. Health Policy. 2005;75(1):59-73. doi:10.1016/j.healthpol.2005.02.007
2. van der Linde M, Salet N, van Leeuwen N, Lingsma HF, Eijkenaar F. Between-hospital variation in indicators of quality of care: a systematic review. BMJ Qual Saf. 2024:1-13. doi:10.1136/bmjqs-2023-016726
3. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. BMJ Qual Saf. 2011;20(10):869-874. doi:10.1136/bmjqs.2010.048058
4. Lingsma HF, Steyerberg EW, Eijkemans MJC, et al. Comparing and ranking hospitals based on outcome: Results from the Netherlands Stroke Survey. QJM. 2009;103(2):99-108. doi:10.1093/qjmed/hcp169
5. Shaw T, Janssen A, Crampton R, et al. Attitudes of health professionals to using routinely collected clinical data for performance feedback and personalised professional development. Med J Aust. 2019;210(S6):S17-S21. doi:10.5694/mja2.50022
6. 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;363:1147-1154. doi:10.1016/S0140-6736(04)15901-1
7. Prang KH, Maritz R, Sabanovic H, Dunt D, Kelaher M. Mechanisms and impact of public reporting on physicians' and hospitals' performance: a systematic review (2000–2020). PLoS One. 2021;16(2):e0247297. doi:10.1371/journal.pone.0247297
8. Hofstede SN, Ceyisakar IE, Lingsma HF, Kringos DS, Marang-van de Mheen PJ. Ranking hospitals: do we gain reliability by using composite rather than individual indicators? BMJ Qual Saf. 2019;28(2):94-102. doi:10.1136/bmjqs-2017-007669
9. de Graaff MR, Hendriks TE, Wouters M, et al. Assessing quality of hepato-pancreato-biliary surgery: nationwide benchmarking. Br J Surg. 2024;111(5):1-9. doi:10.1093/bjs/znae119
10. van Overveld LF, Takes R, Smeele L, Merkx M, Hermens RP; Dutch Head and Neck Audit Group. The Dutch Head and Neck Audit: the first steps. J Head Neck Surg. 2018;1(1):1-8. doi:10.36959/605/528
11. van Overveld LFJ, Braspenning JCC, Hermens RPMG. Quality indicators of integrated care for head and neck cancer. Clin Otolaryngol. 2017;42:322-329. doi:10.1111/coa.12724
12. Donabedian A. The quality of care—how can it be assessed? JAMA. 1988;260(12):1743-1748.
13. van Overveld LFJ, Takes RP, Braspenning JCC, et al. Variation in integrated head and neck cancer care. JNCCN. 2018;16(12):1491-1498. doi:10.6004/jnccn.2018.7061
14. Hoofd-hals chirurgie. Zorginstituut Nederland. https://www.zorginzicht.nl/kwaliteitsinstrumenten/hoofd-hals-chirurgie. Accessed November 5, 2024.
15. van Dijk B, van Oorschot H, Buikema-Grin E, Vermunt M. Kanker van het hoofd-halsgebied in Nederland. IKNL. https://iknl.nl/kanker-van-het-hoofd-halsgebied/mondholte. Published 2024. Accessed October 30, 2024.
16. van Oorschot HD, Hardillo JA, Baatenburg de Jong RJ, et al. Cohort profile: the Dutch oral cavity cancer cohort. BMJ Open. 2025. doi:10.1136/bmjopen-2024-092467
17. Beck N, van Bommel AC, Eddes EH, et al. The Dutch Institute for Clinical Auditing: Achieving Codman's dream on a nationwide basis. Ann Surg. 2020;271:627-631. doi:10.1097/SLA.0000000000003665
18. van der Werf LR, Voeten SC, van Loe CMM, et al. Data verification of nationwide clinical quality registries. BJS Open. 2019;3(6):857-864. doi:10.1002/bjs5.50209

19. Fritz AG, Percy C, Jack A, et al. International Classification of Diseases for Oncology (ICD-O), 3rd ed. 2013.
20. van Oorschot HD, Hardillo JA, van Es RJJ, et al. Surgical complications for oral cavity cancer. Laryngoscope. 2025. doi:10.1002/lary.32033
21. van Oorschot HD, Hardillo JA, van Es RJJ, et al. Hospital-level variation in resection margins after oral cavity cancer surgery. Br J Oral Maxillofac Surg. 2025. doi:10.1016/j.bjoms.2025.06.004
22. van Oorschot HD, de Jel DVC, Hardillo JA, et al. Textbook outcome after oral cancer surgery. Otolaryngol Head Neck Surg. 2025. doi:10.1002/ohn.1337
23. Brierley J, Gospodarowicz M, Wittekind C. TNM Classification of Malignant Tumours, 8th ed. Wiley-Blackwell; 2016.
24. Verheul E, van der Linde M, Lingsma H, et al. Comprehensive evaluation of quality indicators: Dutch Breast Cancer Audit. Int J Health Policy Manag. 2025. In press.
25. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78.
26. Vos EL, Lingsma HF, Jager A, et al. Case-mix and random variation in breast cancer quality indicators. Value Health. 2020;23(9):1191-1199. doi:10.1016/j.jval.2019.12.014
27. Melse PE, Vrolijk JJ, Becherer BE, et al. Data quality assessment of the Dutch Breast Implant Registry. J Plast Reconstr Aesthet Surg. 2024. doi:10.1016/j.bjps.2024.08.051
28. Van Groningen JT, Ceyisakar IE, Gietelink L, et al. Identifying best performing hospitals in colorectal cancer care. Eur J Surg Oncol. 2020;46(6):1144-1150. doi:10.1016/j.ejso.2020.02.024
29. Henneman D, van Bommel ACM, Snijders A, et al. Ranking and rankability of hospital postoperative mortality rates. Ann Surg. 2014;259(5):844-849. doi:10.1097/SLA.0000000000000561
30. Zorginstituut Nederland. Open data MSZ / Ziekenhuizen.
31. van Linschoten RCA, Amini M, van Leeuwen N, et al. Handling missing values in between-hospital comparisons. BMJ Qual Saf. 2023;32(12):742-749. doi:10.1136/bmjqs-2023-016387
32. de Bree R, de Keizer B, Civantos FJ, et al. What is the role of sentinel lymph node biopsy in the management of oral cancer in 2020? Eur Arch Otorhinolaryngol. 2021;278:3181-3191. doi:10.1007/s00405-020-06538-y
33. Katsnelson JY, Tyrell R, Karadsheh MJ, et al. Postoperative complications associated with reconstruction choice in head and neck cancer. J Reconstr Microsurg. 2022;38(5):343-360. doi:10.1055/s-0041-1733922
34. de Jel DVC, van Oorschot HD, Meijer PCA, et al. Nationwide clinical practice variation for reconstructive surgery following oral cavity cancer from the Dutch Head and Neck Audit. Br J Oral Maxillofac Surg. 2024. doi:10.1016/j.bjoms.2024.10.232
35. Hibbard JH. What can we say about the impact of public reporting? Ann Intern Med. 2008;148:161.
36. de Groot IB, Otten W, Smeets HJ, Marang-van de Mheen PJ. Is the impact of hospital performance data greater in patients who have compared hospitals? BMC Health Serv Res. 2011;11:214. doi:10.1186/1472-6963-11-214
37. Faber M, Bosch M, Wollersheim H, Leatherman S, Grol R. Public reporting in health care: how do consumers use quality-of-care information? A systematic review. Med Care. 2009;47(1-8).

PART II

USING OUTCOME PREDICTIONS TO SUPPORT PERSONALISED CARE

6

Chapter 6

External Validation of PREDICT Breast v3.1 for Overall Survival in International Cohorts, Including Young and Invasive Lobular Subgroups

Elfi M. Verheul, Frank Doornkamp, Iurii Petrov, Sabine Siesling, Hester F. Lingsma, Linetta B. Koppert, Lara W.A. Vreven, Adri C. Voogd, Maria Margarete Karsten, Lea Doppelbauer, Pimrapat Gebert, Narsis Kiani, Simona Borstnar, Paul D. P. Pharoah, Elham Hedayati, Ewout W. Steyerberg, David van Klaveren, on behalf of the 4D PICTURE Consortium.

Submitted

ABSTRACT

Purpose: PREDICT Breast is an online tool that provides survival predictions for patients with early-stage breast cancer, for different treatments after surgery. External validation is essential to assess model performance across populations and healthcare settings. We aimed to externally validate PREDICT using clinical practice data from the Netherlands, Sweden, and Slovenia.

Methods: We validated PREDICT in national populations (*Netherlands, N=221,636; Sweden, N=84,928)*, and in two specific subgroups: patients with invasive lobular breast cancer (ILC) *(Netherlands, N=26,834; Sweden, N=10,563; Slovenia, N=341)* and patients aged 40 years or younger *(Netherlands, N=9,995; Sweden, N=2,694)*. We assessed discrimination with the 10-year Area Under the Curve (AUC), and calibration of 10-year mortality predictions through calibration plots, intercepts and slopes.

Results: PREDICT v3.1 discriminated well in the national populations (Netherlands AUC 0.75, 95% CI 0.75–0.76; Sweden 0.75, 95% CI 0.75–0.76), with similar discrimination in ILC patients (Netherlands 0.76, 95% CI 0.74–0.76; Sweden 0.75, 95% CI 0.73–0.77; Slovenia 0.78, 95% CI 0.71-0.83). Calibration showed slight underestimation of mortality risk in the Netherlands (intercept 0.13; slope 1.01), and was near perfect in the Swedish population (intercept – 0.04; slope 1.05). Among ILC patients, we observed some underestimation of mortality (Netherlands intercept 0.20; Sweden intercept 0.10; Slovenia intercept 0.02). In young patients miscalibration was observed (Netherlands, intercept -0.21, slope 0.79; Sweden, intercept -0.08, slope 0.85).

Conclusion: PREDICT v3.1 is generally well calibrated and suitable for clinical use in the evaluated European populations. Efforts to improve PREDICT should focus on more accurate predictions for younger patients.

INTRODUCTION

Shared decision-making (SDM) and personalised care are increasingly recognized as key components of modern cancer care. Prediction models can support SDM[1], and several have been developed for breast cancer (BC)[2 3], with PREDICT being widely used in clinical practice[4].

External validation of prediction models is essential to assess model performance across diverse populations and healthcare settings beyond those used for its development. This is important as patient characteristics, clinical protocols, and treatment practices can vary substantially, both between and within countries, and may also change over time[5]. Validation focuses on discrimination (distinguishing between patients with different outcomes) and calibration (agreement between predicted and observed outcomes) [6]. PREDICT is routinely updated, with the most recent version (3.1, released in 2024) introducing several improvements, including a refitted model based on more recent data (2000–2017), and accounting for both the beneficial effect of radiotherapy on breast cancer mortality and the harmful effects of chemotherapy and radiotherapy on other causes of death[7]. This version has not yet been externally validated in European cohorts beyond the UK.

Histology is not included in v3.1, as it did not demonstrate independent prognostic value during model development, although clinically relevant differences remain between invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). Patients with ILC are typically older, present with more advanced tumours, and show higher ER and lower HER2 expression, with mastectomy more common than breast-conserving surgery [8-12]. More importantly, ER positive ILC is associated with a poorer long-term disease-free and overall survival [9 11] and adjuvant chemotherapy confer less benefit in patients with ILC, although selected patients may still benefit [13 14]. Additionally, prediction models often perform less consistently in young patients [3], including reports of underestimation of mortality by PREDICT in young, node-negative cases [15]. Although women aged 40 years or younger represent a relatively small subgroup, their long life expectancy makes accurate risk prediction essential for guiding treatment. These gaps underscore the need for dedicated validation in both young patients and those with ILC.

We aim to externally validate PREDICT v3.1 using multiple datasets from European cancer registries. We evaluate its performance in Dutch and Swedish national cancer populations and further assess its validity in clinical subgroups where its use may be uncertain: patients with ILC (including Slovenia) and patients aged 40 years and younger.

MATERIAL & METHODS

Data and study population

As part of the 4D project, which aims to enhance data-driven decision-making in oncology [16], this study includes datasets from multiple countries:

- Netherlands Cancer Registry (NCR; Netherlands): Includes all breast cancer patients diagnosed between 2003 – 2022;
- BcBaSe 3.0 (Sweden): research database based on the Swedish National Quality Breast Cancer Register, covering all breast cancer patients in Sweden from 2007 – 2021;
- ILC Database, Institute of Oncology, Ljubljana (Slovenia): Contains all invasive lobular carcinoma (ILC) patients diagnosed between 2003 and 2008 at the Institute of Oncology, Ljubljana.

The inclusion criteria are comparable to those of the PREDICT Breast development set[7 17]. We selected female patients aged between 25 and 85 years with invasive breast cancer, without distant metastasis. Patients who did not undergo surgery, received neo-adjuvant chemotherapy, had tumours larger than 200 mm, or had more than 20 positive lymph nodes were excluded.

For the subgroups, patients with ILC are defined using ICD-O-3.2 code 8520[18]. Patients aged 40 years or younger were defined based on age at the time of diagnosis (biopsy). Given the hypothesised miscalibration in this subgroup, we additionally evaluated calibration across other age groups to assess the miscalibration is specific to younger patients.

Predictors and treatment

The predictors used in the PREDICT algorithm are age at diagnosis, smoking status, ER status, PR status, HER2/ERRB2 status, Ki-67 status (positive defined as more than 10%), invasive tumour size, tumour grade, method of detection (screening or symptomatic), and positive lymph nodes, including micro metastases only when the number of positive nodes is one. While postmenopausal status appears in the online tool to allow selection of bisphosphonate therapy, it is not included in the prognostic model. All predictors were selected from the different databases based on their availability. Since smoking status was not available in any database, we assumed a fixed prevalence of 15% based on a prior Dutch study[19].

Treatment characteristics included in the v3.1 version of the tool are radiotherapy, hormone therapy, chemotherapy, trastuzumab and bisphosphonates. Chemotherapy regimens were categorised as 'standard-dose', referring to anthracycline-based, second-generation regimens such as fluorouracil, epirubicin, cyclophosphamide (FEC), and

'high-dose', referring to third-generation regimens, either high cumulative anthracycline or anthracycline combined with a taxane (e.g. paclitaxel, docetaxel)[20]. If the type of chemotherapy was unknown, patients were assumed to have received high-dose chemotherapy, as this represents the most common practice[21 22]. Furthermore, we assumed that if treatment was given, it was also completed. For radiotherapy, 2 Gray was assigned for left-sided and 0 Gray for right-sided treatment in the Netherlands and Sweden, consistent with the tool's assumptions. For Slovenia, where laterality was missing, 1 Gray was assigned to all patients who underwent radiotherapy. Hormone therapy was assumed to last five years in line with guidelines [23]. As trastuzumab data were unavailable in the Slovenian dataset, we assumed that all HER2-positive tumours received this treatment[24]. Bisphosphonate treatment was set to zero in all databases due to lack of information, which has previously been shown not to significantly impact the results(Table S1)[25].

Outcomes

We focused on 10-year overall survival (OS), considering death from any cause without differentiating underlying causes of mortality. Additional analyses of 5-year outcomes are provided in the supplementary material.

Statistical analysis

Descriptive statistics were reported for each predictor, treatment characteristic, and outcome used in the PREDICT tool for the different cohorts, with means and standard deviations (SD) for continuous variables, and frequencies and percentages for categorical variables.

The PREDICT algorithm v3.1 was obtained from the developer (P.D.P. Pharoah[8]) and used in RStudio to estimate survival probabilities. Variables with missing data that could be selected as 'unknown' by the PREDICT algorithm were treated as such (PR status, HER2 status, Ki-67 status, and mode of detection). Any missing data in the input variables, without the option of 'unknown' in the algorithm (age, ER status, tumour size, tumour grade, positive nodes), were assumed missing at random and imputed using MICE (20 imputed data sets). We used the PredictionTools Package[26], optimized for estimating statistical metrics for survival predictions with imputed data sets, to make calibration plots and estimate discriminative performance metrics.

The discriminative ability of the prediction models was assessed with the AUC (time-dependent area under the ROC curve; AUCt). We chose this AUCt over the C-index, since PREDICT provides estimates on 5 and 10-year survival probabilities rather than survival times[27]. We will hereafter refer to it simply as AUC. Model calibration was visually assessed with calibration plots and numerically assessed with calibration intercepts and slopes. Calibration plots visualize the agreement between predicted and observed outcomes (shown for 10-year mortality, the complement of predicted OS calculated as 1-OS). Ex-

pected survival is plotted on the horizontal axis, and observed survival on the vertical axis. Systematic deviations from the 45-degree line indicate miscalibration and are summarized with a calibration intercept and slope. A calibration intercept below 0 indicates systematic overestimation of risks (predicted risks are too high), while an intercept above 0 indicates underestimation. A slope less than 1 reflects too much spread in predictions, whereas a slope greater than 1 suggests too little. All performance measures were evaluated for the national population (full dataset), as well as for individual sub-cohorts (by country).

We reported the model validations following the TRIPOD checklist (Table S2). For all analyses, we used R statistical software version 4.4.2.

RESULTS

Patient-, tumour- and treatment characteristics

The national study population consisted of 221,636 Dutch and 84,928 Swedish patients (Table 1; Figure S1A,B,C). Subgroups included patients with invasive lobular BC (ILC; Netherlands, N = 26,834; Sweden, N =10,563; Slovenia, N = 341) and those aged 40 years or younger (Netherlands: N = 9,995; Sweden: N = 2,694).

For the national populations, most patient and tumour characteristics were comparable. Differences were observed in screen detection rates, which were higher in Sweden than in the Netherlands (55% vs 42%). More patients received hormone therapy in Sweden (73% vs 51%), but no differences were observed in the proportion of patients with ER positive disease (88%, 87% respectively). In terms of outcome, OS after 10 year was better in Sweden, with 74 956 out of 84,928 patients surviving (88%), compared to 181,152 out of 221,636 patients (82%) in the Netherlands. Almost all patients with ILC had ER positive tumours (Netherlands: 97%; Sweden 98%; Slovenia 96%), resulting in a high proportion receiving hormone therapy. Tumours in young patients (≤ 40 years) were less likely to be ER positive (Netherlands 70% vs 87%; Sweden 71% vs 88%), more likely to be HER2-positive (Netherlands 20% vs 10%, Sweden 23% vs 12%) and more often poorly differentiated (grade 3 Netherlands 52% vs 26%, Sweden 58% vs 30%) compared with the overall national populations, which is consistent with the higher proportion of young patients receiving adjuvant trastuzumab (anti-HER2 therapy) and adjuvant chemotherapy.

Validation of PREDICT Breast v3.1

External validation in the two national populations showed good discrimination at 10 years, with an AUC of 0.75 (95% CI 0.75–0.76) in the Netherlands and 0.75 (95% CI 0.75–0.76) in Sweden (Figure 1). Mortality risk is slightly underestimated in the Netherlands (intercept 0.13; slope 1.01) and shows near perfect calibration in Sweden (intercept

-0.04; slope 1.05). When stratifying by oestrogen receptor (ER) status, underestimation was more pronounced among patients with ER positive tumours in the Netherlands (ER+ intercept 0.18, slope 1.03; ER– intercept –0.05, slope 0.97; Figure S2A), whereas in Sweden patients with ER positive tumours contributed to the near-perfect calibration (ER+ intercept –0.02, slope 1.06; ER– intercept –0.11, slope 0.96; Figure S2B). Calibration plots for 5 year overall mortality showed similar results (Figure S3).

For patients with ILC, discrimination was comparable to the national populations (Netherlands AUC=0.75 with 95% CI 0.74–0.76; Sweden AUC=0.75 with 95% CI 0.73–0.77), and Slovenian data set (AUC=0.78 with 95% CI 0.71-0.83). Calibration plots demonstrated underestimation in the Dutch (intercept 0.20; slope 1.01), Swedish (intercept 0.10; slope 1.09), and Slovenian ILC patients (intercept 0.02; slope 1.03; Figure 2). Calibration plots for 5 years showed similar results (Figure S4).

For patients aged 40 years or younger, PREDICT overestimated mortality risk, particularly in high-risk groups (Netherlands: intercept -0.21, slope 0.79; Sweden: intercept -0.08, slope 0.85; Figure 3, 4), with comparable findings for 5-year overall mortality, but less pronounced (Figure S5A, B). Stratified by ER status, miscalibration among Dutch patients aged 40 years or younger was more pronounced in ER negative compared to ER positive tumours (ER–: intercept –0,51, slope 0.84 ER+: intercept 0.05, slope 0.85; Figure S6A), with comparable findings observed in the Swedish cohort (ER–: intercept –0.51, slope 0.84; ER+: intercept 0.05, slope 0.85; Figure S6B). Additionally, for patients with young triple-negative tumours miscalibration was observed in both cohorts (Netherlands intercept –0.50, slope 1.05; Sweden Intercept -0.19, slope 0.67; Figure S7A,B).

PREDICT was generally well calibrated for patients between 40 and 70 years of age. Above 70, mortality was somewhat underestimated in the Netherlands (intercept 0.23, slope 1.03; Figure 3), and more accurate for patients with ER negative tumours (intercept 0.11, slope 0.97) than for ER positive cases (intercept 0.25, slope 1.04; Figure S6A). In Sweden, for all patients aged >70 years, mortality risks were well calibrated (ER– intercept –0.03, slope 1.12; ER+ intercept –0.01, slope 1.13; Figure S6B). When assessing discrimination within separate age subgroups, the AUC dropped to levels between 0.60 to 0.69 (figure 3,4, Table S3), reflecting the importance of age when predicting mortality.

Table 1: Predictors and treatment characteristics included in PREDICT Breast v3.1 for the validation cohorts from the Netherlands and Slovenia.

	Netherlands (inclusion: 2003-2022)			Sweden (inclusion: 2007-2021)			Slovenia (inclusion: 2003-2008)
	Population N = 221,636	**ILC** N = 26,834	**≤ 40 years** N = 9,996	**Population** N = 84,928	**ILC** N =10,563	**≤ 40 years** N =2,694	**ILC** N = 341
Age at diagnosis (years) *Missing*	61 (12)	63 (11)	36 (4)	62 (12)	65 (11)	36 (3)	62 (12) *1*
Smoker *Missing*	NA*	NA	NA	NA	NA	NA	NA
Post menopausal *Missing*	NA	NA	NA	66,286 (80%) *2,319*	8,846 (86%) *299*	68 (3%) *43*	248 (73%) *1*
ER status, positive *Missing*	186,958 (87%) *6,144*	25,595 (97%) *547*	6,737 (70%) *348*	68,822 (88%) *6,406*	10,363 (98%) *33*	1,766 (71%) *196*	328 (96%)
PR status, positive *Missing*	150,936 (71%) *9,341*	19,914 (77%) *955*	5,915 (62%) *469*	58,095 (74%) *6,498*	8,479 (81%) *40*	1,559 (63%) *200*	277 (81%)
Her2/ERRB2 status, positive *Missing*	20,511 (10%) *24,876*	771 (3%) *2,717*	1,719 (20%) *1,374*	9,118 (12%) *9,204*	437 (4%) *366*	566 (23%) *262*	19 (6%) *2*
Ki-67 status, positive *Missing*	NA	NA	NA	40,240 (76%) *32,275*	4,738 (65%) *3,278*	1,410 (91%) *1,138*	81 (24%)
Invasive tumour size (mm) *Missing*	18 (13) *8,589*	25 (19) *1,159*	20 (13) *485*	18 (15)	26 (20)	20 (17)	25 (20) *3*
Tumour grade *Missing*	1 = 55,592 (26%) 2 = 102,603 (48%) 3 = 54,202 (26%) *9,239*	1 = 4,101 (16%) 2 = 18,936 (76%) 3 = 2,062 (8%) *1,735*	1 = 1,201 (13%) 2 = 3,417 (36%) 3 = 4,932 (52%) *446*	1 = 16,670 (20%) 2 = 41,187 (50%) 3 = 23,928 (30%) *3,143*	1 = 744 (7%) 2 = 8,888 (85%) 3 = 821 (8%) *110*	1 = 223 (9%) 2 = 847 (33%) 3 = 1,461 (58%) *163*	1 = 51 (15%) 2 = 254 (74%) 3 = 36 (11%)
Detected by screening *Missing*	59,665 (42%) *80,385*	7,009 (39%) *8,910*	37 (<1%) *4,714*	46,776 (55%) *258*	5,456 (52%) *27*	87 (3%)	NA

Table 1: Predictors and treatment characteristics included in PREDICT Breast v3.1 for the validation cohorts from the Netherlands and Slovenia. (*continued*)

	Netherlands *(inclusion: 2003-2022)*			**Sweden** *(inclusion: 2007-2021)*			**Slovenia** *(inclusion: 2003-2008)*
Positive nodes, yes *Missing*	66,927 (31%) *4,544*	8,842 (34%) *474*	4,101 (41%) *65*	22,952 (27%)	3,275 (31%)	1,006 (37%)	132 (39%)
Micrometastases only *Missing*	NA	NA	NA	5,046 (6%)	648 (6%)	215 (8%)	30 (9%)
Radiotherapy *Missing*	151,169 (68%)	16,975 (63%)	6,265 (63%)	49,626 (72%) *16,183*	6,263 (71%) *1,769*	1,381 (67%) *642*	150 (44%)
Adjuvant hormonal therapy *Missing*	111,982 (51%)	18,192 (68%)	5,295 (53%)	50,283 (73%) *16,191*	8,145(93%) *1,770*	1,327 (65%) *638*	322 (94%)
Adjuvant trastuzumab *Missing*	12,600 (6%)	437 (2%)	1,469 (15%)	5,977 (7%)	262 (3%)	392 (15%)	1 (<1%)
Adjuvant chemotherapy** *Missing*	No = 155,787 (70%) Standard dose = 9,136 (4.1%) High dose = 41,799 (19%) Type unknown = 14,914 (7%)	No = 20,078 (75%) Standard dose = 955 (4%) High dose = 4,331 (16%) Type unknown = 1,470 (5%)	No = 2,477 (25%) Standard dose = 1,060 (11%) High dose = 4,488 (45%) Type unknown = 1,971 (20%)	22,869 (33%) *16,203*	2,404(27%) *1,771*	1,559 (76%) *640*	No = 263 (77 %) Standard dose = 66 (20%) High dose = 12 (4%)
Survival (5y)***	200,990 (91%)	24,342 (91%)	9,344 (93%)	79,187 (93%)	9,804 (93%)	2,539 (94%)	303 (89%)
Survival (10y)	181,152 (82%)	21,630 (81%)	8,817 (88%)	74 956 (88%)	9,140 (87%)	3,313 (91%)	253 (74%)

Values are n (%), unless for age and invasive tumour size, which are presented as mean (standard deviation); ILC = Invasive Lobular breast Cancer patients; NA = not present in dataset; *81% missing, excluded from predictions; **detailed information about the type of adjuvant chemotherapy was not available for the Swedish cohort; ***results of 5-years analysis included in the supplementary material.

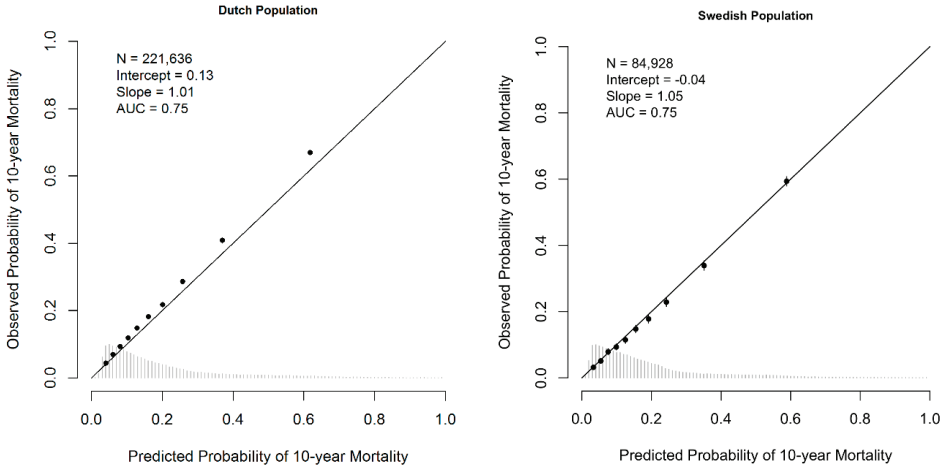
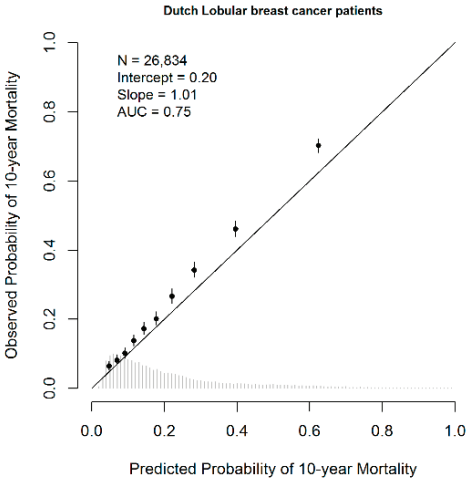
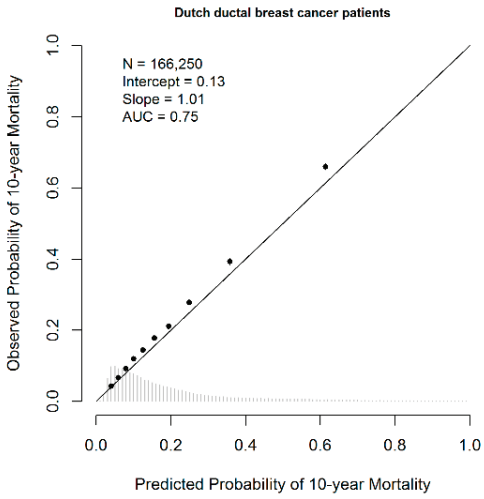
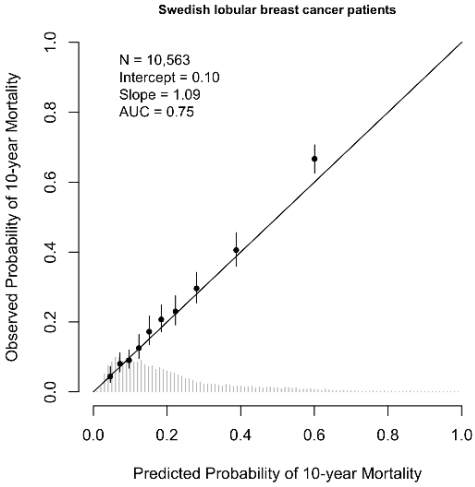


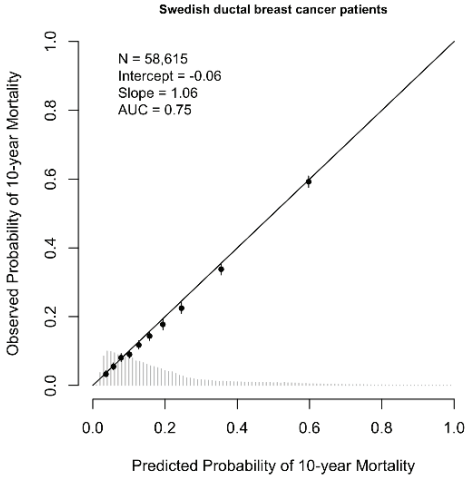
Figure 1: Calibration plots for the Dutch (left) and Swedish (right) populations, showing 10-year overall mortality predictions using PREDICT Breast v3.1.

N = Number of patients included in the external validation set; AUC = Area under the curve at 10 years. The intercept and slope (ideally 0 and 1, respectively) are measures used to evaluate calibration. The Y-axis represents the observed 10-year mortality, and the X-axis represents the predicted 10-year mortality. Each dot indicates the average for a decile of predicted probabilities, with accompanying confidence intervals (95% CI; due to the large sample size, this is negligible in this figure).


Dutch Lobular breast cancer patients
N = 26,834
Intercept = 0.20
Slope = 1.01
AUC = 0.75
Observed Probability of 10-year Mortality
Predicted Probability of 10-year Mortality
0.0
0.2
0.4
0.6
0.8
1.0


Dutch ductal breast cancer patients
N = 166,250
Intercept = 0.13
Slope = 1.01
AUC = 0.75
Observed Probability of 10-year Mortality
Predicted Probability of 10-year Mortality
0.0
0.2
0.4
0.6
0.8
1.0


Swedish lobular breast cancer patients
N = 10,563
Intercept = 0.10
Slope = 1.09
AUC = 0.75
Observed Probability of 10-year Mortality
Predicted Probability of 10-year Mortality
0.0
0.2
0.4
0.6
0.8
1.0


Swedish ductal breast cancer patients
N = 58,615
Intercept = -0.06
Slope = 1.06
AUC = 0.75
Observed Probability of 10-year Mortality
Predicted Probability of 10-year Mortality
0.0
0.2
0.4
0.6
0.8
1.0

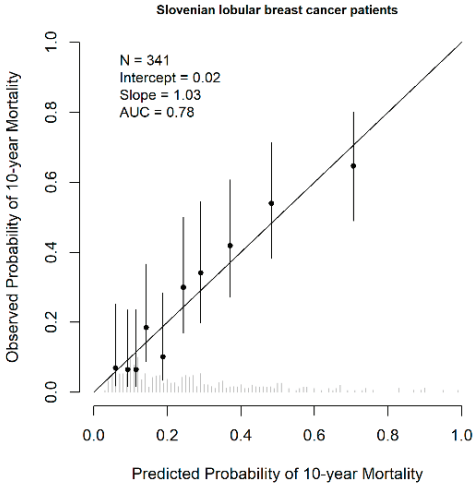

Slovenian lobular breast cancer patients
N = 341
Intercept = 0.02
Slope = 1.03
AUC = 0.78
Observed Probability of 10-year Mortality
Predicted Probability of 10-year Mortality
0.0
0.2
0.4
0.6
0.8
1.0

Figure 2: Calibration plots for Dutch (left), Swedish (middle), and Slovenian (right) patients with invasive lobular breast cancer (top row) and invasive ductal breast cancer (bottom row*), showing 10-year overall mortality predictions using PREDICT Breast v3.1.

*N = Number of patients included in the external validation set; AUC = Area under the curve at 10 years. The intercept and slope (ideally 0 and 1, respectively) are measures used to evaluate calibration. The Y-axis represents the observed 10-year mortality, and the X-axis represents the predicted 10-year mortality. Each dot indicates the average for a decile of predicted probabilities, with accompanying 95% confidence intervals. * For the Slovenian cohort, we only had access to patients with invasive lobular breast cancer, not those with ductal breast cancer.*

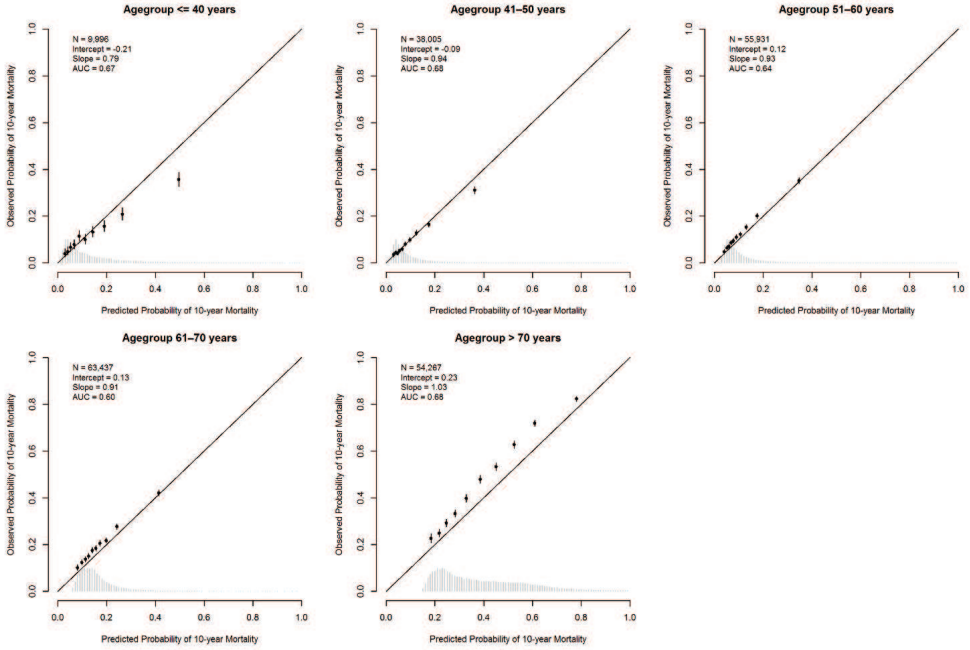


Figure 3: Calibration plots stratified by age group for 10-year overall mortality in Dutch breast cancer patients using PREDICT Breast v3.1.

N = Number of patients included in the external validation set; AUC = Area under the curve at 10 years. The intercept and slope (ideally 0 and 1, respectively) are measures used to evaluate calibration. The Y-axis represents the observed 10-year mortality, and the X-axis represents the predicted 10-year mortality. Each dot indicates the average for a decile of predicted probabilities, with accompanying 95% confidence intervals.

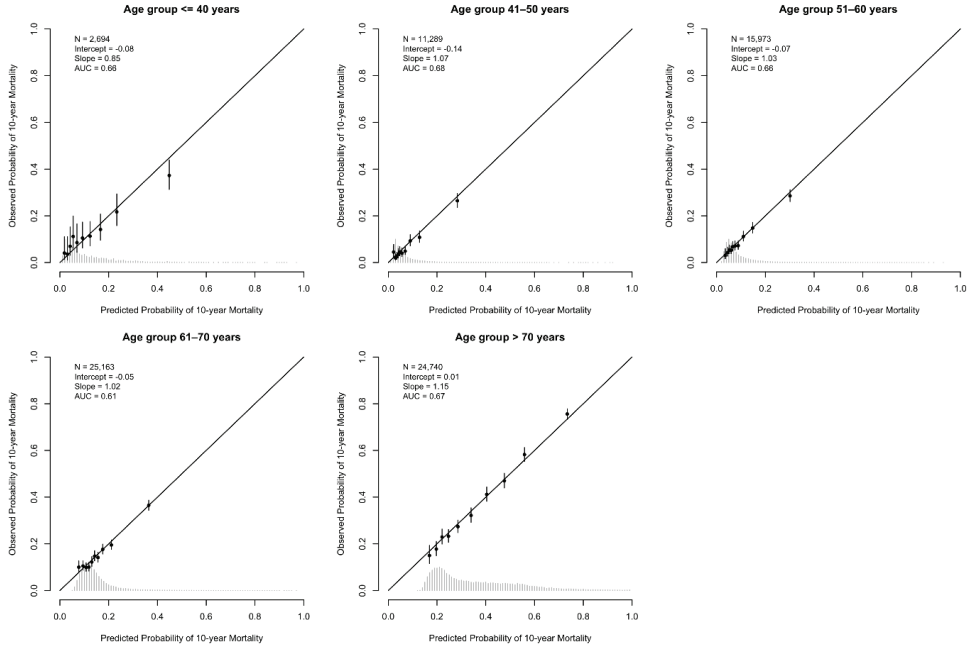


Figure 4: Calibration plots stratified by age group for 10-year overall mortality in Swedish breast cancer patients using PREDICT Breast v3.1.

N = Number of patients included in the external validation set; AUC = Area under the curve at 10 years. The intercept and slope (ideally 0 and 1, respectively) are measures used to evaluate calibration. The Y-axis represents the observed 10-year mortality, and the X-axis represents the predicted 10-year mortality. Each dot indicates the average for a decile of predicted probabilities, with accompanying 95% confidence intervals.

DISCUSSION

We validated the PREDICT Breast v3.1 tool in national cohorts from the Netherlands and Sweden, and in subgroups of patients with invasive lobular carcinoma (ILC) and those aged 40 years or younger. The model showed excellent calibration of mortality risks in the overall Swedish national cohort and slight underestimation in the overall Dutch national cohort. For ILC, mortality was consistently underestimated in the Dutch Swedish and Slovenian cohorts, although overall calibration remained acceptable. In patients aged 40 years or younger, the model was somewhat miscalibrated and overestimated mortality risk in high-risk groups. Overall, these findings support its suitability for clinical use, where PREDICT supports shared-decision making by providing estimates of both survival outcomes and expected treatment benefit. However, further improvements are recommended for the ILC and young patient subgroups.

National populations

PREDICT v3.1 demonstrated good calibration and discrimination in the national cohorts, consistent with recent external validation studies[28]. However, we observed a slight underestimation of mortality in the Netherlands, compared with a near perfect calibration in Sweden. Although hormone therapy is incorporated in PREDICT, differences in treatment implementation, adherence, or recording between countries may influence the correspondence between predicted and observed outcomes and could partly contribute to the calibration differences. Hormone therapy use was higher in Sweden (73%) than in the Netherlands (51%). In the UK development dataset, use ranged from 40–60% depending on the region[7]. As hormone therapy uptake in the Netherlands falls within this range, differences in uptake alone are unlikely to fully explain the observed calibration pattern. Background mortality may also contribute, as life expectancy is higher in Sweden[29], but further analyses focusing on breast cancer-specific survival and background mortality are needed to test this hypothesis.

The most notable difference between the national cohorts was the higher proportion of screen-detected cases in Sweden. This likely reflects national screening policies, with screening starting at 40 years of age in Sweden compared to 50 in the Netherlands. Despite this, the age-at-diagnosis distribution did not indicate earlier diagnoses in Sweden. However, since screen detection is already included as a predictor in the model, such differences are unlikely to fully explain the observed calibration discrepancy between the two populations.

ILC patients

Although questions have been raised about the applicability of the PREDICT tool for patients with ILC and no previous validation studies specifically focus on this subgroup, we observed reasonable discrimination and calibration, supporting its use for 5- and 10-year predictions. A possible explanation for the better-than-expected calibration is that previously observed differences in chemotherapy benefit between invasive ductal carcinoma (IDC) and ILC are largely attributable to the higher prevalence of ER-positive and HER2-negative tumours in ILC, rather than histological type itself 30. As ER and HER2 status are included in the model, these underlying factors are already accounted for in the predictions. Nevertheless, calibration could be further improved, as underestimation of mortality is still observed, most pronounced in the Netherlands (intercept 0.20).

Young breast cancer patients

For patients aged 40 years and younger, PREDICT v3.1 shows miscalibration, particularly by overestimating mortality risk for young patients at high risk. This trend is seen in both countries and is more pronounced in patients with ER negative and triple negative tumours. Similar findings have been reported in other external validation studies of earlier PREDICT versions[15 31]. This miscalibration appears to be age group specific and may be

related to the fact that, in younger patients, mortality is more likely to be predominantly breast cancer related, whereas in older patients, background mortality plays a larger role. It may also reflect their limited representation in the development cohorts, as well as biological differences in tumour behaviour and treatment response. Further research, including competing risk modelling, is needed to refine predictions. Yet accurate prediction is particularly important for younger patients, for whom life expectancy and the long-term consequences of treatment are crucial considerations in shared decision-making.

Evolution of treatment recommendations and future research

When PREDICT was developed, all patients who had received neoadjuvant therapy were excluded [17]. However, the increasing use of neoadjuvant chemotherapy has led questions regarding adjuvant therapy recommendations in this group. In the current online version of the tool (PREDICT v3.1), patients are advised to input tumour size measured prior to neoadjuvant chemotherapy. This variable, however, was not available in our dataset, making validation in this context impossible. In patients receiving neoadjuvant chemotherapy, using pathological tumour size confirmed the expected underestimation and underscored the need for validation using pre-treatment size (Figure S8). In addition, more treatment strategies have evolved over time, with changes in chemotherapy regimens and the introduction of novel therapies such as CDK4/6 inhibitors[32 33]. These advancements are not yet reflected in PREDICT, underscoring the importance of continuously updating the model to remain relevant for modern clinical practice. Future research should focus on improving model accuracy for specific subgroups and maintaining alignment with evolving treatment options and diverse patient populations. Priorities include validating in patients receiving neoadjuvant therapy, improving modelling of age effects and better representation of underrepresented groups in development cohorts. Genomic markers like Mammaprint have modestly improved 5-year breast cancer mortality predictions[25], and future studies should evaluate their impact on long-term outcomes. Finally, while PREDICT provides survival estimates based on average treatment effects of adjuvant therapies, generally considered constant on the relative risk scale across subgroups[34 35], future research could explore potential interactions with specific variables[36].

Strengths and Limitations

A major strength of this study is the use of large, population-based registries from the Netherlands and Sweden, covering multiple years of data and including all surgically treated breast cancer patients in the respective countries. Another key strength is the focus on ILC, a subgroup for which the applicability of current prediction tools has not been well established. Moreover, data from three countries could be used, although the Slovenian cohort consists of previously collected data from an earlier treatment period, which may limit representativeness for current clinical practice.

Certain limitations should be acknowledged. In this study, no distinction could be made between breast cancer specific mortality and overall mortality, unlike in the PREDICT tool, making it difficult to determine whether performances reflect breast cancer specific or background mortality. Additionally, smoking was not available, and therefore a fixed smoking prevalence of 15% based on Dutch population data was assumed. Having smoking information available would change the absolute individual mortality estimates, but the impact on overall discrimination would only be marginal[37]. Furthermore, data on 15-year survival outcomes are lacking. Such long-term information is especially relevant for patients with ILC, as deviations from average population predictions tend to become more apparent over time[8], and for younger patients, given their longer life expectancy. Finally, input data were not entirely harmonised between the two national cohorts; for example, Ki-67 was only available in the Swedish dataset and its assessment is known to be subject to interobserver variability[38]. While sensitivity analyses suggested that treating Ki-67 as unknown had minor impact on performance, differences in data availability should be considered when interpreting results across both populations.

Conclusion

PREDICT Breast v3.1 is generally well calibrated for the evaluated European national populations. Although PREDICT demonstrated reasonable performance for lobular breast cancer patients, calibration could be further improved. For patients aged 40 and younger, an overestimation of mortality risk in high-risk patients was observed. Improved modelling of age effects may enhance the overall predictive performance of the tool for younger patients.

Acknowledgements

The authors thank all registrar and healthcare professionals for data registration.

4D PICTURE Consortium:

Participants of the 4D picture consortium: **Kasper Frank** (Center for Shared Decision Making, Vejle/Lillebaelt University Hospital of Southern Denmark), **Yassin Engelberts** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands), **Lára R. Hallsson** (Institute of Public Health, Medical Decision Making and Health Technology Assessment; Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL – University for Health Sciences and Technology, Hall in Tirol, Austria), **Toby Hackmann** (Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Michel W. J. M. Wouters** (Medical Decision Making, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Yufeng Liu** (Department of Linguistics and English Language, Lancaster University, Lancaster, United Kingdom), **Uwe Siebert** (Institute of Public Health, Medical Decision Making and Health Technology Assessment; Department of Public Health, Health Services

Research and Health Technology Assessment, UMIT TIROL – University for Health Sciences and Technology, Hall in Tirol, Austria; Center for Health Decision Science, Departments of Epidemiology & Health Policy and Management, Harvard T.H. Chan School of Public Health; Institute for Technology Assessment and Department of Radiology, Harvard Medical School, Boston, MA, USA), **Iñaki Álvarez Busto** (Medical Oncology, Instituto de Investigación Sanitaria de Aragón (IIS-A); Hospital Universitario Miguel Servet, Zaragoza, Spain; Department of Medical Oncology, Miguel Servet University Hospital, Facultad de Medicina, Universidad de Zaragoza, Instituto Investigación Sanitaria Aragón, Zaragoza, Spain), **Tamara Wit** (Medical Decision Making, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Lennart Koch** (Institute of Public Health, Medical Decision Making and Health Technology Assessment; Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL – University for Health Sciences and Technology, Hall in Tirol, Austria), **Dirk Snelders** (Department of Design, Organisation and Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, the Netherlands), **Anne Stiggelbout** (Medical Decision Making, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Ida Korfage** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands), **Carlos Romero-Piqueras** (Fractal Strategy, Zaragoza, Spain), **Judith Rietjens** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands; Department of Design, Organisation and Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, the Netherlands), **Liesbeth C. de Wreede** (Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Daisy Monika Lal** (School of Computing and Communication, Lancaster University, United Kingdom), **Kathrina Dankl** (Lab for Social Design, Design School Kolding, Kolding, Denmark), **Claudia Cruz Oliveira** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands), **Amara Callistus Nwosu** (International Observatory on End of Life Care, Lancaster University, Lancaster, United Kingdom; Department of Palliative Care, Lancaster Medical School, Lancaster, United Kingdom; Marie Curie North West, Liverpool, United Kingdom), **Teresa Puértolas** (Department of Medical Oncology, Miguel Servet University Hospital, Facultad de Medicina, Universidad de Zaragoza, Instituto Investigación Sanitaria Aragón, Zaragoza, Spain).

REFERENCES

1. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res. 2018;2:11.
2. Hueting TA, van Maaren MC, Hendriks MP, Koffijberg H, Siesling S. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. J Clin Epidemiol. 2022;152:238-47.
3. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. BMC Cancer. 2019;19(1):230.
4. Candido Dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res. 2017;19(1):58.
5. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med. 2023;21(1):70.
6. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
7. Grootes I, Wishart GC, Pharoah PDP. An updated PREDICT breast cancer prognostic model including the benefits and harms of radiotherapy. NPJ Breast Cancer. 2024;10(1):6.
8. Oesterreich S, Nasrazadani A, Zou J, Carleton N, Onger T, Wright MD, et al. Clinicopathological Features and Outcomes Comparing Patients With Invasive Ductal and Lobular Breast Cancer. J Natl Cancer Inst. 2022;114(11):1511-22.
9. Chen Z, Yang J, Li S, Lv M, Shen Y, Wang B, et al. Invasive lobular carcinoma of the breast: A special histological type compared with invasive ductal carcinoma. PLoS One. 2017;12(9):e0182397.
10. Delpech Y, Coutant C, Hsu L, Barranger E, Iwamoto T, Barcenas CH, et al. Clinical benefit from neoadjuvant chemotherapy in oestrogen receptor-positive invasive ductal and lobular carcinomas. Br J Cancer. 2013;108(2):285-91.
11. Pestalozzi BC, Zahrieh D, Mallon E, Gusterson BA, Price KN, Gelber RD, et al. Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 International Breast Cancer Study Group clinical trials. J Clin Oncol. 2008;26(18):3006-14.
12. Truin W, Roumen RM, Siesling S, van der Heiden-van der Loo M, Duijm LE, Tjan-Heijnen VC, Voogd AC. Patients with invasive lobular breast cancer are less likely to undergo breast-conserving surgery: a population based study in the Netherlands. Ann Surg Oncol. 2015;22(5):1471-8.
13. Truin W, Voogd AC, Vreugdenhil G, van der Heiden-van der Loo M, Siesling S, Roumen RM. Effect of adjuvant chemotherapy in postmenopausal patients with invasive ductal versus lobular breast cancer. Ann Oncol. 2012;23(11):2859-65.
14. Corso G, Shen S, Criscitiello C, Mukhtar R, Gamble L, Rocco EG, et al. Invasive lobular carcinoma: Strategies and perspectives from the lobular breast cancer research group. Cancer Treat Rev. 2025;140:103001.
15. Wang Y, Broeks A, Giardiello D, Hauptmann M, Jozwiak K, Koop EA, et al. External validation and clinical utility assessment of PREDICT breast cancer prognostic model in young, systemic treatment-naive women with node-negative breast cancer. Eur J Cancer. 2023;195:113401.
16. Rietjens JAC, Griffioen I, Sierra-Perez J, Sroczynski G, Siebert U, Buyx A, et al. Improving shared decision-making about cancer treatment through design-based data-driven decision-support tools and redesigning care paths: an overview of the 4D PICTURE project. Palliat Care Soc Pract. 2024;18:26323524231225249.

17. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res. 2010;12(1):R1.
18. (WHO) W. International classification of diseases for oncology (ICD-O) – 3rd edition. 2013.
19. Verheul EM, van Klaveren D, Lingsma HF, Vos E, Hoornweg MJ, Siesling S, Koppert LB; NBCA consortium. High-impact complications after breast cancer surgery in the Dutch national quality registry: evaluating case-mix adjustment for hospital comparisons. BJS Open. 2024 Dec 30;9(1):zrae147. doi: 10.1093/bjsopen/zrae147. PMID: 39903734; PMCID: PMC11793075.
20. Early Breast Cancer Trialists' Collaborative G, Peto R, Davies C, Godwin J, Gray R, Pan HC, et al. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. Lancet. 2012;379(9814):432-44.
21. Federation of Medical Specialists. (2022). Guideline on breast cancer – Systemic treatment [Borstkanker – Systemische therapie]. Richtlijnendatabase. Retrieved september 29, 2025, from https://richtlijnendatabase.nl/richtlijn/borstkanker/behandeling_invasief_carcinoom/adjuvante_systemische_behandeling/systemische_therapie.html
22. Regionala cancercentrum i samverkan. (2022). National breast cancer care guideline [Nationellt vårdprogram för bröstcancer]. Stockholm: RCC. Retrieved september 29, 2025, from https://kunskapsbanken.cancercentrum.se/globalassets/cancerdiagnoser/brost/vardprogram/nationellt-vardprogram-brostcancer.pdf [
23. Federation of Medical Specialists. (2022). Guideline on breast cancer – endocrine therapy [Borstkanker – Endocriene therapie]. Richtlijnendatabase. Retrieved July 30, 2025, from https://richtlijnendatabase.nl/richtlijn/borstkanker/behandeling_invasief_carcinoom/adjuvante_systemische_behandeling/endocriene_therapie.html
24. Loibl S, Andre F, Bachelot T, Barrios CH, Bergh J, Burstein HJ, et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol. 2024;35(2):159-82.
25. Engelhardt EG, Binuya MAE, Pharoah PDP, Poncet C, Rutgers EJT, Piccart M, et al. Prognostication and treatment predictions for estrogen receptor positive early-stage breast cancer: incorporating the 70-gene signature into the PREDICT prognostication model. Breast. 2025;83:104542.
26. Maas, C. H. M. (2025). PredictionTools: R scripts. GitHub. https://github.com/CHMMaas/PredictionTools/tree/main/R
27. McLernon DJ, Giardiello D, Van Calster B, Wynants L, van Geloven N, van Smeden M, et al. Assessing Performance and Clinical Usefulness in Prediction Models With Survival Outcomes: Practical Guidance for Cox Proportional Hazards Models. Ann Intern Med. 2023;176(1):105-14.
28. Hsiao YW, Wishart GC, Pharoah PDP, Peng PC. Assessing the Performance of the PREDICT Breast Version 3.0 Prognostic Tool in Patients With Breast Cancer in the United States. J Natl Compr Canc Netw. 2025;23(6):227-33.
29. Statistics Netherlands (CBS). European Scale 2019: Life Expectancy. CBS Longreads. Available from: https://longreads.cbs.nl/european-scale-2019/life-expectancy/ [Accessed 29 October 2025].
30. Lips EH, Mukhtar RA, Yau C, de Ronde JJ, Livasy C, Carey LA, et al. Lobular histology and response to neoadjuvant chemotherapy in invasive breast cancer. Breast Cancer Res Treat. 2012;136(1):35-43.

31. Gray E, Marti J, Brewster DH, Wyatt JC, Hall PS, Group SA. Independent validation of the PREDICT breast cancer prognosis prediction tool in 45,789 patients using Scottish Cancer Registry data. Br J Cancer. 2018;119(7):808-14.
32. Hussain M, Brezden-Masley C, Chia S, Curigliano G, Webster M, Henning JW. Clinician's guide: expert insights on the use of CDK4/6 inhibitors in patients with early breast cancer. Ther Adv Med Oncol. 2025;17:17588359251326710.
33. Parikh PM, Vora A, Yadav R, Kapoor A, Sahoo T, Rajappa S, et al. Consensus Guidelines for the Use of Cyclin-Dependent Kinase (CDK) 4/6 Inhibitors in the Management of Hormone Receptor Positive (HR+ve), Her2-ve Early Breast Cancer (EBC). South Asian J Cancer. 2025;14(1):45-52.
34. Early Breast Cancer Trialists' Collaborative Group . Electronic address bocoau, Early Breast Cancer Trialists' Collaborative G. Anthracycline-containing and taxane-containing chemotherapy for early-stage operable breast cancer: a patient-level meta-analysis of 100 000 women from 86 randomised trials. Lancet. 2023;401(10384):1277-92.
35. Early Breast Cancer Trialists' Collaborative G. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet. 2005;365(9472):1687-717.
36. Dennis JM, Young KG, Cardoso P, Gudemann LM, McGovern AP, Farmer A, et al. A five-drug class model using routinely available clinical features to optimise prescribing in type 2 diabetes: a prediction model development and validation study. Lancet. 2025;405(10480):701-14.
37. Pencina MJ, D'Agostino RB, Massaro JM. Understanding increments in model performance metrics. Lifetime Data Anal. 2013;19(2):202-18.
38. Focke CM, Burger H, van Diest PJ, Finsterbusch K, Glaser D, Korsching E, et al. Interlaboratory variability of Ki67 staining in breast cancer. Eur J Cancer. 2017;84:219-27.

7

Chapter 7

Comparative Validation of PREDICT Versions 3.1 and 2.2 for Overall Survival in the Dutch Breast Cancer Population

Lara W.A. Vreven, Elfi M. Verheul, Marissa C. van Maaren, Frank Doornkamp, Robert-Jan Schipper, Sabine Siesling, Paul D.P. Pharoah, Vivianne C.G. Tjan-Heijnen, Adri C. Voogd.

Published in: The Breast, 2025.

ABSTRACT

Background: PREDICT Breast is a clinical decision-support tool estimating prognosis and the absolute benefit of adjuvant systemic therapies in early breast cancer. PREDICT v2.2 is recommended in Dutch guidelines. Both v2.2 and the recently updated v3.1 have not been validated in the Dutch population. This study compares the predictive performance of PREDICT v3.1 and v2.2 for 10-year OS in Dutch breast cancer patients.

Methods: Women diagnosed between 2005 and 2013 with primary invasive breast cancer were selected from the Netherlands Cancer Registry. Ten-year OS predictions from v2.2 and v3.1 were compared with observed OS for the overall cohort and 36 subgroups defined by oestrogen receptor (ER) status, HER2-status, age, and tumour stage. Discrimination (ability to distinguish patients with different outcomes) and calibration (agreement between predicted and observed outcomes) of both models were assessed.

Results: Among 101,282 patients, both versions showed moderate discrimination (AUC v2.2 = 0.768; v3.1 = 0.775) and calibration (v2.2 intercept: 0.07; slope: 1.09; v3.1 intercept: 0.12, slope: 1.00). V3.1 slightly overestimated (1.9%), whereas v2.2 slightly underestimated (1.6%) 10-year OS. Across subgroups, v3.1 generally outperformed v2.2 except in patients aged >75 years, where v2.2 provided more accurate estimates. In ER-/HER- patients aged 50-75 years, v3.1 overestimated (1.5-2.8%) and v2.2 underestimated (2.8-5.3%) 10-year OS.

Conclusion: Both PREDICT v2.2 and v3.1 accurately predict 10-year OS in Dutch breast cancer patients, with small differences between versions that vary by subgroup. No single model is optimal for all patients highlighting the need for subgroup-specific recalibration and careful interpretation when applying PREDICT.

INTRODUCTION

In Europe, approximately 600,000 patients were diagnosed with breast cancer in 2022[1]. This highlights the need for accurate prediction tools to support personalized treatment decisions[2]. PREDICT Breast is a clinical prediction model that supports decision-making on adjuvant systemic therapy for women with early invasive breast cancer by predicting baseline survival and the added benefit of adjuvant systemic therapies on 5-, 10-, and 15-year overall survival (OS) based on tumour- and patient-related characteristics[3].

Genomic assays such as Oncotype DX and MammaPrint are widely used to guide chemotherapy decisions. They were originally developed and validated as prognostic tools estimating baseline risk of distant recurrence[4,5,6].Accumulating data, however, suggests that these assays may also carry predictive information on chemotherapy benefit in hormone receptor-positive, HER2-negative early breast cancer[7,8]. At the same time, the ASTER study[9] and Beltran-Bless et al.[10] support that genomic scores often remain predominantly prognostic, and the extent to which they provide independent treatment-predictive information remains debated. This highlights the importance of validated clinical models like PREDICT and the integration of clinicopathological and genomic information.

The PREDICT algorithm, originally released in 2010 using eastern UK data from 1993 to 2003, estimates 5- and 10-year OS[3]. Since its initial release, PREDICT has been continuously updated to improve accuracy and incorporate new clinical advancements. V2.0 (2017) added predictors such as age at diagnosis, the exact tumour size, number of positive lymph nodes, and the presence of micrometastases, while subsequent updates (v2.1 and v2.2) included bisphosphonates therapy, 15-year OS estimates, and refined hormone therapy options[11]. V3.0 (2024) incorporated updated hazard ratios, progesterone receptor status, radiotherapy effects on breast cancer-specific mortality, and treatment effects on other causes of mortality[12], with v3.1, further refining baseline hazards for patients undergoing primary surgery.

The most recent validated version in the Dutch breast cancer population is PREDICT v2.0, which demonstrated good overall accuracy, indicating that predicted risks aligned reasonably well with observed outcomes and that the model could adequately distinguish between patients with different outcomes (e.g. alive vs dead). However, the model performed worse in oestrogen receptor (ER) negative patients, patients aged ≥75 years, T3 tumours, and those receiving both endocrine therapy and chemotherapy[13]. This supports the need to externally validate the updated version of PREDICT in predefined subgroups.

As PREDICT v2.2 is currently recommended in Dutch clinical guidelines[14], this study aims to evaluate whether PREDICT v3.1 provides improved predictive performance compared to v2.2 for 10-year OS in the Dutch breast cancer population. In addition, clinically rel-

evant subgroups defined by ER status, HER2 status, age, and tumour stage were used to validate both PREDICT models.

METHODS

Design

In this population-based external validation cohort study, data on patient-, tumour-, and treatment-related characteristics from the Netherlands Cancer Registry (NCR), which is hosted by the Netherlands Comprehensive Cancer Organisation (IKNL), were used. It is a comprehensive, population-based database that includes data on all newly diagnosed cancer cases in the Netherlands, registered by trained data managers. Included variables corresponded to those in PREDICT Breast: age, smoking status, ER status, progesterone status, HER2 status, Ki-67 status, tumour size and grade, mode of detection, and number of positive lymph nodes.

Tumour staging followed the tumour, node, and metastasis (TNM) classification system, with the use of the 6th or 7th edition depending on the year of diagnosis[15,16]. Chemotherapy classification changed from generation-based in v2.2 to standard-dose anthracycline-based and high-dose anthracycline and/or taxane-based regimes in v3.1. Vital status and date of death were derived from the Municipal Personal Records database, which contained complete records until February 2024. Because cause of death was not recorded, we analysed all-cause mortality without distinguishing between breast cancer-specific and other-cause mortality.

Patients

In this study, all women with primary invasive breast cancer in the Netherlands, diagnosed from 2005 to 2013, who underwent surgery were included. Patients who received neoadjuvant therapy, had no pathologically confirmed tumour, had metastatic disease at diagnosis, were younger than 25 or older than 85 years, or were recorded as alive but with incomplete follow-up information were excluded. These inclusion criteria match those of the development cohorts of PREDICT v2.2 and v3.1[3,12].

Statistical analysis

The outcome of interest was 10-year OS, defined as time from diagnosis to death from any cause, last observation or end of follow-up. Predicted OS was calculated using PREDICT v2.2 and v3.1 functions obtained from the developer (P.D.P.P). We compared the 10-year predicted survival probabilities with the observed 10-year survival in the data using Kaplan-Meier estimates.

For missing data, the PREDICT algorithm allows specific variables to be missing (Ki-67 status, progesterone status and mode of detection). For the other variables we used multiple imputation (MICE)[17] with 5 imputations for tumour size, number of positive lymph nodes, differentiation grade, HER2 status, and ER status. A single dataset was created by averaging numeric variables and taking the majority for categorical variables. Smoking status was unavailable for 101,207 patients (>99.9%) in the validation cohort, therefore smoking status was coded as non-smoking for the entire cohort.

Calibration (agreement between predicted and observed outcomes) and discrimination (the ability to distinguish between patients with different outcomes, e.g. alive vs dead) were assessed[18] for the total cohort and 36 subgroups based on ER status (positive or negative), HER2 status (positive or negative), age (<50, 50–75 or >75), and TNM stage (I, II or III). A 95% confidence interval was calculated for observed OS based on Kaplan Meier estimates. Confidence intervals for the predictions were not computed, as PREDICT does not consider uncertainties in its predictions.

Simplified calibration plots were used to visually assess how well model predictions align with the observed outcome frequencies. Calibration slope was used to assess over- or underfitting. A slope <1 indicates overfitting: small risks are underestimated, and large risks are overestimated, while a slope >1 indicates the opposite (underfitting). If the slope equals 1 but the intercept differs from 0, predictions are systemically too high (intercept <0) or too low (intercept >0) on average[19].

Discriminative accuracy, which reflects the model's ability to distinguish between different outcomes (i.e. dead vs alive at 10 years), was assessed using the area under the receiver operating curve (AUC). An AUC closer to 1.0 indicates excellent discrimination, while a value of 0.5 suggests no predictive capability beyond random chance. AUC values were interpreted as follows: ≥0.80 is good, 0.69–0.79 is moderate, and <0.69 is poor[20].

A subgroup analysis was performed for patients who did not receive adjuvant therapy, representing the baseline on which adjuvant treatment decisions are made. Calibration (agreement between predicted and observed outcomes) and discrimination (the ability to distinguish between patients with different outcomes) were assessed[18]. All statistical analyses were performed in R version 4.4.2. Validation and reporting were carried out according to the TRIPOD criteria (Supplementary Table 1)[21].

RESULTS

Patient characteristics

A total of 128,380 women with primary invasive breast cancer diagnosed between 2005 and 2013 were identified from the NCR (Figure 1). In total 101,282 (78.9%) patients met eligibility criteria (Figure 1; Table 1). Adjuvant therapy use differed between subgroups, with chemotherapy and trastuzumab most common in young HER2+ patients, and hormone therapy in ER+ subgroups (Supplementary Table 2). Patients not receiving adjuvant therapy were older and had higher 10-year survival compared with the total population (Supplementary Table 3).

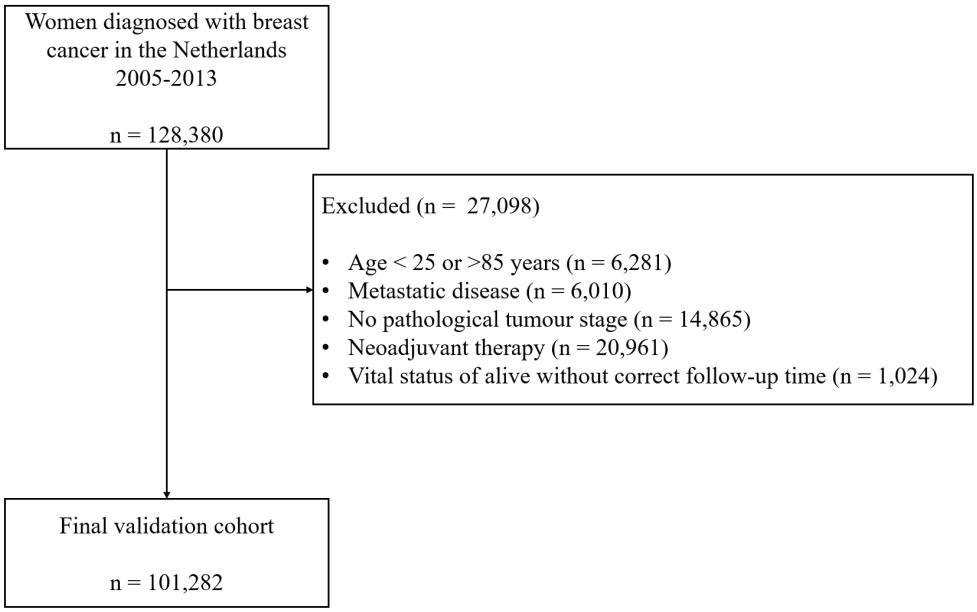


Figure 1. Patient selection for the Dutch study cohort for the validation of PREDICT v2.2 and v3.1.

Table 1. Baseline characteristics, treatment and 10-year survival in the Dutch validation cohort.

Characteristic	n = 101,282
Year of diagnosis	2005-2013
Age, median (IQR)	60.0 (51.0-69.0)
Menopausal status, n (%)*	
Pre-menopausal	7,874 (23.9)
Post-menopausal	25,136 (76.1)
Missing	68,272
Tumour size (mm), median (IQR)	16.0 (11.0 – 23.0)
Missing	2,170

Table 1. Baseline characteristics, treatment and 10-year survival in the Dutch validation cohort. (*continued*)

Characteristic	n = 101,282
Lymph nodes, n (%)	
≥ 1	34,377 (34.4%)
0	65,470 (65.6%)
Missing	1,435
Tumour stage, n (%)	
Stage I	52,969 (52.3)
Stage II	38,233 (37.7)
Stage III	10,080 (10.0)
Differentiation, n (%)	
Grade 1	24,193 (24.9)
Grade 2	44,155 (45.5)
Grade 3	29.792 (29.6)
Missing	4,142
Oestrogen receptor status, n (%)	
Positive	83,804 (83.9)
Negative	16,054 (16.1)
Missing	1,424
HER2 status, n (%)	
Positive	12,265 (12.1)
Negative	83,713 (82.9)
Unknown	5,039 (5.0)
Missing	265
Progesterone status, n (%)	
Positive	66,467 (65.8)
Negative	30,974 (30.7)
Unknown	3,563 (3.5)
Missing	278
Mode of detection, n (%)	
Clinically detected	12,417 (44.3)
Screening detected	15,620 (55.7)
Missing	73,245
Adjuvant therapy, n (%)	
Chemotherapy	38,498 (38.0)
Generation 2[†]	13,738 (13.6)
Generation 3[†]	24,757 (24.4)
Hormone therapy	51,737 (51.1)
Trastuzumab	7,684 (7.6)
Radiotherapy	67,045 (66.2)
Vital status at 10-year, n (%)	
Alive	77,739 (76.8)
Death	23,543 (23.2)

*Abbreviations: IQR = interquartile range, HER2 = human epidermal growth factor receptor 2. *Menopausal status was used to determine eligibility for bisphosphonate therapy and was not included as predictor in the model. [†]v3.1: Generation 2 = standard-dose anthracycline; Generation 3 = high-dose anthracycline/taxane-based*

Discriminatory accuracy and ability for 10-year overall survival

The discriminatory accuracy for 10-year OS was moderate in the total population for both PREDICT versions. The AUC of v2.2 was 0.768 (95% CI: 0.765-0.771), while for v3.1 it was slightly higher at 0.775 (95% CI: 0.772-0.778). The difference in AUCs was 0.007 (95% CI: 0.006-0.008), indicating a small but consistent improvement in discriminatory performance with v3.1.

A similar pattern was observed in patients not receiving adjuvant therapy, with an AUC of 0.818 (95% CI: 0.811-0.825) for v2.2 and 0.826 (95% CI: 0.818 – 0.833) for v3.1.

Calibration for 10-year overall survival

At 10-year, v3.1 slightly overestimated OS in the total cohort, with a mean predicted survival of 78.7% compared to an observed survival of 76.8% (difference: +1.9%; Table 2). In contrast, v2.2 slightly underestimated OS with a predicted survival of 75.2% (difference: -1.6%; Table 2). Calibration plots for the total cohort demonstrated good agreement between predicted and observed OS for both models, with slopes of 1.09 (v2.2) and 1.00 (v3.1) and intercepts of –0.07 and 0.12, respectively (Figure 2).

Overall, both PREDICT models demonstrated good calibration across the 36 predefined subgroups, with calibration slopes close to 1 and intercepts close to 0 in the majority of the cases (Figures 3-6). V3.1 demonstrated better calibration for 10-year OS than v2.2 in the majority of the subgroups (Table 2).

Among ER-/HER2- patients aged 50–75 years, the models demonstrated opposite calibration patterns across all tumour stages, which was similar to the calibration observed in the overall population. V3.1 tended to overestimate 10-year OS by 1.5%–2.8%, whereas v2.2 underestimated survival by 2.8%–5.3%. However, notable exceptions were observed. In certain subgroups v2.2 outperformed v3.1. In patients aged >75 years, v3.1 tended to overestimate survival compared with v2.2. For example, in ER+/HER2- patients with stage II disease, v3.1 overestimated survival more by 12.2% versus 8.8% with v2.2. Similarly, in ER-/HER2-patients with stage I disease, overestimation was 14.4% for v3.1 compared to 9.5% for v2.2. Comparable patterns were noted in ER+/HER2+ patients.

For patients not treated with adjuvant therapy, both versions overestimated 10-year OS, with 4.1% (v2.2) and 6.7% (v3.1) (Supplementary Table 4). The calibration plot shows good agreement between predicted and observed survival, with slopes of 1.09 (v2.2) and 0.96 (v3.1) and intercepts of 0.21 and 0.33, respectively (Supplementary Figure 1).

Table 2: Observed and predicted 10-year overall survival by subgroups based on ER status, HER2 status, age, and tumour stage for PREDICT v2.2 and v3.1.

					PREDICT v2.2		PREDICT v3.1			
N (%)				Observed, % (95% CI)	Predicted, %	Difference, %	Predicted, %	Difference, %		
Total cohort					101282 (100.0)	76.8 (76.5-77.0)	75.2	-1.6	78.7	1.9
ER-status	**HER2-status**	**Age**	**Stage**							
ER+	HER2+	< 50	I		877 (0.9)	94.4 (92.9-95.9)	90.5	-3.9	92.7	-1.7
			II		1163 (1.1)	91.1 (89.4-92.7)	84.9	-6.2	87.9	-3.2
			III		424 (0.4)	79.0 (75.2-83.0)	59.9	-19.1	69.1	-9.9
		50-75	I		2422 (2.4)	84.6 (83.1-86.0)	82.1	-2.4	85.4	0.9
			II		1849 (1.8)	78.7 (76.9-80.6)	76.4	-2.3	79.1	0.4
			III		546 (0.5)	65.9 (62.1-70.0)	54.8	-11.1	61.0	-5.0
		>75	I		153 (0.1)	46.4 (39.1-55.0)	46.9	0.5	51.5	5.1
			II		323 (0.3)	31.0 (26.3-36.4)	37.8	6.9	39.4	8.4
			III		114 (0.1)	20.2 (14.0-29.1)	21.0	0.9	22.4	2.2
ER+	HER2-	< 50	I		6502 (6.4)	94.2 (93.6-94.8)	92.2	-2.0	94.0	-0.2
			II		6354 (6.3)	89.3 (88.6-90.1)	87.4	-1.9	89.5	0.2
			III		1748 (1.7)	73.4 (71.4-75.5)	67.1	-6.3	73.9	0.5
		50-75	I		33331 (32.9)	84.6 (84.2-5.0)	82.9	-1.6	86.8	2.2
			II		17175 (17.0)	78.45 (77.8-79.1)	78.6	0.2	81.7	3.2
			III		3972 (3.9)	60.8 (59.3-62.3)	60.2	-0.5	65.4	4.6
		>75	I		3062 (3.0)	44.4 (42.7-46.2)	48.9	4.5	54.6	10.2
			II		4068 (4.0)	33.8 (32.3-35.2)	42.6	8.8	46.0	12.2

Table 2: Observed and predicted 10-year overall survival by subgroups based on ER status, HER2 status, age, and tumour stage for PREDICT v2.2 and v3.1. (*continued*)

				PREDICT v2.2		PREDICT v3.1			
			III	1020 (1.0)	21.2 (18.8-23.8)	27.6	6.4	29.3	8.1
ER-	HER2+	< 50	I	353 (0.3)	90.9 (88.0-94.0)	83.1	-7.8	86.5	-4.4
			II	537 (0.5)	85.3 (82.3-88.3)	75.7	-9.5	76.0	-9.3
			III	297 (0.3)	69.0 (64.0-74.5)	47.8	-21.2	48.5	-20.6
		50-75	I	1227 (1.2)	84.5 (82.5-86.6)	72.2	-12.3	80.3	-4.2
			II	1216 (1.2)	78.0 (75.8-80.4)	65.3	-12.7	69.8	-8.2
			III	509 (0.5)	64.6 (60.6-68.9)	37.9	-26.8	44.6	-20.1
		>75	I	87 (0.1)	28.7 (20.6-40.0)	35.4	6.7	38.8	10.1
			II	213 (0.2)	30.5 (24.9-37.4)	26.1	-4.4	23.6	-6.9
			III	120 (0.1)	8.3 (4.6-15.1)	7.6	-0.7	6.5	-1.8
ER-	HER2-	< 50	I	1367 (1.4)	89.9 (88.3-91.5)	84.3	-5.6	86.2	-3.7
			II	1830 (1.8)	80.3 (78.5-82.2)	76.7	-3.6	75.7	-4.6
			III	373 (0.4)	52.8 (48.0-58.1)	46.1	-6.7	47.1	-5.7
		50-75	I	3286 (3.2)	78.1 (76.7-79.5)	73.3	-4.8	80.9	2.8
			II	2795 (2.8)	67.7 66.0-69.5	65.0	-2.8	69.2	1.5
			III	697 (0.7)	40.9 (37.4-44.7)	35.6	-5.3	42.0	1.1
		>75	I	302 (0.3)	31.5 (26.6-37.2)	41.0	9.5	45.9	14.4
			II	710 (0.7)	26.1 (23.0-29.5)	32.1	6.1	31.6	5.6
			III	260 (0.3)	11.9 (8.6-16.6)	32.3	0.5	11.4	-0.6

Abbreviations: N = total number, CI = confidence interval, ER = oestrogen receptor, HER2 = human epidermal growth factor receptor 2

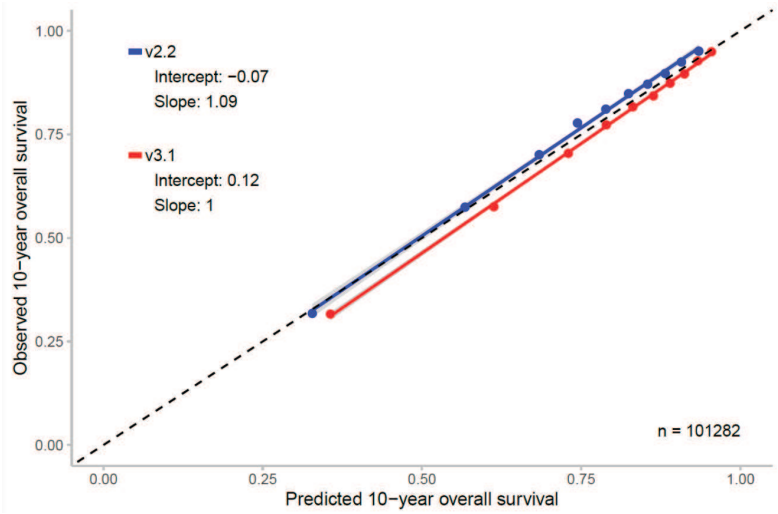


Figure 2: Observed and predicted 10-year overall survival for the Dutch validation cohort of PREDICT v2.2 and v3.1 with the 95% confidence interval.

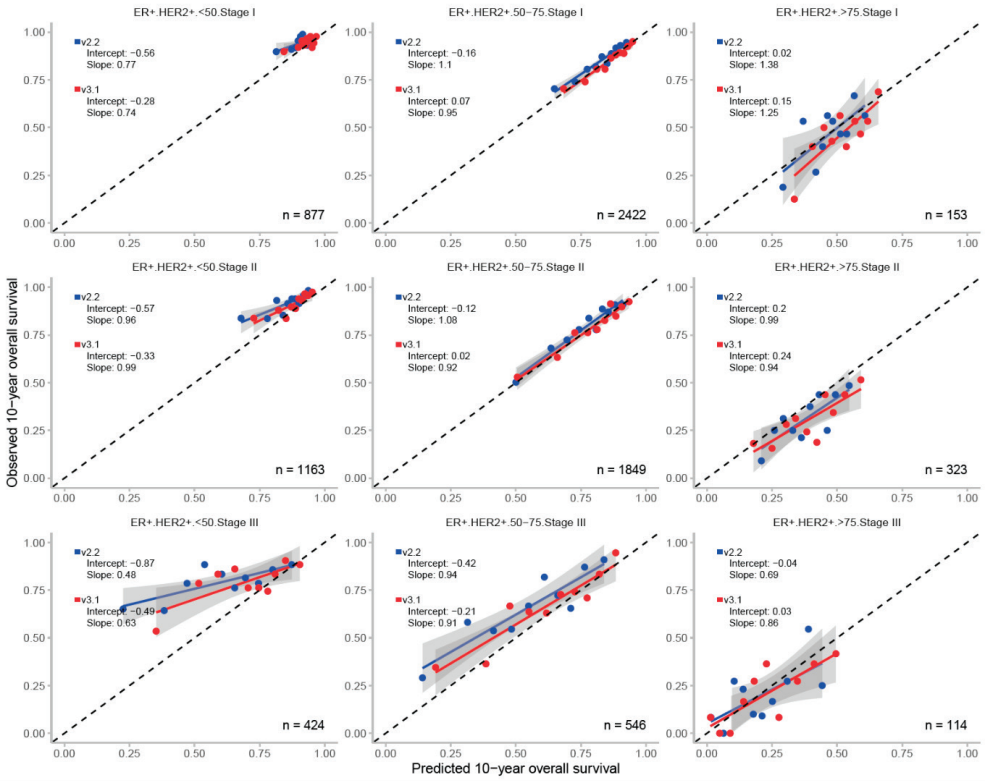


Figure 3: Observed and predicted 10-year overall survival for patients with oestrogen receptor (ER) positive and human epidermal growth factor receptor 2 (HER2) positive breast cancer, stratified by age and tumour stage with the 95% confidence interval.

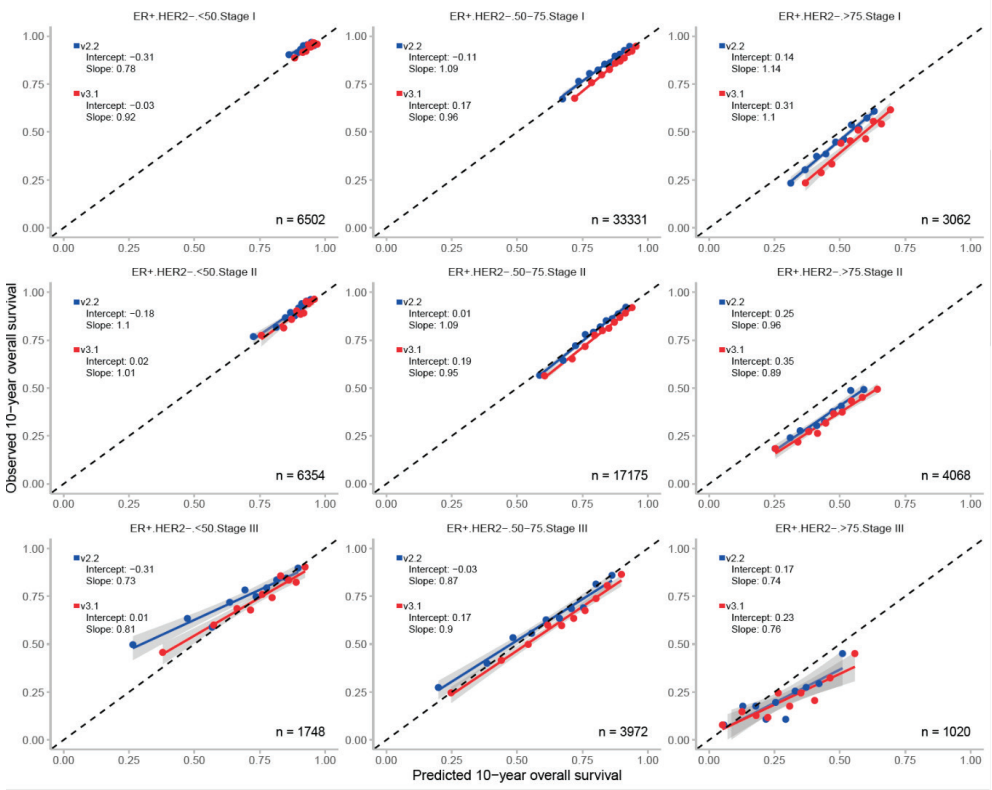


Figure 4: Observed and predicted 10-year overall survival for patients with oestrogen receptor (ER) positive and human epidermal growth factor receptor 2 (HER2) negative breast cancer, stratified by age and tumour stage with the 95% confidence interval.

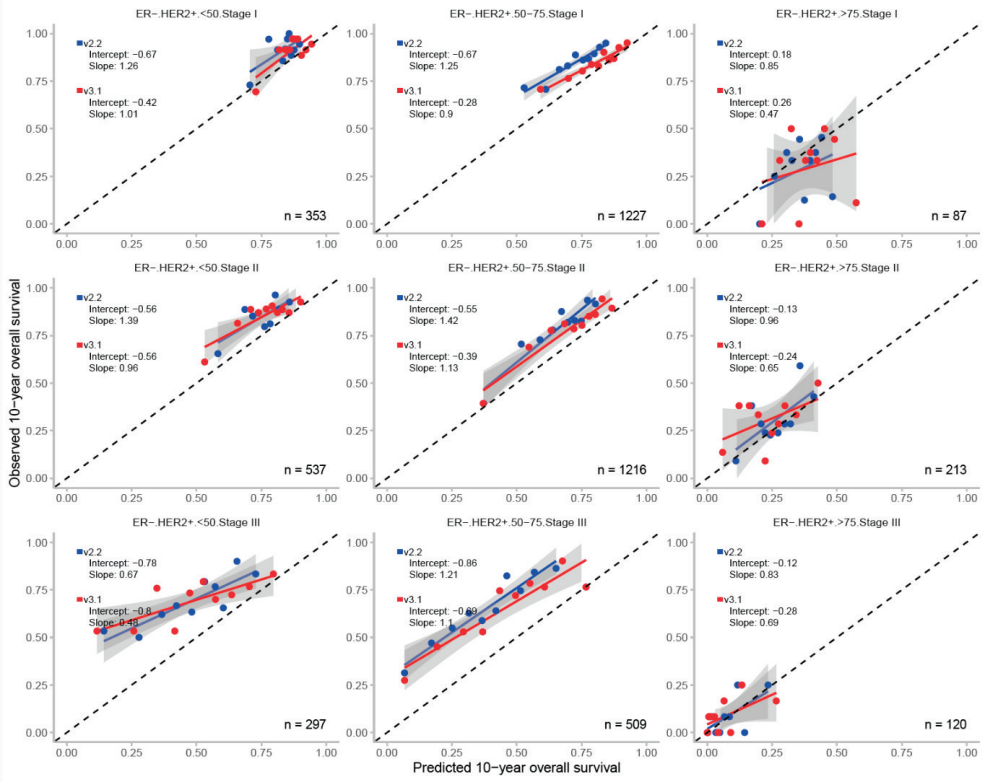


Figure 5: Observed and predicted 10-year overall survival for patients with oestrogen receptor (ER) negative and human epidermal growth factor receptor 2 (HER2) positive breast cancer, stratified by age and tumour stage with the 95% confidence interval.

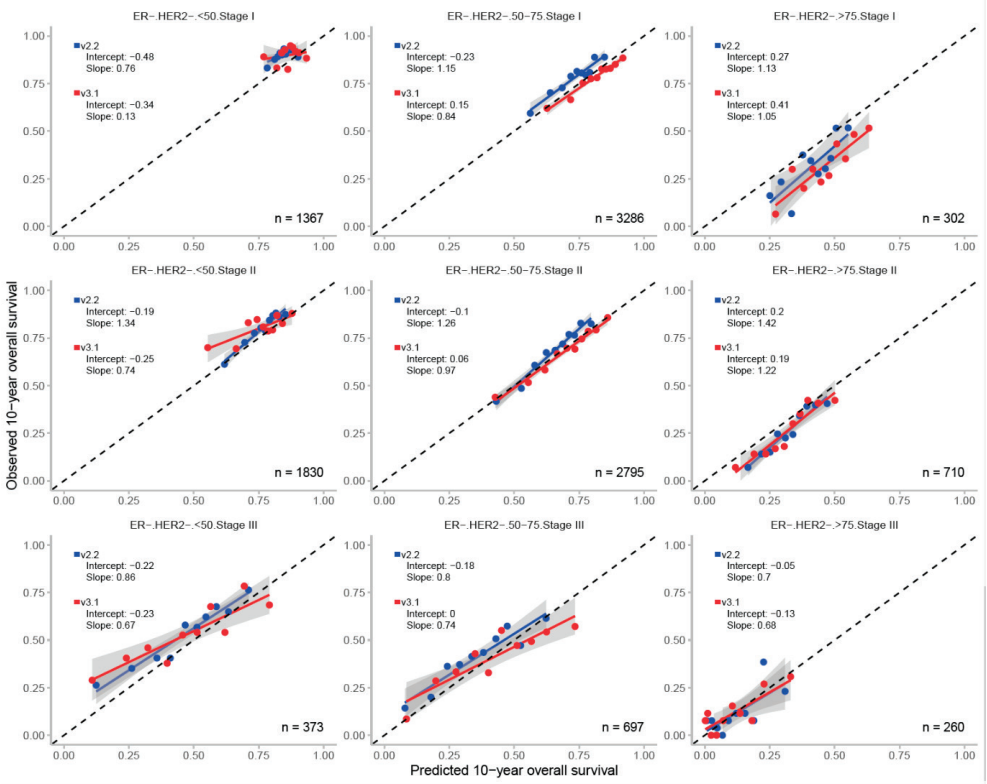


Figure 6: Observed and predicted 10-year overall survival for patients with oestrogen receptor (ER) negative and human epidermal growth factor receptor 2 (HER2) negative breast cancer, stratified by age and tumour stage with the 95% confidence interval.

DISCUSSION

In this study, we compared the performance of PREDICT v3.1 and PREDICT v2.2 in an external validation predicting 10-year OS in a large Dutch breast cancer cohort. Both versions demonstrated moderate discriminatory accuracy (the ability to distinguish between patients with different outcomes) and good calibration (agreement between predicted and observed outcomes) in the overall population. V3.1 slightly overestimated OS by 1.9%, while v2.2 underestimated OS by 1.6%. Across the subgroups v3.1 outperformed v2.2 in the majority of the 36 predefined subgroups, although notable exceptions were observed. In older patients, aged >75 years, v2.2 tended to have more accurate predictions than v3.1, particularly in ER+/HER2- patients. Among ER-/HER2- patients aged 50–75 years, the models showed opposite calibration trends with v3.1 overestimating and v2.2 underestimating survival. These patterns suggest that while both versions are generally well-calibrated, certain patient populations may benefit from tailored recalibration. PREDICT estimates both prognosis and the benefit of adjuvant systemic therapies, providing accessible and complementary clinical value to primarily prognostic genomic assays. Just like these assays, PREDICT has limitations in certain subgroups, emphasizing the need for ongoing validation and careful interpretation in clinical practice.

Compared to previous PREDICT validations, we observed similar discriminatory performance, with AUC comparable with earlier studies worldwide (0.709–0.780) reported for various PREDICT versions[13,22,23,24].

While PREDICT consistently demonstrates moderate discrimination, calibration is more important than discrimination for absolute risk estimation[25]. V2.2 showed slightly better calibration than v3.1 for 10-year OS in the overall Dutch cohort. Consistent with our findings, the Dutch validation study of PREDICT v2.0[13], using patients diagnosed in 2005, reported a 10-year OS difference of −1.0% compared with −1.6% (v2.2) and 2.0% (v3.1) in our study, suggesting that all versions performed well across the Dutch population. Chen et al.[14] found improved calibration for v3.0 compared to v2.2 in a Chinese cohort for the 5-year OS. Direct comparisons are challenging due to differences in model versions, survival outcomes (5 vs. 10 years), healthcare systems, and patient populations. Availability and uptake of (insured) treatment options differ across countries. For instance, trastuzumab was rapidly implemented after approval in the Netherlands[26], while uptake was slower elsewhere[27], influencing validation outcomes. Moreover, although PREDICT v3.1 distinguishes standard-from high-dose anthracycline regimens, high-dose anthracyclines are now mostly reserved for high-risk patients. Therefore, this distinction is of limited relevance for most contemporary patients.

Calibration in specific subgroups indicated lower predictive accuracy for patients aged >75 years and ER- tumours, in combination with other tumour characteristics. These

findings are consistent with validation studies of earlier PREDICT versions[13,24,28] and may reflect the underrepresentation of these patients implying that caution remains warranted. In contrast, v3.1 showed clear improvement for patients <75 years and HER2+ tumours, for whom v2.2 underestimated OS. Chen et al. previously reported improved accuracy in v3.0 for this subgroup[24], reflecting updated treatment effects that capture improved prognosis associated with advances in HER2-targeted therapy[29]. Accordingly, the improved calibration in v3.1 for HER2+ patients aligns with these updates. Nevertheless, although the lower estimated benefit of various adjuvant systemic treatments[12] may reflect updated evidence, careful interpretation is necessary to prevent overly conservative decisions that might inadvertently lead to undertreatment and compromise long-term outcomes. In patients not receiving any adjuvant systemic therapy, both versions showed good discrimination but overestimated OS, which may indicate patient selection.

Predictions for high-risk patients, such as those with HER2+ tumours or advanced tumour stage, often deviated from observed survival. Clinically, however, this is less relevant, as these patients almost invariably receive adjuvant treatment regardless of predicted risk. Calibration was considerably better in low- or intermediate-risk patients, where predictions more closely aligned with observed outcomes and may therefore support clinical decision-making for adjuvant therapy[30].

Differences and similarities between our Dutch cohort and the UK development dataset were observed[12]. Age, tumour size, tumour grade, and prevalence of ER-positive disease were largely comparable, whereas treatment patterns differed more notably. Only 51.1% of patients in our cohort received endocrine therapy, compared with 60% in eastern UK and 40% in West Midlands[12]. Differences in endocrine therapy uptake may affect calibration, as PREDICT incorporates treatment benefit. Additional factors, including patient preferences, shared decision-making, and limitations in registry-based data not capturing treatment adherence may contribute to cohort differences.

Patients who received neoadjuvant therapy are excluded in PREDICT. The use of neoadjuvant therapy in the Netherlands has increased over time from 9% in 2005 to 51% in 2023[31], meaning that patients who still undergo upfront surgery, as in our cohort, represent a selected subgroup that no longer reflects routine clinical care in 2025. Although survival is expected to be similar in patients receiving neoadjuvant therapy, different predictive factors may be required, as seen in the INFLUENCE model[32]. Therefore, future validation studies of PREDICT in neoadjuvant treated patients are necessary to determine how these patients can be accurately integrated into the PREDICT tool. Additionally, this selection bias may contribute to the observed overestimation of survival in subgroups that are now more frequently treated with neoadjuvant therapy and are therefore partly excluded from validation.

Calibration and discrimination of both PREDICT models were generally adequate at the population level. However, caution is warranted for certain subgroups. For instance, older patients and certain tumour subtypes may benefit from recalibration or careful interpretation of predictions. Tools such as PORTRET, developed for older patients, address limitations of commonly used models in this group[33]. These findings highlight the need to consider subgroup-specific performance when applying PREDICT in clinical practice.

Strengths and limitations

This is the first study to compare PREDICT v2.2 and v3.1 in estimating 10-year OS using a large, population-based Dutch cohort. A major strength of this study is the large sample size, enabling the stratification across clinically relevant subgroups, which gives a more nuanced evaluation of model performance and provides concrete recommendations based on patient and tumour characteristics.

Although our sample size was large, some subgroups (particularly patients aged >75 years and ER-/HER2+ patients aged <50 years) remained small. Results for these subgroups should be interpreted with caution. As seen in many prediction models, performance often decreases at the extremes, and preferably we would have had larger samples for more stable results[34]. However, their inclusion is crucial for the model's applicability for these subgroups.

Our findings should be interpreted in the context of contemporary clinical practice. While PREDICT is widely used, its limitations are most evident in aggressive tumour subtypes, extreme age groups, and neoadjuvant settings. Like genomic assays, model performance may vary across time and populations, highlighting the need for ongoing validation and careful interpretation.

Additionally, some assumptions were necessary as not all data was available. Smoking status was unavailable, however coding all patients as smokers versus non-smokers had minor effect on the results, therefore all patients were classified as non-smokers. Moreover, information on Ki-67 status was completely missing in our data and was therefore set to 'unknown'. Nevertheless, improvements in predictive performance can be expected if information on Ki-67 is known[35]. Heart dose (in grays) from radiotherapy was missing and therefore assumed to be zero, despite variation across Dutch institutions (0.80–1.82 Gy)[36], as doses <2.5 Gy ensure adequate heart protection[37]. Bisphosphonate use was set to 'not used', although recommended for postmenopausal women receiving adjuvant systemic therapy, as it modestly improves OS[38]. However, this assumption is unlikely to bias results towards worse outcomes. In PREDICT, the duration of hormone therapy can be chosen to be 5 or 10 years. However, we did not have information about the duration of hormone therapy in the dataset and therefore we assumed a duration of 5-years for every patient, following the Dutch guidelines[39].

CONCLUSION

Both PREDICT v3.1 and v2.2 demonstrate moderate overall discrimination and good calibration for predicting 10-year OS in the Dutch breast cancer population. Within subgroups, differences in calibration between versions are minimal but vary across subgroups, indicating that no single model is optimal for all patient groups. These findings suggest that neither model consistently outperforms the other across all subgroups, highlighting the need for subgroup-specific recalibration and caution when interpreting predictions, particularly in subgroups where model's performance may be limited. It should be noted that primary validation was performed in the Dutch population. Further studies and integration of additional datasets are needed to assess the relevance and generalizability of PREDICT and improve predictive accuracy across diverse populations.

Supplementary Information

Supplementary material is available at: https://www.thebreastonline.com/article/S0960-9776(25)00900-2/fulltext

Acknowledgements

We thank the Netherlands Cancer Registry for providing the data, and we are grateful to the registration clerks for their efforts in collecting and maintaining the registry.

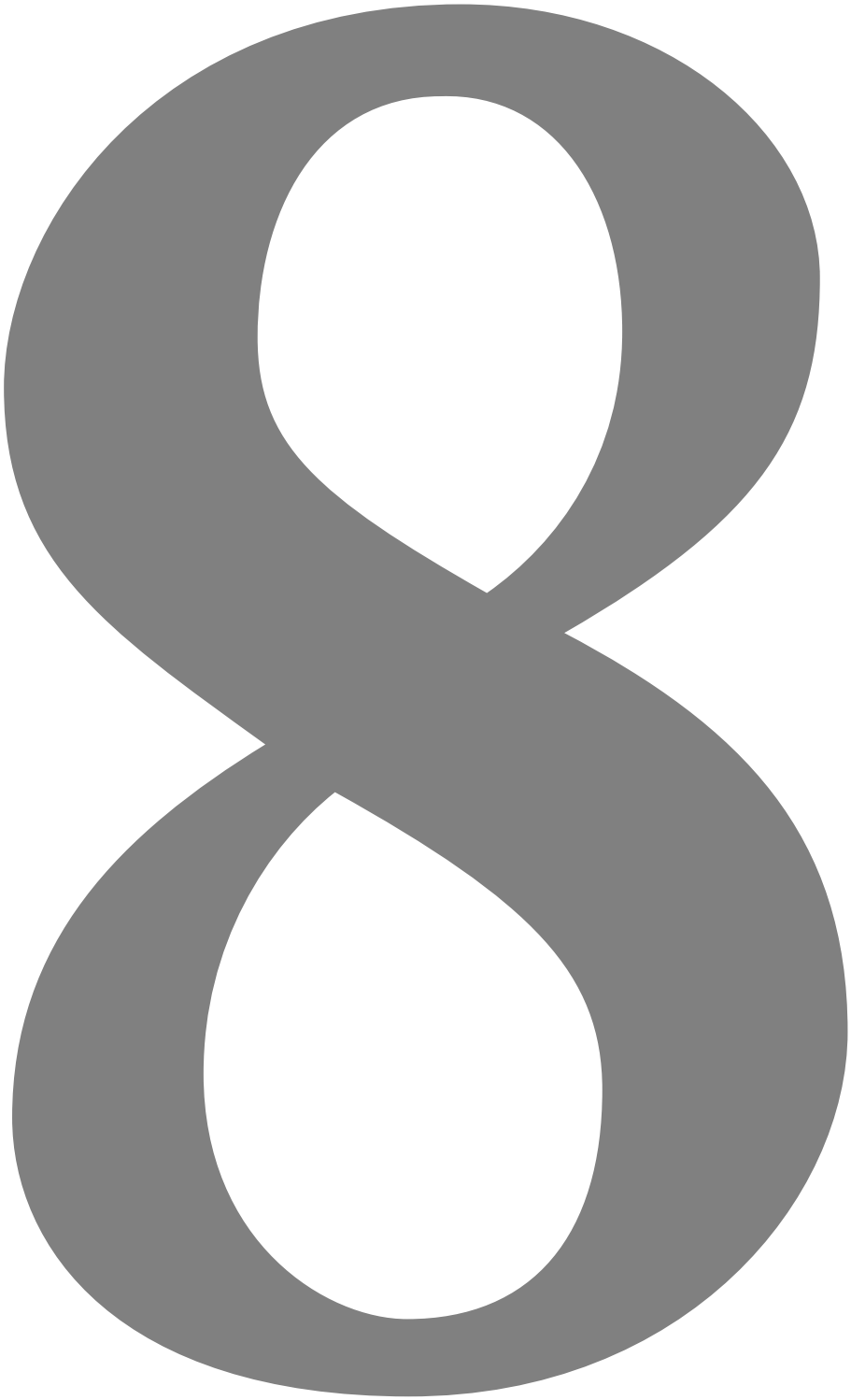
REFERENCES

1. Europa Donna Breast Cancer Facts. 2024 https://www.europadonna.org/breast-cancer/ [cited 2025 29 July].
2. Pak L.M., Morrow M. Addressing the problem of overtreatment in breast cancer. Expert Rev Anticancer Ther. 2022;22(5):535–548. doi: 10.1080/14737140.2022.2064277.
3. Wishart G.C., Azzato E.M., Greenberg D.C., Rashbass J., Kearins O., Lawrence G., et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res. 2010;12(1) doi: 10.1186/bcr2464.
4. Wang D., Wang X., Yang X. External validation of the oncotype DX breast cancer recurrence score nomogram and development and validation of a novel machine learning-based model to predict postoperative overall survival and guide adjuvant chemotherapy in ER positive, Her-2 negative breast cancer patients: a retrospective cohort study. Front Oncol. 2025;15 doi: 10.3389/fonc.2025.1586262.
5. Berton Giachetti PPM., Carnevale Schianca A., Trapani D., Marra A., Toss A., Marchio C., et al. Current controversies in the use of oncotype DX in early breast cancer. Cancer Treat Rev. 2025;135 doi: 10.1016/j.ctrv.2025.102887.
6. van 't Veer L.J., Dai H., van de Vijver M.J., He Y.D., Hart A.A., Mao M., et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–536. doi: 10.1038/415530a.
7. Brufsky A.M., Hoskins K.F., Conter H.J., Kelemen P., Habibi M., Samian L., et al. MammaPrint predicts chemotherapy benefit in HR+HER2- early breast cancer: FLEX registry real-world data. JNCI Cancer Spectr. 2025;9(5) doi: 10.1093/jncics/pkaf079.
8. Livraghi L., Martella F., Ghilli M., Angiolini C., Magnanini S., Moretti E., et al. Prospective study on oncotype DX(R) assay to assess recurrence risk in early ER-positive HER2-negative breast cancer patients with uncertain biological behavior by standard parameters and its impact on treatment recommendation: the POST trial. Eur J Cancer. 2024;213 doi: 10.1016/j.ejca.2024.115108.
9. Brain E., Mir O., Bourbouloux E., Rigal O., Ferrero J.M., Kirscher S., et al. Adjuvant chemotherapy and hormonotherapy versus adjuvant hormonotherapy alone for women aged 70 years and older with high-risk breast cancer based on the genomic grade index (ASTER 70s): a randomised phase 3 trial. Lancet. 2025;406(10502):489–500. doi: 10.1016/S0140-6736(25)00832-3.
10. Beltran-Bless A.A., Pond G.R., Bayani J., Barker S.L., Spears M., Mallon E., et al. Does RSClin provide additional information over classic clinico-pathologic scores (PREDICT 2.1, INFLUENCE 2.0, CTS5)? Breast. 2025;83 doi: 10.1016/j.breast.2025.104528
11. Candido Dos Reis F.J., Wishart G.C., Dicks E.M., Greenberg D., Rashbass J., Schmidt M.K., et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res. 2017;19(1):58. doi: 10.1186/s13058-017-0852-3.
12. Grootes I., Wishart G.C., Pharoah P.D.P. An updated PREDICT breast cancer prognostic model including the benefits and harms of radiotherapy. npj Breast Cancer. 2024;10(1):6. doi: 10.1038/s41523-024-00612-y.
13. van Maaren M.C., van Steenbeek C.D., Pharoah P.D.P., Witteveen A., Sonke G.S., Strobbe L.J.A., et al. Validation of the online prediction tool PREDICT v. 2.0 in the Dutch breast cancer population. Eur J Cancer. 2017;86:364–372. doi: 10.1016/j.ejca.2017.09.031.
14. Federatie Medisch Specialisten Borstkanker 2025. https://richtlijnendatabase.nl/richtlijn/borstkanker/startpagina_-_borstkanker.html [cited 2025 1 June].

15. Sobin L.H.W.C. sixth ed. Wiley; New York: 2002. International union against cancer, TNM classification of malignant tumours.
16. Sobin L.H., Compton C.C. TNM seventh edition: what's new, what's changed: communication from the international union against cancer and the American joint committee on cancer. Cancer. 2010;116(22):5336–5339. doi: 10.1002/cncr.25537.
17. van Buuren S., Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Software. 2011;45(3):1–67. doi: 10.18637/jss.v045.i03.
18. Steyerberg E.W., Vickers A.J., Cook N.R., Gerds T., Gonen M., Obuchowski N., et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138. doi: 10.1097/EDE.0b013e3181c30fb2
19. Van Calster B., Vickers A.J. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Mak. 2015;35(2):162–169. doi: 10.1177/0272989X14547233.
20. .de Hond A.A.H., Steyerberg E.W., van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health. 2022;4(12):e853–e855. doi: 10.1016/S2589-7500(22)00188-1.
21. Collins G.S., Reitsma J.B., Altman D.G., Moons K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55–63. doi: 10.7326/M14-0697.
22. Wong H.S., Subramaniam S., Alias Z., Taib N.A., Ho G.F., Ng C.H., et al. The predictive accuracy of PREDICT: a personalized decision-making tool for Southeast Asian women with breast cancer. Medicine (Baltim) 2015;94(8) doi: 10.1097/MD.0000000000000593
23. Wishart G.C., Bajdik C.D., Azzato E.M., Dicks E., Greenberg D.C., Rashbass J., et al. A population-based validation of the prognostic model PREDICT for early breast cancer. Eur J Surg Oncol. 2011;37(5):411–417. doi: 10.1016/j.ejso.2011.02.001.
24. Chen E., Chen C., Chen Y., You J., Jin C., Huang Z., et al. Insights into the performance of PREDICT tool in a large mainland Chinese breast cancer cohort: a comparative analysis of versions 3.0 and 2.2. Oncologist. 2024;29(8):e976–e983. doi: 10.1093/oncolo/oyae164
25. Alba A.C., Agoritsas T., Walsh M., Hanna S., Iorio A., Devereaux P.J., et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA. 2017;318(14):1377–1384. doi: 10.1001/jama.2017.12126
26. 26.de Munck L., Schaapveld M., Siesling S., Wesseling J., Voogd A.C., Tjan-Heijnen V.C., et al. Implementation of trastuzumab in conjunction with adjuvant chemotherapy in the treatment of non-metastatic breast cancer in the Netherlands. Breast Cancer Res Treat. 2011;129(1):229–233. doi: 10.1007/s10549-011-1451-0.
27. Ades F., Senterre C., Zardavas D., de Azambuja E., Popescu R., Piccart M. Are life-saving anticancer drugs reaching all patients? Patterns and discrepancies of trastuzumab use in the European Union and the USA. PLoS One. 2017;12(3) doi: 10.1371/journal.pone.0172351.
28. Yi-Wen Hsaio, Wishart G.C., Pharoah P.D.P., Peng Pei-Chen. Validation of the PREDICT breast version 3.0 prognostic tool in US breast cancer patients. 2024.
29. Swain S.M., Shastry M., Hamilton E. Targeting HER2-positive breast cancer: advances and future directions. Nat Rev Drug Discov. 2023;22(2):101–126. doi: 10.1038/s41573-022-00579-0.
30. Strasser-Weippl K., Goss P.E. Competing risks in low-risk breast cancer. Am Soc Clin Oncol Educ Book. 2013:32–39. doi: 10.14694/EdBook_AM.2013.33.32.
31. Netherlands Comprehensive Cancer Organisation (IKNL) Borstkanker in Nederland 2025. https://iknl.nl/borstkankercijfers [cited 2025 6 Feb].
32. Van Maaren M.C., Hueting T.A., van Uden D.J.P., van Hezewijk M., de Munck L., Mureau M.A.M., et al. The INFLUENCE 3.0 model: updated predictions of locoregional recurrence and con-

tralateral breast cancer, now also suitable for patients treated with neoadjuvant systemic therapy. Breast. 2025;79 doi: 10.1016/j.breast.2024.103829.

33. van der Plas-Krijgsman W.G., Giardiello D., Putter H., Steyerberg E.W., Bastiaannet E., Stiggelbout A.M., et al. Development and validation of the PORTRET tool to predict recurrence, overall survival, and other-cause mortality in older patients with breast cancer in the Netherlands: a population-based study. Lancet Healthy Longev. 2021;2(11):e679–e680. doi: 10.1016/S2666-7568(21)00258-0.
34. Collins G.S., Ogundimu E.O., Altman D.G. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35(2):214–226. doi: 10.1002/sim.6787.
35. Kanyilmaz G., Yavuz B.B., Aktan M., Karaagac M., Uyar M., Findik S. Prognostic importance of Ki-67 in breast cancer and its relationship with other prognostic factors. Eur J Breast Health. 2019;15(4):256–261. doi: 10.5152/ejbh.2019.4778.
36. Maliko N., Stam M.R., Boersma L.J., Vrancken Peeters M., Wouters M., KleinJan E., et al. Transparency in quality of radiotherapy for breast cancer in the Netherlands: a national registration of radiotherapy-parameters. Radiat Oncol. 2022;17(1):73. doi: 10.1186/s13014-022-02043-0.
37. Piroth M.D., Baumann R., Budach W., Dunst J., Feyer P., Fietkau R., et al. Heart toxicity from breast cancer radiotherapy : current findings, assessment, and prevention. Strahlenther Onkol. 2019;195(1):1–12. doi: 10.1007/s00066-018-1378-z
38. Eisen A., Somerfield M.R., Accordino M.K., Blanchette P.S., Clemons M.J., Dhesy-Thind S., et al. Use of adjuvant bisphosphonates and other bone-modifying agents in breast cancer: ASCO-OH (CCO) guideline update. J Clin Oncol. 2022;40(7):787–800. doi: 10.1200/JCO.21.02647.
39. Trayes K.P., Cokenakes S.E.H. Breast cancer treatment. Am Fam Physician. 2021;104(2):171–178.



Chapter 8

Development and Validation of Prediction Models for Health-Related Quality of Life Outcomes After Breast Cancer Surgery and Reconstruction

Elfi M. Verheul, Maria Margarete Karsten, Pimrapat Gebert, Lea Doppelbauer, Simona Borstnar, Sabine Siesling, Hester F. Lingsma, Noëlle J. M. C. Vrancken-Peeters, Anne M. Stiggelbout, Marc A. M. Mureau, Linetta B. Koppert, David van Klaveren, on behalf of the 4D PICTURE Consortium.

Published in: European Journal of Surgical Oncology (EJSO), 2026

ABSTRACT

Background: Predictions of Health-Related Quality of Life (HRQoL) outcomes could support realistic recovery expectations after breast cancer (BC) surgery. We aimed to develop and validate prediction models for HRQoL outcomes after BC surgery.

Methods: We used three datasets of BC patients from Berlin, Germany; Ljubljana, Slovenia; and Rotterdam; Netherlands. We included non-metastasised patients who were surgically treated for an initial diagnosis of BC and completed pre- and postoperative validated questionnaires. We used linear mixed models to analyse 15 domains of the EORTC QLQ-C30 and EORTC QLQ-BR23 over a two-year horizon. Baseline domain score (measured pre-operatively), age, BMI, smoking, TN stage, receptor status, neoadjuvant chemotherapy, axillary surgery and surgery type (breast-conserving, mastectomy, and immediate implant-based reconstruction) were included as predictors. Predictive performance at validation was assessed by the proportion of variance explained (marginal R^2; mR^2).

Results: We included N=795 patients from Germany for development and N=623 from Slovenia and N=417 from Netherlands for validation. The largest proportion of variance was explained by the prediction models for sexual functioning (SF, mR^2 35%), physical functioning (PF, mR^2 29%), body image (BI, mR^2 26%), and cognitive functioning (CF, mR^2 25%). The models captured meaningfully different trends over time for different outcomes and surgery types. The predictive performance of the models was largely driven by the baseline domain score. Performance was reasonable at external validation, with r^2 values of 19–33% for PF, 10–17% for CF, 15–18% for BI, and 22–28% for SF, although some other outcomes (e.g. breast symptoms and role functioning) showed miscalibration, indicating a need for recalibration.

Conclusion: HRQoL after breast cancer surgery can be predicted using simple models with baseline domain scores and surgery type, demonstrating a new opportunity for Patient-Reported Outcome Measures (PROMs) in personalized care.

INTRODUCTION

Breast cancer (BC) is the most commonly diagnosed cancer among women worldwide[1], affecting millions of individuals and their families each year. Its impact extends beyond physical health, influencing various aspects of Health-Related Quality of Life (HRQoL)[2 3]. Although there is no universally accepted definition of HRQoL, it is usually considered a multidimensional concept reflecting an individual's perception of physical, mental, social, and autonomy-related domains[4]. These domains can be measured using Patient-Reported Outcome Measures (PROMs)[5].

Multiple studies have shown that BC patients experience a decline in HRQoL after treatment, with the extent of this decline affected by the type of surgery[2 3]. Patients undergoing breast reconstruction or breast-conserving surgery report higher HRQoL outcomes than those who undergo mastectomy, particularly in the domains of physical health and body image[6 7]. However, HRQoL outcomes may also vary based on patient-related factors, such as age at diagnosis, cancer stage, and baseline HRQoL scores[8 9]. Additionally, individual perspectives and personal priorities significantly shape a patient's overall experience, yet these factors are sometimes overlooked by healthcare professionals[10].

To support both patients and clinicians to set realistic recovery expectations based on individualized information on outcomes, various prediction models have been developed. Many of these models focus on clinical outcomes, such as mortality, recurrence, lymph node involvement, and treatment effects[11 12], and are primarily intended to support decision-making. Our aim, however, is not to compare outcomes, but to provide individualized predictions of HRQoL once the surgical procedure for a breast cancer patient has been chosen. While HRQoL is increasingly recognized as a key outcome for patients, valid models predicting these outcomes remain lacking. Such predictions are urgently needed to help patients form realistic recovery expectations tailored to their specific type of surgery.

We aimed to develop and validate models to predict HRQoL outcomes over time after different types of breast cancer surgery. We used data from different European cohorts to develop and validate prediction models with HRQoL baseline score, patient-, tumour-, and pre-operative treatment characteristics as potential predictors.

METHODS

Data and study population

We utilised three existing retrospective databases including patients from the *PRO routine* program of Charité[13] (Berlin cohort, Germany; Nov 1, 2016 – May 1, 2023; N = 795), a pilot

study on individualized rehabilitation of breast cancer patients of Institute of Ljubljana[14] (Ljubljana cohort, Slovenia; Dec 1, 2019 – March 1, 2023; N = 623) and the value-based healthcare programme of Erasmus MC [15 16] (Rotterdam cohort, Netherlands; Nov 1, 2015 – Jan 1, 2022; N = 417; Figure S1).

The Berlin database was used as development cohort, as it contains a median of six measurements per patient, ranging from the preoperative assessment to a maximum of two years postoperatively. The Ljubljana and Rotterdam databases served as validation cohorts, as they include only three measurements: preoperatively, and at 6 and 12 months postoperatively.

The inclusion criteria for model development were female patients aged ≥18 years with ductal carcinoma in situ (DCIS) and/or invasive breast cancer who underwent surgery and had completed at least one patient-reported outcome measure (PROM) preoperatively and postoperatively. Patients with distant metastases or disease recurrence were excluded. Additionally, patients undergoing autologous reconstruction were excluded due to insufficient sample size in the development cohort. Consequently, we distinguished three types of surgery in our analyses: breast-conserving therapy/oncoplastic reconstruction; mastectomy: and mastectomy with immediate implant-based reconstruction.

Candidate predictors

Candidate predictors that were know pre-operatively were selected, as these are available at the time the tool will be used. This means that adjuvant therapies were excluded. Predictors were chosen based on the literature, expert opinion, and their availability in the databases we reused.

- Patient characteristics: age at surgery, Body Mass Index (BMI), smoking, and preoperative PROM score, which is referred to as the baseline score. For each model, only the baseline score of the corresponding outcome was selected;
- Tumour characteristics: (clinical) T and N stage, receptor status (HER2 positive, HR positive/HER2 negative, or triple negative);
- Treatments: surgery type, neoadjuvant chemotherapy, axillary surgery.

We used pathological T- and N-stage as proxies for clinical T- and N-stage in the Rotterdam validation cohort, because clinical T- and N-stage were not available. We had no access to variables such as socio-economic status. However, we included baseline PRO scores.

Outcome assessments

Health-Related Quality of Life (HRQoL) was assessed using PROMs preoperatively and at various follow-up time points postoperatively. Two different validated PROMs were used: the EORTC QLQ-C30[17], a cancer-generic questionnaire, and the EORTC QLQ-BR23[18], a

breast cancer-specific questionnaire. We selected 15 domains that are relevant for breast cancer surgery:

- *Functional outcomes*: Global health status, Physical functioning, Role functioning, Emotional functioning, Cognitive functioning, Social functioning; (*EORTC QLQ-C30) and* Body image, Sexual functioning, Sexual enjoyment (*EORTC QLQ-BR23);*
- *Symptoms:* Fatigue, Pain, Insomnia (*EORTC QLQ-C30) and* Systemic side effects, Breast symptoms, Arm symptoms (*EORTC QLQ-BR23)*.

All outcomes are scores between 0 and 100, where higher functional domain scores reflect better functioning and quality of life, while higher symptom scores indicate greater symptom severity.

Data analysis

Descriptive statistics were reported for each cohort and each predictor, with means and standard deviations (SD) for continuous variables, and frequencies and percentages for categorical variables. Comparisons of baseline demographic and clinical characteristics between cohorts were performed using χ^2 tests for categorical variables and ANOVA or Kruskal-Wallis tests for continuous variables. We used boxplots to present the distribution of each observed outcome within cohorts at 0, 6, and 12 months. We used single regression-based imputation to complete missing predictor values in each cohort separately. The imputation model incorporated, in addition to the candidate predictors, hormone therapy, adjuvant chemotherapy, targeted therapy, and radiotherapy.

We developed 15 linear mixed models corresponding to 15 domains of the EORTC QLQ-C30 and EORTC QLQ-BR23 questionnaires over a two-year horizon. We included random intercepts to account for patient-level variation. We modelled non-linearity of outcome trends over time with natural splines with 6 knots. Each model included an interaction between time and type of surgery to account for different outcome trends over time across surgery types. Temporal dependence among repeated measurements was modelled using a first-order autoregressive [AR(1)] correlation structure. Since future HRQoL measurements are unavailable for predictions at baseline, the random intercept was excluded when making predictions. The model fit was evaluated using marginal R^2 (mR^2), and predictor importance was measured by the partial mR^2. The strongest predictors were used to fit more parsimonious core models for each outcome.

We used the squared Pearson's correlation coefficient r^2 rather than R^2, which is sensitive to average outcome differences, to measure model performance at apparent and external (Berlin cohort) and external validation (Ljubljana and Rotterdam cohorts) at 6 and 12 months. The Root Mean Squared Error (RMSE) was used to measure the accuracy of the predictions. We used calibration plots of mean observed outcomes versus mean

predicted outcomes in tenths of predicted outcomes, to visualize the alignment between predictions and observed outcomes. Systematic deviations from the 45-degree line were interpreted as potential miscalibration: an intercept greater than 0 or less than 0 indicated systematic over- or underprediction requiring intercept recalibration, a slope less than 1 or greater than 1 indicated over- or underestimation of prediction variability requiring slope recalibration. Furthermore, we used boxplots to examine the variation in observed outcomes within deciles of predicted scores, providing insight into the variability among patients with comparable predicted outcomes.

We developed prediction models following the TRIPOD checklist (Table S1). For all analyses, we used R statistical software version 4.3.2.

RESULTS

Patient characteristics and demographics

In total, 1,835 patients were included (Berlin N=795; Ljubljana N=623; Rotterdam N=417, Table 1; Figure S1). The mean age was 53.3 years (SD = 11.7) with a BMI of 25.4 (SD = 5.0). A higher proportion of advanced T and N stages were observed in the Ljubljana cohort (≥cT2 46%) compared to Berlin (≥cT2 37%) and Rotterdam (≥cT2 37%). Breast-conserving surgery/oncoplastic reconstruction was performed in 73% of the patients in Berlin, which was more than in the cohorts of Rotterdam (57%) and Ljubljana (61%). Rotterdam had a larger mastectomy proportion (30% vs 8% in Berlin and 16% in Ljubljana) and fewer immediate implant-based reconstructions (13% vs 19% in Berlin and 23% in Ljubljana). Furthermore, baseline scores differed between the cohorts in most HRQoL domains, but not in a consistent direction, and no cohort systematically scored higher or lower across domains (Table 1; Figure S2A, S2B).

Model development and performance

The baseline score was by far the strongest predictor, followed by type of surgery (Figure 1). Elapsed time since surgery also played an important role in predicting the trend. For most models the contribution of other baseline predictors was limited (Figure S3, table S2).

The largest proportion of variance was explained by the models for sexual functioning (mR^2 35%), physical functioning (mR^2 29%), body image (mR^2 26%), and cognitive functioning (mR^2 25%). The core models, only including baseline score, type of surgery, and time since surgery, resulted in slightly lower mR^2 values (Table 2). The trend over time since surgery varied between the different types of surgery for each outcome (Figure 2, S4, S5, S6, S7).

Apparent and external validation

Validation at 12 months showed moderate r^2 for most outcomes across all three cohorts (Table 2) indicating that the models were able to explain a proportion of the variance in HRQoL trajectories. However, for some outcomes the r^2 was low (e.g. breast symptoms: 6%, 4%, and 3%). The RMSE ranged from 14 to 38 points on the 100-point scale across all cohorts (Table 2), reflecting the average prediction error for each outcome. Calibration for the validation cohorts was moderate for most outcomes (Figure 3, S8–S10), but for some outcomes recalibration of the intercept, slope, or both could further improve performance. While average observed outcomes aligned well with average predicted outcomes across tenths of predicted values, outcomes at the individual level still showed considerable variation (Figure S8-S10). Validation at 6 months showed similar results to the 12-month predictions (Figure S11-S13).

Table 1: Baseline characteristics of all breast cancer patients included in this study, along with the distribution across the three cohorts: the Berlin development cohort and the validation cohorts from Ljubljana and Rotterdam.

Characteristic	All patients N = 1,835[1]	Berlin N = 795[1]	Ljubljana N = 623[1]	Rotterdam N = 417[1]	p-value[2]
Age at surgery (years) [min, max]	53.3 (11.7) [24-88]	54.1 (12.0) [24-88]	51.8 (9.2) [26-86]	54.0 (14.2) [27-73]	0.001
BMI (kg/m)	25.4 (5.0)	25.1 (4.9)	25.7 (5.3)	25.6 (4.5)	0.017
(missing)	140	47	4	89	
Smoking	251 (15%)	121 (16%)	94 (15%)	36 (10%)	0.025
(missing)	119	45	12	62	
T stage[3]					<0.001
cT0	43 (3%)	0 (0%)	0 (0%)	37 (8.9%)	
cT1	751 (50%)	422 (54%)	329 (53%)	241 (58%)	
cT2/3/4	574 (39%)	288 (37%)	286 (46%)	97 (23%)	
cTis	112 (8%)	70 (9.0%)	6 (1.0%)	42 (10%)	
(missing)	17	15	2	0	
N stage[3]					<0.001
cN0	1,328 (74%)	628 (83%)	386 (62%)	314 (75%)	
cN1/2/3	465 (26%)	129 (17%)	233 (38%)	103 (25%)	
(missing)	40	38	2	0	
Receptor Status					0.7
HER2 positive	219 (13%)	87 (12%)	84 (14%)	48 (12%)	
HR positive and HER2 negative	1,259 (73%)	532 (74%)	447 (74%)	280 (72%)	
triple negative	236 (14%)	101 (14%)	76 (13%)	59 (15%)	
(missing)	121	75	16	30	
Neoadjuvant chemotherapy	478 (26%)	239 (30%)	158 (25%)	81 (19%)	<0.001
Surgery type					<0.001

Table 1: Baseline characteristics of all breast cancer patients included in this study, along with the distribution across the three cohorts: the Berlin development cohort and the validation cohorts from Ljubljana and Rotterdam. (*continued*)

Characteristic	All patients N = 1,835[1]	Berlin N = 795[1]	Ljubljana N = 623[1]	Rotterdam N = 417[1]	p-value[2]
BCS/oncoplastic	1,199 (66%)	582 (73%)	381 (61%)	236 (57%)	
Implant-based reconstruction	342 (19%)	148 (19%)	141 (23%)	53 (13%)	
Mastectomy	285 (16%)	65 (8%)	98 (16%)	122 (30%)	
(missing)	9	0	3	6	
Axillary surgery					<0.001
No	110 (6.0%)	99 (12%)	11 (1.8%)	0 (0%)	
SLNB/RISAS	1,376 (75%)	606 (76%)	448 (72%)	322 (77%)	
ALND	348 (19%)	89 (11%)	164 (26%)	95 (23%)	
(missing)	1	1	0	0	
Overall Quality of life (baseline)	68.6 (21.3)	65.0 (21.1)	68.5 (22.2)	75.6 (18.2)	<0.001
Physical functioning (baseline)	88.4 (16.3)	86.1 (17.8)	91.2 (13.1)	88.8 (16.9)	<0.001
Role functioning (baseline)	85.4 (23.7)	83.5 (25.3)	89.2 (19.8)	83.1 (25.3)	<0.001
Emotional functioning (baseline)	65.6 (25.4)	56.7 (26.3)	71.9 (23.0)	73.2 (21.5)	<0.001
Cognitive functioning (baseline)	84.3 (21.1)	80.7 (23.6)	88.0 (18.2)	85.6 (18.8)	<0.001
Social functioning (baseline)	81.9 (24.7)	75.1 (27.8)	87.5 (19.8)	86.4 (21.9)	<0.001
Body image (baseline)	86.7 (20.7)	81.7 (23.4)	91.6 (16.7)	88.7 (18.4)	<0.001
(missing)[4]	55	37	2	16	
Sexual functioning (baseline)	27.3 (25.9)	27.9 (28.2)	28.6 (23.9)	24.3 (24.1)	0.011
(missing)[4]	68	37	6	25	
Sexual enjoyment (baseline)	51.6 (35.6)	68.0 (27.2)	56.7 (30.9)	30.2 (36.9)	<0.001
(missing)[4]	612	392	209	11	
Fatigue (baseline)	25.3 (24.6)	30.8 (27.3)	18.9 (19.9)	24.2 (23.0)	<0.001
Pain (baseline)	18.1 (23.4)	24.2 (26.4)	12.5 (19.2)	14.6 (20.2)	<0.001
Insomnia (baseline)	32.6 (31.4)	38.6 (33.7)	27.9 (29.1)	28.1 (28.3)	<0.001
(missing)[4]	1	0	1	0	
Systemic therapy side effects (baseline)	14.0 (15.7)	18.3 (17.9)	11.2 (12.3)	10.2 (13.7)	<0.001
(missing)[4]	51	37	1	13	

Table 1: Baseline characteristics of all breast cancer patients included in this study, along with the distribution across the three cohorts: the Berlin development cohort and the validation cohorts from Ljubljana and Rotterdam. (*continued*)

Characteristic	All patients N = 1,835[1]	Berlin N = 795[1]	Ljubljana N = 623[1]	Rotterdam N = 417[1]	p-value[2]
Breast symptoms (baseline)	16.0 (18.8)	17.9 (20.2)	15.6 (19.1)	13.0 (14.8)	0.002
(missing)[4]	49	37	1	11	
Arm symptoms (baseline)	11.1 (18.1)	14.6 (20.9)	8.8 (15.2)	8.1 (15.3)	<0.001
(missing)[4]	49	37	1	11	

BMI = Body Mass Index; HER2 = Human Epidermal growth factor Receptor 2; HR = Hormone Receptor; Triple negative = negative for hormone receptors and HER2; BCS = Breast Conserving Surgery; SLNB = Sentinel Lymph Node Biopsy; ALND = Axillary Lymph Node Dissection; RISAS = Radioactive Iodine Seed Localization, [1]Values are n (%), unless for age, BMI, and all baseline scores, which are presented as mean (standard deviation); [2]P-values compare baseline characteristics between the three cohorts. The "All patients" column is descriptive only.[3]In the Rotterdam cohort, the pathological T and N stage are presented; [4]If the baseline score (score measured pre-operatively) for an outcome is missing, the patient will be excluded from the development or validation of the corresponding model.

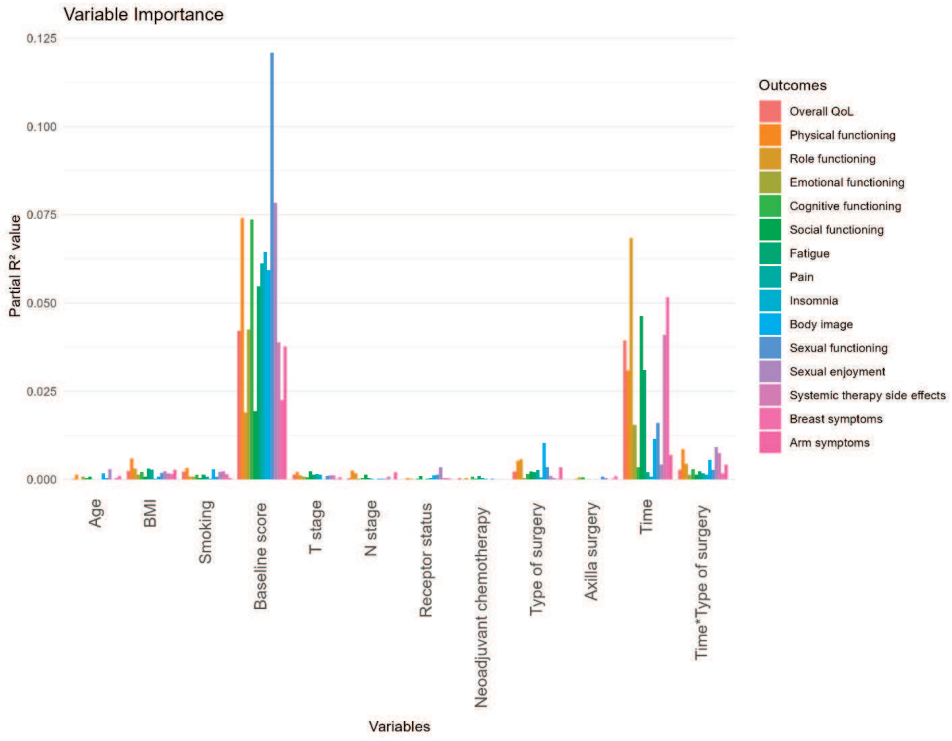


Figure 1: The contribution of each variable to the model's performance measured by the partial mR^2. The partial mR^2 represents the proportion of variance in the outcome explained by each predictor. *QoL = Quality of life, Time*Type of surgery reflects the interaction between time and the type of surgery in the models.*

Table 2: Performance and validation measures for the full and core models in the development cohort (Berlin) and validation cohorts (Ljubljana and Rotterdam). *The mR^2 represents the marginal R^2 of the full and core model and indicates the proportion of variance explained by the predictor variables (fixed effects) in each model. The core model includes baseline score, time, type of surgery and the interaction of time and type of surgery. T12 reflects the timepoint at 12 months (1 year). For this timepoint, we present the squared correlation coefficient (cor^2), which we will refer to as r^2, and Root Mean Square Error (RMSE).*

	Development cohort: Berlin				**Validation cohort 1: Ljubljana**		**Validation cohort 2: Rotterdam**	
Functional scales	**mR^2 full**	**mR^2 core**	**r^2T12**	**RMSE T12**	**r^2T12**	**RMSE T12**	**r^2T12**	**RMSE T12**
Global health status	17%	16%	14%	19	9%	21	13%	18
Physical functioning	29%	25%	25%	16	19%	15	33%	13
Role functioning	15%	12%	13%	26	3%	25	11%	26
Emotional functioning	17%	16%	17%	23	14%	22	13%	21
Cognitive functioning	25%	24%	24%	23	10%	22	17%	20
Social functioning	13%	10%	8%	27	9%	23	11%	22
Body image	26%	24%	25%	25	15%	22	18%	23
Sexual functioning	35%	35%	28%	24	22%	20	28%	20
Sexual enjoyment	21%	20%	10%	25	15%	*26*	21%	38
Symptom scales								
Fatigue	21%	19%	20%	24	14%	23	23%	22
Pain	21%	19%	19%	26	11%	24	13%	24
Insomnia	21%	21%	22%	30	5%	34	11%	29
Systemic side effects	17%	16%	12%	16	9%	14	12%	14
Breast symptoms	12%	11%	6%	20	4%	20	3%	19
Arm symptoms	15%	13%	12%	22	9%	20	12%	20

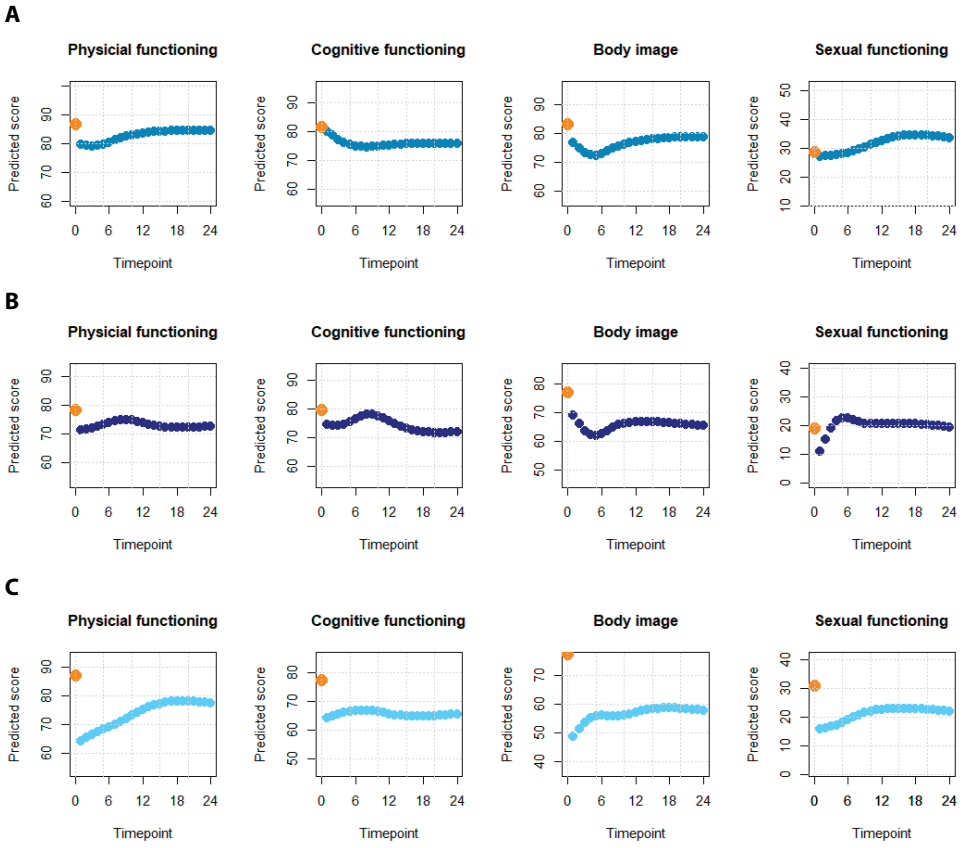


Figure 2: Predictions from the core models, using baseline score and surgery type, for four Health-Related Quality of Life outcomes in an average patient undergoing (A) breast-conserving with or without oncoplastic surgery, (B) mastectomy, or (C) mastectomy with immediate implant-based reconstruction. *For each surgery type, the model uses the mean baseline score across patients in that group. As the distribution of baseline scores differs across groups, the baseline score of the average patient and the corresponding predicted values differ accordingly. These predictions illustrate expected outcomes over 24 months for representative patients with typical baseline values: (A) PF: 87, CF: 82, BI: 83, SEF: 28; (B) PF: 79, CF: 79, BI: 77, SEF: 19; (C) PF: 87, CF: 77, BI: 78, SEF: 31. The orange dots indicate the baseline scores, while blue dots represent predicted scores from 1 to 24 months after surgery.*

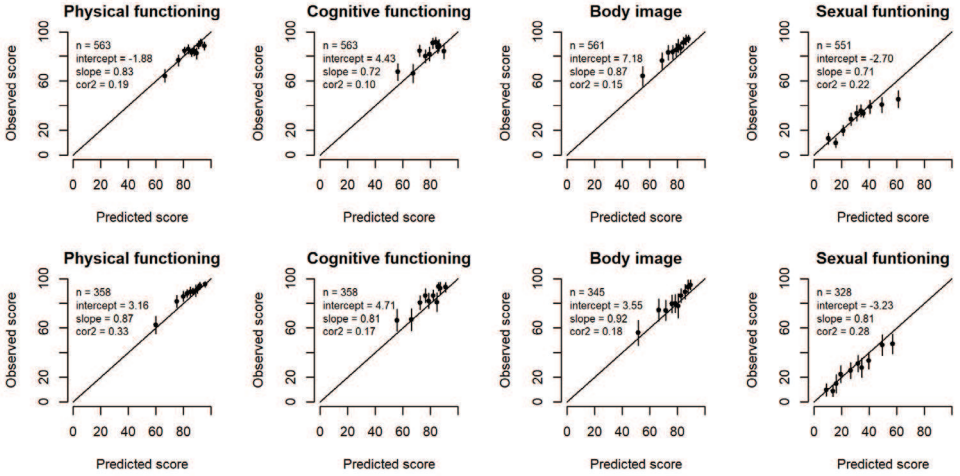


Figure 3: Calibration plots at the 12-month timepoint for the Ljubljana validation cohort (top row) and the Rotterdam validation cohort (bottom row). *The x-axis represents the predicted scores, and the y-axis represents the observed scores. The diagonal line indicates perfect calibration. Ideally, the intercept equals 0 and the slope equals 1. An intercept > 0 suggests underestimation, while an intercept < 0 suggests overestimation. A slope < 1 indicates underfitting (predictions are too close to the mean, overestimating low scores and underestimating high scores), whereas a slope > 1 indicates overfitting (predictions are too extreme, underestimating low scores and overestimating high scores). N denotes the number of patients; and r^2 is the squared Pearson correlation between predicted and observed values.*

DISCUSSION

Summary of main results

We developed 15 models to predict trends in Health-Related Quality of Life (HRQoL) outcomes over the elapsed time since surgery, up to two years following breast cancer surgery. The baseline HRQoL score was by far the strongest predictor. With simple models that included only the baseline score and type of surgery, we were able to predict the HRQoL outcomes at different time points over a two-year horizon, particularly for sexual functioning, physical functioning, body image, and cognitive functioning. The models showed reasonable predictive performance across most outcomes, also in external cohorts, supporting their potential for use in clinical expectation management.

Results in the context of published literature and implications for clinical practice

HRQoL predictions integrating repeated measurements are relatively new in the field[9]. Most existing HRQoL models either use a binary cut-off or predict a single timepoint[19 20]. A prostate cancer study developed a tool incorporating repeated measurements across five functional domains (e.g., sexual functioning, urinary incontinence) and reported

R^2 values between 20% and 39%[21], which is comparable to most of our models. Their variability, as indicated by root mean squared error (RMSE), was also consistent with our findings. In psychosocial sciences, an R^2 between 10% and 50% is considered acceptable, as long as certain predictors show statistical significance[22]. These moderate level of performance reflect the inherent complexity of HRQoL's multidimensional nature[23], making accurate prediction a challenge. Most of our models are acceptable, but not all outcomes are predictable. For example, the model for breast symptoms with squared correlations of 6%, 4%, and 3%, respectively, in the Berlin, Ljubljana, and Rotterdam cohort at 12 months, suggests that this outcome cannot be reliably predicted using the variables included in the model. Several factors may explain the limited predictive ability of certain domains. First, these outcomes may exhibit substantial inter-individual variability that is not captured by the included predictors alone. Second, postoperative complications, which are known to strongly influence these domains[24 25], were not available in our datasets. Finally, some outcomes, such as insomnia, were measured with a single item, which may have introduced additional measurement error.

Dorr et al. showed that performance can be improved with dynamic predictions[26], which means the inclusion of new PRO measurements collected after baseline. However, this approach is inconsistent with the aim of our model, which is to provide personalized predictions regarding HRQoL outcomes before the surgery is actually performed. Additionally, Dorr et al. found that including additional clinical and demographic variables did not improve predictions[26]. Helmrich et al. reported that patient characteristics (e.g., pre-operative mental status, psychosocial factors) are more important than disease-related characteristics[19]. These findings are reflected in the comparable mR^2 values between the full and core models, which only include baseline scores and surgery type. The strong predictive role of baseline HRQoL likely reflects that it captures individual differences in perceived health and coping that clinical or demographic variables do not. Baseline scores therefore function not only as a measure of preoperative status but also as a proxy for stable psychological and contextual factors, which may explain why adding further variables offered little additional predictive value. Consistent with our findings, previous studies have shown that baseline measurements are the strongest predictors of postoperative HRQoL outcomes, including baseline fatigue for predicting post-treatment fatigue[20 27] and baseline anxiety and depression for explaining more variance in quality of life than treatment itself[28]. From a practical point of view, we argue that relying only on the core variables used in our simple models improves usability for both patients and physicians. This simplification reduces the number of required variables, minimizing data collection effort and time, while the performance decline remains marginal.

On average, calibration plots showed that the models performed well. However, due to variability between patients, there is a risk of informing patients either too optimistically or too pessimistically. This concern is particularly relevant for patients with very high or

very low baseline scores. Since the baseline score is included as a predictor, patients with extreme baseline values tend to be slightly adjusted toward the mean (Figure S7). Clinically, this miscalibration may lead to unrealistic expectations if predictions are taken too literally, highlighting the importance of carefully interpreting prediction uncertainty, especially for patients at the extremes of the baseline range.

In addition to within-cohort heterogeneity, we also observed notable heterogeneity between cohorts, as reflected in the external calibration plots. For example, body image was underestimated in both validation cohorts due to higher observed scores compared to the development cohort. These baseline differences were substantial and not fully captured by the model predictions. These differences may relate to variations in personal values, cultural norms, healthcare organization, and PROMs implementation[29 30]. As we lacked information on cultural background, ethnicity, and individual value systems, we could not further assess these factors. Nevertheless, the Reborn study in Japanese women reported patterns of postoperative HRQoL that differ from those described in several Western cohorts[31]. These findings underscore the importance of cultural differences and suggest that our models, developed and validated in predominantly European populations, may not be directly generalizable to more diverse non-European populations.

The prediction models are developed for expectation management. Only variables that are known with certainty before surgery are included as predictors to ensure usability. Therefore, neoadjuvant chemotherapy is included, whereas adjuvant treatments such as radiotherapy, hormone therapy, and adjuvant chemotherapy are not. Once the type of surgery has been selected, the models can be used in consultation with the patient, preferably by a healthcare provider familiar with their interpretation, to discuss predicted outcomes and help set realistic expectations for postoperative recovery. However, the models are not suitable for comparing surgery types across patients, since confounding by indication leads to non-comparable groups and limits the validity of outcome comparisons.

Strengths and limitations

The strengths of this study include a robust, relatively new methodological approach with repeated HRQoL measurements. Access to three international cohorts enabled external validation, which was often lacking in previous research[9]. Furthermore, adherence to TRIPOD guidelines ensured transparency and rigor. Through international collaboration and data reuse, we included sufficiently large samples across three surgery types. Although reusing data limited our flexibility in selecting predictors, both our core model and the existing literature indicate that including additional tumour-related variables does not improve model performance[19 21 26]. For the Rotterdam cohort, pathological T and N staging was used instead of clinical staging. Neoadjuvant therapy was included as a predictor, and among patients who did not receive it, pathological and clinical staging are expected

to differ in only ~30% of cases. Given the ~70% concordance and the minimal impact of tumour-related predictors, this discrepancy is considered negligible. The outcomes of insomnia and sexual enjoyment were based on a single question with a 4-point Likert scale. Given the methodological complexity and interpretation challenges of our modelling work, we pragmatically applied linear mixed models to these outcomes rather than a proportional odds model. Although multiple imputation is generally preferred for handling missing data, we opted for single imputation as a pragmatic and acceptable alternative[32].

Notably, the cohorts are not fully representative of their respective countries, as inclusion was limited to non-recurrent breast cancer patients treated in university hospitals. For example, the mean age in the Rotterdam cohort was 54 years, and only 57% underwent breast-conserving surgery, compared to a national mean age of 65 and a breast-conserving rate slightly above 70%[33]. Additionally, in the Berlin cohort non-adherence to PROMs was assessed, which was associated with younger age, more comorbidities, no chemotherapy, and lower baseline physical functioning[34], further limiting representativeness of the full population.

Future Research

HRQoL outcomes are strongly influenced by psychosocial factors and personality traits[35]. Therefore, future research should focus on incorporating these patient-related predictors into models to determine whether they improve model performance. Additionally, it may be valuable to identify when a patient falls into a high-risk group for a specific outcome, as this could guide the need for interventions. Specifically for BC, more data is needed on immediate autologous and delayed reconstructions to include these patient groups in the models as well.

CONCLUSION

We developed 15 prediction models for different HRQoL domains after BC surgery over a two-year horizon. We found that baseline score is the most important predictor of the trajectory over time and needs to be taken into account when discussing postoperative expectations. As PROMs are increasingly integrated into routine clinical practice, we demonstrate that using baseline scores in a simple model presents an opportunity to improve personalized care.

Supplementary Information

Supplementary material is available at: https://www.ejso.com/article/S0748-7983(26)00085-5/fulltext

4D PICTURE Consortium:

Participants of the 4D picture consortium: **Marieke Bak** (Institute for History and Ethics of Medicine, Technical University of Munich, Munich, Germany), **Liesbeth C. de Wreede** (Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Kasper Frank** (Center for Shared Decision Making, Vejle/Lillebaelt University Hospital of Southern Denmark), **Laura Hartman** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands), **Elham Hedayati** (Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden), **Toby Hackmann** (Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Lára R. Hallsson** (Institute of Public Health, Medical Decision Making and Health Technology Assessment, UMIT TIROL – University for Health Sciences and Technology, Hall in Tirol, Austria), **Yassin Engelberts** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands), **Yufeng Liu** (Department of Linguistics and English Language, Lancaster University, Lancaster, United Kingdom), **Daisy Monika Lal** (School of Computing and Communications, Lancaster University, Lancaster, United Kingdom), **Ida J. Korfage** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands), **Amara Callistus Nwosu** (International Observatory on End of Life Care, and Department of Palliative Care, Lancaster Medical School, Lancaster University, Lancaster, United Kingdom; Marie Curie Hospice Liverpool, Liverpool, United Kingdom), **Claudia Cruz Oliveira** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands), **Barbara Perić** (Institute of Oncology Ljubljana, and Medical Faculty, University of Ljubljana, Ljubljana, Slovenia), **Roberto Pazo-Cid** (Department of Medical Oncology, Miguel Servet University Hospital, Facultad de Medicina, Universidad de Zaragoza, Zaragoza, Spain; Instituto de Investigación Sanitaria Aragón (IIS-A); Hospital Universitario Miguel Servet, Zaragoza, Spain), **Carlos Romero-Piqueras** (Fractal Strategy, Zaragoza, Spain), **Judith Rietjens** (Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands; Department of Design, Organisation and Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, the Netherlands), **Dirk Snelders** (Department of Design, Organisation and Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, the Netherlands), **Uwe Siebert** (Institute of Public Health, Medical Decision Making and Health Technology Assessment, UMIT TIROL – University for Health Sciences and Technology, Hall in Tirol, Austria; Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard T.H. Chan School of Public Health; MGH-ITA, Department of Radiology, Harvard Medical School, Boston, MA, USA), **Gaby Sroczynski** (Institute of Public Health, Medical Decision Making and Health Technology Assessment, UMIT TIROL – University for Health Sciences and Technology, Hall in Tirol, Austria), **Ewout Steyerberg** (Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands), **Erik van Mulligen** (Department of Medical Informatics, Erasmus MC,

University Medical Center Rotterdam, Rotterdam, the Netherlands), **Tamara Wit** (Medical Decision Making, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands), **Michel W. Wouters** (Medical Decision Making, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands).

REFERENCES

1. Organization WH. Global breast cancer initiative implementation framework: assessing, strengthening and scaling-up of services for the early detection and management of breast cancer. Executive summary. Geneva; 2023.
2. Mokhtari-Hessari P, Montazeri A. Health-related quality of life in breast cancer patients: review of reviews from 2008 to 2018. Health Qual Life Outcomes. 2020;18(1):338.
3. Grasic Kuhar C, Gortnar Cepeda T, Kurzeder C, Vetter M. Changes in the quality of life of early breast cancer patients and comparison with the normative Slovenian population. Radiol Oncol. 2023;57(2):211-9.
4. Hinz A, Zenger M, Schmalbach B, Brahler E, Hofmeister D, Petrowski K. Quality of Life Domains in Breast Cancer Survivors: The Relationship Between Importance and Satisfaction Ratings. Front Psychol. 2022;13:923537.
5. Salas M, Mordin M, Castro C, Islam Z, Tu N, Hackshaw MD. Health-related quality of life in women with breast cancer: a review of measures. BMC Cancer. 2022;22(1):66.
6. Zehra S, Doyle F, Barry M, Walsh S, Kell MR. Health-related quality of life following breast reconstruction compared to total mastectomy and breast-conserving surgery among breast cancer survivors: a systematic review and meta-analysis. Breast Cancer. 2020;27(4):534-66.
7. Ng ET, Ang RZ, Tran BX, Ho CS, Zhang Z, Tan W, et al. Comparing Quality of Life in Breast Cancer Patients Who Underwent Mastectomy Versus Breast-Conserving Surgery: A Meta-Analysis. Int J Environ Res Public Health. 2019;16(24).
8. Janz NK, Mujahid MS, Hawley ST, Griggs JJ, Alderman A, Hamilton AS, et al. Racial/ethnic differences in quality of life after diagnosis of breast cancer. J Cancer Surviv. 2009;3(4):212-22.
9. van der Heijden TGW, de Ligt KM, Hubel NJ, van der Mierden S, Holzner B, van de Poll-Franse LV, et al. Exploring the role of health-related quality of life measures in predictive modelling for oncology: a systematic review. Qual Life Res. 2025;34(2):305-23.
10. Boman LE, Sandelin K, Wengstrom Y, Silen C. Patients' participation during treatment and care of breast cancer - a possibility and an imperative. Eur J Oncol Nurs. 2018;37:35-42.
11. Min N, Wei Y, Zheng Y, Li X. Advancement of prognostic models in breast cancer: a narrative review. Gland Surg. 2021;10(9):2815-31.
12. Hueting TA, van Maaren MC, Hendriks MP, Koffijberg H, Siesling S. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. J Clin Epidemiol. 2022;152:238-47.
13. Karsten MM, Speiser D, Hartmann C, Zeuschner N, Lippold K, Kiver V, et al. Web-Based Patient-Reported Outcomes Using the International Consortium for Health Outcome Measurement Dataset in a Major German University Hospital: Observational Study. JMIR Cancer. 2018;4(2):e11373.
14. Bešić N. Pilot study on individualized integrated rehabilitation of breast cancer patients 2019-2022.: ARIS; [Available from: https://cris.cobiss.net/ecris/si/en/project/17955.
15. Vrancken Peeters N, Kaplan ZLR, Clarijs ME, Mureau MAM, Verhoef C, van Dalen T, et al. Health-related quality of life (HRQoL) after different axillary treatments in women with breast cancer: a 1-year longitudinal cohort study. Qual Life Res. 2024;33(2):467-79.
16. van Egdom LSE, Lagendijk M, van der Kemp MH, van Dam JH, Mureau MAM, Hazelzet JA, Koppert LB. Implementation of Value Based Breast Cancer Care. Eur J Surg Oncol. 2019;45(7):1163-70.
17. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365-76.

18. Sprangers MA, Groenvold M, Arraras JI, Franklin J, te Velde A, Muller M, et al. The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study. J Clin Oncol. 1996;14(10):2756-68.
19. Helmrich I, van Klaveren D, Dijkland SA, Lingsma HF, Polinder S, Wilson L, et al. Development of prognostic models for Health-Related Quality of Life following traumatic brain injury. Qual Life Res. 2022;31(2):451-71.
20. Adiprakoso D, Katsimpokis D, Oerlemans S, Ezendam NPM, van Maaren MC, van Til JA, et al. Development of a prediction model for clinically-relevant fatigue: a multi-cancer approach. Qual Life Res. 2025;34(1):231-45.
21. Laviana AA, Zhao Z, Huang LC, Koyama T, Conwill R, Hoffman K, et al. Development and Internal Validation of a Web-based Tool to Predict Sexual, Urinary, and Bowel Function Longitudinally After Radiation Therapy, Surgery, or Observation. Eur Urol. 2020;78(2):248-55.
22. Ozili PK. The acceptable R-square in empirical modelling for social science research. MPRA Paper 2023;115769.
23. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. J Nurs Scholarsh. 2005;37(4):336-42.
24. Kouwenberg CAE, de Ligt KM, Kranenburg LW, Rakhorst H, de Leeuw D, Siesling S, et al. Long-Term Health-Related Quality of Life after Four Common Surgical Treatment Options for Breast Cancer and the Effect of Complications: A Retrospective Patient-Reported Survey among 1871 Patients. Plast Reconstr Surg. 2020;146(1):1-13.
25. Gopie JP, Timman R, Hilhorst MT, Hofer SO, Mureau MA, Tibben A. The short-term psychological impact of complications after breast reconstruction. Psychooncology. 2013;22(2):290-8.
26. Dorr MC, Andrinopoulou ER, Sewnaik A, Berzenji D, van Hof KS, Dronkers EAC, et al. Individualized Dynamic Prediction Model for Patient-Reported Voice Quality in Early-Stage Glottic Cancer. Otolaryngol Head Neck Surg. 2024;170(1):169-78.
27. Courtier N, Gambling T, Enright S, Barrett-Lee P, Abraham J, Mason MD. A prognostic tool to predict fatigue in women with early-stage breast cancer undergoing radiotherapy. Breast. 2013;22(4):504-9.
28. Harris J, Purssell E, Cornelius V, Ream E, Jones A, Armes J. Development and internal validation of a predictive risk model for anxiety after completion of treatment for early stage breast cancer. J Patient Rep Outcomes. 2020;4(1):103.
29. Chang EM, Gillespie EF, Shaverdian N. Truthfulness in patient-reported outcomes: factors affecting patients' responses and impact on data quality. Patient Relat Outcome Meas. 2019;10:171-86.
30. Anderson M, van Kessel R, Wood E, Stokes A, Fistein J, Porter I, et al. Understanding factors impacting patient-reported outcome measures integration in routine clinical practice: an umbrella review. Qual Life Res. 2024;33(10):2611-29.
31. Seki H, Komiya T, Sowa Y, Kato M, Nishida Y, Isaka H, et al. Breast satisfaction and health-related quality of life following total mastectomy, breast-conserving surgery, or immediate breast reconstruction in Japanese patients with breast cancer: multicentre cross-sectional controlled study (Reborn). BJS Open. 2025;9(5).
32. van Linschoten RCA, Amini M, van Leeuwen N, Eijkenaar F, den Hartog SJ, Nederkoorn PJ, et al. Handling missing values in the analysis of between-hospital differences in ordinal and dichotomous outcomes: a simulation study. BMJ Qual Saf. 2023;32(12):742-9.
33. Verheul EM, van Klaveren D, Lingsma HF, Vos E, Hoornweg MJ, Siesling S, et al. High-impact complications after breast cancer surgery in the Dutch national quality registry: evaluating case-mix adjustment for hospital comparisons. BJS Open. 2024;9(1).

34. Gebert P, Hage AM, Blohmer JU, Roehle R, Karsten MM. Longitudinal assessment of real-world patient adherence: a 12-month electronic patient-reported outcomes follow-up of women with early breast cancer undergoing treatment. Support Care Cancer. 2024;32(6):344.
35. Wintraecken VM, Vulik S, de Wild S, Dirksen C, Koppert LB, de Vries J, Smidt ML. A descriptive systematic review of the relationship between personality traits and quality of life of women with non-metastatic breast cancer. BMC Cancer. 2022;22(1):426.

9

Chapter 9

General Discussion

MAIN FINDINGS

The overall aim of this thesis was to assess how clinical and patient-reported outcome data can be used to improve the quality of cancer care in two ways: (1) by using outcome data for continuous and reliable comparison of hospital performance to drive quality improvement, and (2) by developing prediction models to support personalized care and shared decision-making. This final chapter reflects on the main findings (Text box 1), explores their implications for policy and practice, and outlines priorities for future research.

Textbox 1: Main findings per research question

What is currently known about cancer outcome performance comparison and the use of case-mix adjustment models in European countries?
Performance comparison between hospitals (i.e. benchmarking) is performed in a few countries, with a predominant focus on short-term clinical outcomes, often without case-mix adjustment. There is no standardization of case-mix adjustment models and only limited transparency regarding their application.

How can we assess if quality indicators are valid and reliable? Are there differences in benchmarking hospital performance between high- and low-incidence cancers?
We considered case-mix adjustment as a key aspect of validity and highlighted the need to assess the validity of each Quality Indicator (QI) individually. Reliability is a concern in both low-incidence diseases and high-incidence diseases (due to the large number of hospitals and often subgroups within QIs), requiring multi-year or aggregated reporting for public QIs. We provided a framework for assessing individual QIs on their validity and reliability.

Is the most widely used prediction model in breast cancer care valid for guiding personalised, evidence-based treatment decisions across different geographical regions and specific patient groups?
BREAST Predict, which is a widely used model in breast cancer care, showed good performance in the general population and in patients with lobular breast cancer in Sweden and the Netherlands. Comparison of *BREAST Predict* versions 2.2 and 3.1 in the Dutch population confirmed the consistent performance of the model. Improved modelling of age effects may further enhance its predictive accuracy for younger patients.

Can we develop models to predict individual trajectories of health-related quality of life (HRQoL)?
Baseline Patient-Reported Outcome (PRO) scores are the strongest predictor of HRQoL trajectories, showing potential for personalised expectation management. We developed 15 models predicting HRQoL trajectories over the elapsed time since surgery, up to two years following breast cancer surgery. The models showed reasonable predictive performance across most outcomes, also in external cohorts.

PART I: USING OUTCOME INFORMATION TO STIMULATE CONTINUOUS QUALITY IMPROVEMENT

Availability and standardisation of outcome data

Valid quality indicators (QIs), based on standardized data definitions, are essential for fair hospital comparisons both nationally and internationally[1-4]. Our review (Chapter 2) showed that long-term and patient-reported outcomes are underrepresented in current quality indicators. As a result, current assessments capture only part of cancer care quality. This finding is supported by a previous review of 89 breast cancer quality indicators across 22 European and North American documents, which additionally demonstrated a substantial heterogeneity in the selection and definitions of QI outcomes between countries [5]. Although professional initiatives such as the European Society of Breast Cancer Specialists (EUSOMA)[6] and policy-led initiatives such as the European Commission Initiative on Breast Cancer (ECIBC)[7 8] have proposed standardized QI sets, their implementation into national quality programmes and clinical audits remains limited.

Across regions, cancer quality programmes differ substantially in scope, data infrastructure and maturity, ranging from well-established audit systems in Western Europe and Oceania to early-stage frameworks in Latin America and Asia[9]. Less mature programmes can learn from established systems, avoid reinventing the wheel, and build directly on existing harmonized outcome sets, such as those developed by the International Consortium for Health Outcomes Measurement (ICHOM)[10]. A concrete example of international collaboration is provided by the breast implant registries working together under the umbrella of 'International Collaboration of Breast Registry Activities' (ICOBRA)[11]. These registries regularly exchange experiences on governance and data processes, enabling mutual learning, avoiding duplication of effort and aligning outcome definitions across countries[12].

Case-mix adjustment

We showed that there is no consensus on how to develop, evaluate and report case-mix models. This lack of standardization undermines both the validity of comparisons be-

tween hospitals and the possibility of benchmarking across countries. When case mix has an impact, the extent of this impact differs per QI. In general, short-term outcomes such as complications are less sensitive to case-mix variation than long-term outcomes such as mortality[13]. The frequent inclusion of treatment variables in case-mix models, without strong scientific justification (Chapters 2 and 3), raises questions about whether QI scores represent quality of care. Ideally, treatment-related variables should not be included in the case-mix adjustment model, as they can potentially mask aspects of quality of care. For example, adjusting for chemotherapy administration can mask variations in guideline adherence. Chapter 3 illustrates this dilemma for the QI "complications after surgery", where we in the research team extensively discussed whether surgery type should be included as a case-mix variable. Including surgery type may mask differences in surgical decision-making and thus aspects of quality of care. However, adjustment can be justified when the aim is to compare complication rates within each surgery type, provided that the choice of surgery is evaluated separately through another QI. This example underlines that case-mix adjustment cannot follow a one-size-fits-all approach: the selection of variables and the appropriateness of adjustment must be carefully evaluated for each individual indicator. Equally important is the transparent evaluation and reporting of model performance and its impact at the hospital level, which is essential for fair interpretation and appropriate use of benchmarking results. Chapter 3 demonstrates that the effect of case-mix adjustment is negligible: variables that initially appeared to influence complications were evenly distributed across hospitals and did not affect comparative performance. This highlights the need to assess the impact of case-mix adjustment at the hospital level, in addition to overall model performance. Consequently, there is a clear need for international guidelines to ensure case-mix evaluations are scientifically robust and transparent, enabling benchmarking across indicators, hospitals, and countries.

Balancing validity and reliability

The QI "complications after surgery" in breast cancer patients illustrates in chapter 4 the tension between validity (defined as case-mix adjustment) and reliability. In addition to being valid, QIs must also be reliable, meaning that observed differences should reflect true differences in quality of care rather than random variation[3 14]. Choosing a more detailed level (e.g. stratification) can increase validity and accountability but reduces reliability due to smaller sample sizes[15]. Chapter 4 illustrates this trade-off: stratification by surgery type improved validity, as outcomes were compared within more homogeneous patient groups, but reduced reliability due to smaller patient numbers per hospital. If QI are not reliable, they should not be transparently published in their current form, as the differences between hospitals are related to noise, and therefore misleading the stakeholder. This was already pointed out a decade ago [1 2 14], but awareness of this is still lacking in current practice.

Improving the reliability

We proposed reporting at an aggregated level, with a focus on multi-year reporting, as a solution to improve reliability. Other forms of aggregation are also possible when relevant, such as clustering hospitals into regions[16]. To support the principle that the trade-off between validity and reliability must be evaluated for every QI, I developed a practical guide in chapter 4 with four measurable criteria: feasibility, discriminative ability, validity, and reliability. Feasibility (data availability) and discriminative ability (whether there are observed differences between hospitals) are important prerequisites, but the main emphasis lies on validity and reliability. The guide provides a hands-on approach for assessing QIs and enables informed decisions about this balance. I applied it to both a high-incidence disease (breast cancer, Chapter 4) and a low-incidence disease (oral cavity cancer, Chapter 5), and showed that reliability is a recurring concern. Combined with the sample size calculations in Chapter 5, the guide can be used to determine the minimum number of years needed for a QI to achieve sufficient reliability.

Reliability is crucial when detecting outliers (i.e., hospitals that perform significantly better or worse). The potential for Type I errors (false positives) and Type II errors (false negatives) must be considered [17,18]. A Type I error occurs when a hospital is falsely labelled as an outlier even though it is not, which can cause reputational harm and misdirect improvement efforts. A Type II error occurs when a poorly performing hospital is not identified, delaying potential interventions. Lower reliability increases the risk of both type 1 error and type 2 error. Random effects models offer an additional statistical approach to estimate hospital performance in a more conservative way, especially for smaller hospitals, by "shrinking" their performance estimates toward the overall mean[19]. While this may mask quality of care differences in smaller hospitals, it also decreases the risk of falsely labelling hospitals as outliers. For external reporting, this is particularly important, as the consequences of Type I errors are severe for hospitals.

Transparent reporting of QIs

When QIs are valid and reliable, they must be presented in a way that is understandable for stakeholders. In the Netherlands, QIs are still often presented in public reports as point estimates without an explanation or visualisation of uncertainty (Figure 1) [20]. To address this, the SKI-tool (Spiegelinformatie Kwaliteitsindicatoren), developed by Dutch Hospital Data (DHD) and financed by the Dutch Healthcare Institute (Zorginstituut Nederland), has made publicly available national QIs easier accessible [21,22]. This tool allows users to search by condition or treatment and provides visualisations, such as funnel plots (Figure 2 in the introduction), that enable benchmarking of hospitals and clinics at both national and regional level. Although this represents an important step forward, challenges remain. Users must familiarise themselves with a new platform, and interpretation can still be difficult. Funnel plots are a good way to make statistical uncertainty visible[23,24], but they can be easily misinterpreted without adequate knowledge and explanation[25]. Addition-

ally, the tool also includes bar charts, which are easier to understand, but do not display uncertainty and may therefore be misleading.

With the current push for transparent reporting, as also emphasised in the Dutch outcome-oriented care agreement (UGZ) of the Ministry of Health (VWS)[26], it is increasingly important that publicly reported QIs are based on rigorous, transparent methodology and presented in a clear and understandable way. Ideally, visualisation and communication should be tailored to different stakeholder groups (e.g. clinicians, policymakers, and patients), each of whom requires different levels of contextual explanation and statistical detail.

B	C	N	O	Q	S	T	U	W
Verslag	Organisatie	Indicatorset	Thema	Indicator	IndicatorW	Teller	Noemer	Indicator_nvt
2023	Ziekenhuis Rijnstate	Hoofd-hals	Ongeplande heroperatie	percentage	0.00	0	44	0
2023	Medisch Centrum Leeuwarden	Hoofd-hals	Ongeplande heroperatie	percentage	5.40	3	54	0
2023	UMC Groningen	Hoofd-hals	Ongeplande heroperatie	percentage	6.20	6	104	0
2023	Maastricht UMC, Maastricht	Hoofd-hals	Ongeplande heroperatie	percentage	4.60	4	85	0
2023	Franciscus Gasthuis & Vlietland	Hoofd-hals	Ongeplande heroperatie	percentage				1
2023	Erasmus MC	Hoofd-hals	Ongeplande heroperatie	percentage	5.10	8	148	0
2023	Haaglanden Medisch Centrum	Hoofd-hals	Ongeplande heroperatie	percentage	5.10	5	40	0
2023	HagaZiekenhuis locatie Den Haag	Hoofd-hals	Ongeplande heroperatie	percentage				1
2023	LUMC	Hoofd-hals	Ongeplande heroperatie	percentage	3.70	2	56	0
2023	UMC Utrecht	Hoofd-hals	Ongeplande heroperatie	percentage	3.40	5	144	0
2023	St. Het Nederlands Kanker instituut	Hoofd-hals	Ongeplande heroperatie	percentage	7.20	5	73	0
2023	Medisch Spectrum Twente	Hoofd-hals	Ongeplande heroperatie	percentage	0.00	0	36	0
2023	Radboud UMC	Hoofd-hals	Ongeplande heroperatie	percentage	1.90	2	111	0
2023	Canisius Wilhelmina Ziekenhuis	Hoofd-hals	Ongeplande heroperatie	percentage				1
2023	Elisabeth Tweesteden Ziekenhuis	Hoofd-hals	Ongeplande heroperatie	percentage	6.40	3	57	0
2023	Amsterdam UMC	Hoofd-hals	Ongeplande heroperatie	percentage	8.77	10	114	0
2023	Noordwest Ziekenhuisgroep	Hoofd-hals	Ongeplande heroperatie	percentage	2.90	1	31	0

Figure 1: Example of public reporting of an outcome indicator in the Netherlands on zorginzicht: "Percentage of curatively treated patients who underwent resection of a primary head and neck malignancy and required reoperation within 30 days." Column B shows the reporting year; Column C the hospital names; Column N the disease group; Column O the QI theme; Column Q the indicator value (e.g. percentage or median); Column S the scoring method (here, % of reoperations), calculated from Column T (numerator) and Column U (denominator); Column W marks hospitals where the indicator does not apply, for example when all relevant patients are referred elsewhere. Only the most relevant columns are shown; the full document can be found at https://www.zorginzicht.nl/openbare-data.

Missing values in outcomes

Missing values in outcomes are a challenging issue in hospital comparison. Three methods that are commonly used are imputation, complete case analysis and "the missing-to-zero" approach. (Multiple) imputation, which uses an underlying regression model to estimate the most likely value, is often applied for missing case-mix variables [27]. When applied to missing outcomes, however, this approach may be perceived as unfair, as it relies on modelled rather than observed data. Complete case analysis, meaning the removal of all patients with missing outcomes, reduces the number of patients and thereby increases statistical uncertainty (and decreases reliability). Additionally, it may introduce bias if the data are not missing at random. The current approach at the Dutch Institute of Clinical Auditing (DICA) is the "missing to zero" method, where all missing outcomes are coded as 0 (not achieved). The idea is that this creates an incentive to properly register

data. In Chapter 4, we performed a sensitivity analysis to compare the impact for a QI with 20% of outcome data missing. We showed that the different approaches lead to different benchmarks and thus can lead to different hospitals identified as outliers. This highlights the importance of considering the most appropriate method for handling missing outcome data, taking into account the underlying reason for missingness[28]. Ideally, decisions on how to handle missing outcome data are made collaboratively by a data manager, a clinically trained professional, and a methodologist, ensuring both clinical context and statistical validity are addressed. In all cases, transparent reporting of the percentage of missing outcome data and the strategies used to address it is essential to ensure the credibility and comparability of hospital performance.

PROMs for benchmarking

The International Consortium for Health Outcomes Measurement (ICHOM) promotes the use of standardized outcome definitions and explicitly integrates Patient-Reported-Outcomes (PROs) as measured by Patient Reported Outcome Measures (PROMs) and long-term outcomes into their outcome sets [10]. While these outcomes are valuable additions to QI sets, the measurement and incorporation of PROMs is challenging. For example, preoperative (baseline) PROM scores strongly influence postoperative outcomes (Chapter 8), raising questions about the validity of benchmarking with PROMs due to their strong dependence on case-mix. Additionally, PROM data are frequently missing, making them even more difficult to interpret as valid measures of quality of care.

PART II: USING OUTCOME PREDICTIONS TO SUPPORT PERSONALISED CARE

Shared decision-making (SDM) requires outcome information that is both accurate and relevant for the individual patient. Prediction models have the potential to strengthen SDM by providing personalised estimates of risks and benefits that can be discussed with the patient.

External validation of prediction models

To ensure that developed prediction models are accurate in populations beyond the development cohort, external validation is essential. Therefore, we validated the Breast PREDICT 3.1 tool in Chapters 6 and 7. This is a widely used decision tool in breast cancer care that predicts survival and the additional benefit of systemic treatments added to surgery. Our analyses showed that the tool is well calibrated for both Dutch and Swedish populations and that it can also be applied to patients with lobular breast cancer. This is a novel insight that provides practical guidance for medical oncologists, as it was previously unclear whether the tool could be used reliably for lobular breast cancer patients.

Nevertheless, we also found that the model was less accurate for women under 40 years of age. This is particularly concerning, as the tool is even more valuable for younger patients, for whom treatment decisions are often more complex and long-term outcomes more critical. An adjustment to Breast PREDICT for younger patients seems necessary to ensure well-calibrated predictions.

Although Breast PREDICT showed excellent performance for most patient groups, there is a tendency to continually develop new prediction models, most of which never make it into clinical practice [29]. Hueting et al. recently demonstrated in their systematic review of 922 breast cancer prediction models, the majority were at high risk of bias, showed substantial overlap in predictors and outcomes, and were insufficiently reported to allow application in new patients [30]. The large number of overlapping models risks creating noise rather than clarity**.** Rather than prioritizing the continuous development of new models, greater emphasis should be placed on externally validating those prediction tools that are already in clinical use or that have shown promise in prior studies. Validation tests the generalisability of a model to new settings or populations, thereby enhancing clinical trust, sustainability, improving applicability, and promoting reuse and efficiency. Both the development and validation of models should be reported using the TRIPOD guidelines for transparent reporting [31]. Guidelines that we still miss for the development of case-mix models.

Development of prediction models

Development of new models is justified in different circumstances: when prediction models are not yet available for a particular outcome of interest; when experimenting with novel algorithms or methodological approaches; when it is unclear which predictors are most informative, and a simpler model is desired or, in some cases, when validation studies indicate poor performance of existing models. The prediction of HRQoL measured by PROMs is a clear example where development is necessary, as very few models exist despite their growing relevance in cancer care. As PROMs are increasingly integrated into routine clinical practice, HRQoL prediction models provide a new opportunity to improve personalised care.

We developed HRQoL prediction models for breast cancer surgery and reconstruction (Chapter 8). We used linear mixed models to analyse 15 different HRQoL over a two-year horizon, predicting trends over time. Our analyses showed that, combining the pre-operative PROM score of a specific domain with surgery type, achieved comparable performance to more complex models that also included numerous patient and tumour characteristics. These relatively simple models can meaningfully support personalised expectation management, offering patients more realistic information about their likely recovery patterns. It is important to note that the model is not suitable for guiding treatment decisions, due to possible unmeasured confounding by indication in the

development data, which were observational rather than derived from a randomized controlled trial. The models provide predictions after the decision on type of surgery is made, because the predictions are based on the selective group of patients who received the same type of surgery in the past. Additionally, the strong predictive value of baseline PRO scores underlines the crucial importance of measuring them systematically. Without baseline data, interpretation of change over time becomes unreliable, both in prediction modelling and in comparative effectiveness research such as randomised controlled trials[32]. The strong influence of baseline PRO scores indicates that variation in quality of life is driven predominantly by patient characteristics reflected in the baseline score, with treatment contributing to a much lesser extent.

Although some models demonstrate reasonable performance, the question remains whether they are truly 'good enough' for clinical implementation. The multidimensional nature of HRQoL makes prediction inherently challenging, yet such models still provide greater individual accuracy than stratified "patients-like-me" approaches, which rely on subgroup averages based on a limited set of predictors. However, successful implementation requires consensus within the field: Do these models provide sufficient added value for patients? Are PROM-based prediction models considered reliable and useful by clinicians, and if so, across all domains or only specific PRO outcomes? In addition, implementation would require integration into electronic patient records without increasing the time burden for clinicians. Ultimately, the key question is whether the added value justifies the associated costs.

A recent systematic review of HRQoL prediction models in oncology found that the development of robust HRQoL prediction models is still at an early stage[33]. Most models were at high risk of bias or treated HRQoL as a predictor rather than an outcome. Our study addresses this gap by systematically developing advanced longitudinal HRQoL prediction models for breast cancer surgery patients, incorporating multiple HRQoL domains and outcomes, and thereby laying methodological groundwork for future clinical applications.

A further challenge for HRQoL prediction models is the selection and non-response bias of patients who complete PROMs. Respondents are often younger and healthier, meaning that model development may not fully capture the trajectories of more vulnerable or less represented groups [34 35]. This bias needs to be explicitly acknowledged and addressed before PROM-based prediction models can be safely and fairly applied in clinical practice.

LIMITATIONS

Several analyses regarding QIs relied on data from national audits in the Netherlands. This may limit the generalisability of the findings to other countries or healthcare systems with different clinical practices and data definitions. At the same time, the Dutch audits are highly comprehensive, of high data quality, and represent one of the most advanced registry infrastructures internationally, which strengthens the validity of the analyses.

Regarding the external validation of prediction models, we were dependent on data availability, meaning that not all predictors were consistently present in all databases we used. We addressed this by carefully evaluating potential solutions for missing variables and assessing the impact of their absence. For example, for the international validation studies of PREDICT the KI67 marker was available in Sweden but not in the Netherlands. Fortunately, the impact of this variable on the calibration plots was negligible. For the development of HRQoL models, we missed variables on socio-economic status and personal traits, such as 'optimism' and 'trait anxiety' [36]. Therefore, we relied on pre-operative baseline scores, which may also capture individual patient characteristics, though future work should explore this in more depth.

MAIN RECOMMENDATIONS AND FUTURE PERSPECTIVES

The main recommendations arising from this thesis are presented in Textbox 2.

Textbox 2: Main recommendations

1. Healthcare systems and registries should adopt a methodological framework for developing and assessing QIs. As outlined by the practical guide (Chapter 4), the framework should move beyond expert opinion and explicitly include appropriate case-mix adjustment and transparent reporting to ensure valid and reliable comparisons.
2. Public reporting of hospital-level QIs should be based on multiple years of data to reduce random variation and ensure reliable comparisons. In addition, alternative forms of aggregation, such as reporting at the regional level, can be used to balance statistical robustness with clinical relevance and interpretability.
3. Prioritise external validation of existing prediction tools in relevant regions and subgroups over the development of new tools.
4. Health-Related Quality of Life (HRQoL) prediction models offer a new opportunity to improve personalised care, and should further refined

Future directions for research include generating outcome information through innovative methods, with artificial intelligence (AI) supporting automated data collection and analysis while internationally using the same outcome definitions. Automation can reduce registration burden and make it possible to collect more relevant variables and outcomes. This efficiency gain could also make it feasible to expand registrations, register for more diseases, and more easily register in countries that currently lack registry infrastructures. In addition, automation may reduce the need for manual data entry, freeing up resources that can be redirected towards data quality assurance, methodological expertise, and clinical interpretation. Beyond data collection, AI can also support analysis, although this does not necessarily lead to better results and reinforces the need for trained professionals with methodological expertise to ensure validity. At the same time, governance, privacy, and ethical safeguards must be strengthened to ensure responsible data use and successful implementation will depend on sufficient digital infrastructure and adoption by hospitals.

In terms of methodological rigour, knowledge from this thesis should be applied to transforming outcome data into valid and reliable outcome indicators. Therefore, QIs should be routinely assessed for validity and reliability, a process that needs more attention and could also be supported by AI. Additionally, attention should be given to the QI set as a whole, where completeness, overlap, and actionability are essential to ensure a balanced picture of cancer care quality [37]. The ultimate goal is continuous learning and improvement, where international collaboration and the exchange of best practices will be crucial to preserve the value of registries and accelerate quality improvement across borders. These ambitions are also strongly aligned with the Dutch Integral Care Agreement (IZA) and Outcome-Based Care (UGZ), which emphasise outcome-driven healthcare and provide a policy framework to accelerate their implementation.

Prediction models require continuous validation and updating, which should be embedded and implemented more systematically. Extending validation beyond overall prognostic performance towards validation of treatment effects in *Breast PREDICT* could improve the identification of patients who benefit most from specific therapies and those who may safely avoid them[38 39]. PROM-based prediction modelling, which is still largely confined to research settings, also requires further development. Careful attention to potential bias, model calibration, and equity across patient subgroups, as well as ethical considerations, is essential before routine clinical implementation becomes feasible.

Conclusion

This thesis highlights both the potential and the challenges of using outcome data to improve quality of cancer care. It underscores the importance of methodologically robust approaches and the need for transparent reporting to enable valid comparisons of hospital performance. With the increasing emphasis on public reporting of hospital performance, the validity and reliability of quality indicators is even more important. Validity is also crucial for outcome prediction for personalised care. External validation ensures accuracy and generalisability across patient groups and healthcare settings. It should be prioritised over the development of new tools, as this is more efficient and, when the performance is good, strengthens trust in their use in clinical practice. Finally, prediction models for health-related quality of life (HRQoL) represents an important but challenging next step. Such models offer a new opportunity to improve personalised care but require further refinement and validation.

REFERENCES

1. van Dishoeck AM, Koek MB, Steyerberg EW, van Benthem BH, Vos MC, Lingsma HF. Use of surgical-site infection rates to rank hospital performance across several types of surgery. Br J Surg. 2013;100(5):628-36; discussion 37.
2. Lingsma, H. (2010, November 26). Measuring quality of care: Methods and applications to acute neurological diseases. Retrieved from http://hdl.handle.net/1765/21742.
3. Vos EL, Lingsma HF, Jager A, Schreuder K, Spronk P, Vrancken Peeters M, et al. Effect of Case-Mix and Random Variation on Breast Cancer Care Quality Indicators and Their Rankability. Value Health. 2020;23(9):1191-9.
4. Fischer, C. (2015). Quality indicators for hospital care: Reliability and validity [Doctoral dissertation, Erasmus University Rotterdam]. Erasmus University Repository. http://hdl.handle.net/1765/79066.
5. Maes-Carballo M, Gomez-Fandino Y, Reinoso-Hermida A, Estrada-Lopez CR, Martin-Diaz M, Khan KS, Bueno-Cavanillas A. Quality indicators for breast cancer care: A systematic review. Breast. 2021;59:221-31.
6. Rubio IT, Marotti L, Biganzoli L, Aristei C, Athanasiou A, Campbell C, et al. EUSOMA quality indicators for non-metastatic breast cancer: An update. Eur J Cancer. 2024;198:113500.
7. Schreuder K, Bult TJ, Stroop B, Koppert LB, Bijlsma RM, Bantema-Joppe EJ, et al. European quality indicators developed by the European Commission Initiative on Breast Cancer: a first nationwide assessment for the Dutch setting. Breast Cancer Res Treat. 2024;203(3):523-31.
8. European Commission Initiative on Breast Cancer. Development of the scheme. https://healthcare-quality.jrc.ec.europa.eu/ecibc/methodologies/quality-assurance#inline-nav-2. Accessed October 17.
9. . Siesling, S, Verheul, EM, et all (2026) Chapter 07 Defining quality breast cancer care around the globe In D Verhoeven et al, Breast cancer: Global quality care (2e ed) Oxford University Press
10. Ong WL, Schouwenburg MG, van Bommel ACM, Stowell C, Allison KH, Benn KE, et al. A Standard Set of Value-Based Patient-Centered Outcomes for Breast Cancer: The International Consortium for Health Outcomes Measurement (ICHOM) Initiative. JAMA Oncol. 2017;3(5):677-85.
11. Melse PE, Moes I, Becherer BE, et al.; DBIR collaborators. Dutch Breast Implant Registry (DBIR) Annual Report 2022: Version 2023.01. Leiden (NL): Dutch Institute for Clinical Auditing; 2024 Feb. 9, COLLABORATIONS. Available from: https://www.ncbi.nlm.nih.gov/books/NBK604485/
12. Spronk PER, Begum H, Vishwanath S, Crosbie A, Earnest A, Elder E, et al. Toward International Harmonization of Breast Implant Registries: International Collaboration of Breast Registry Activities Global Common Data Set. Plast Reconstr Surg. 2020;146(2):255-67.
13. Fischer C, Lingsma HF, van Leersum N, Tollenaar RA, Wouters MW, Steyerberg EW. Comparing colon cancer outcomes: The impact of low hospital case volume and case-mix adjustment. Eur J Surg Oncol. 2015;41(8):1045-53.
14. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. BMJ Qual Saf. 2011;20(10):869-74.
15. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. JAMA. 2004;292(7):847-51.
16. Z L Rana Kaplan NvL, David van Klaveren, Otto Visser, Eduardus F M Posthuma, Danielle van Lammeren-Venema, Tjeerd J F Snijders, Catharina H M J van Elssen, Anna van Rhenen, Peter A von dem Borne, Nicole M A Blijlevens, Jan J Cornelissen, Marc H G P Raaijmakers, Arjan A van de Loosdrecht, Gerwin Huls, Valery E P P Lemmens, Hester F Lingsma, Avinash

G Dinmohamed. Regional disparities in the use of intensive chemotherapy for AML in the Netherlands: does it influence survival? BMJ Oncology. 2024;3;e00026.
17. Pavlou M, Ambler G, Omar RZ, Goodwin AT, Trivedi U, Ludman P, de Belder M. Outlier identification and monitoring of institutional or clinician performance: an overview of statistical methods and application to national audit data. BMC Health Serv Res. 2023;23(1):23.
18. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127-31.
19. Varewyck M, Goetghebeur E, Eriksson M, Vansteelandt S. On shrinkage and model extrapolation in the evaluation of clinical center performance. Biostatistics. 2014;15(4):651-64.
20. Nederland Z. Open data Medisch-specialistische zorg (MSZ) / Ziekenhuizen en Zelfstandige Behandelcentra 2023 [updated 05-06-2024. Available from: https://www.zorginzicht.nl/openbare-data/open-data-medisch-specialistische-zorg-msz-ziekenhuizen-en-zelfstandige-behandelcentra#verslagjaar-2023.
21. Zorginstituut Nederland. www.zorginstituutnederland.nl/wat-wij-doen/werkwijzen-en-procedures/bevorderen-kwaliteit-en-inzichtelijkheid-van-de-zorg/aanleveren-en-openbaarmaken-van-kwaliteitsgegevens
22. Dutch Hospital Data (DHD). SKI tool. DHD. Retrieved September 24, from https://ski.dhd.nl. [
23. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med. 2005;24(8):1185-202.
24. van Dishoeck AM, Looman CW, van der Wilden-van Lier EC, Mackenbach JP, Steyerberg EW. Displaying random variation in comparing hospital performance. BMJ Qual Saf. 2011;20(8):651-7.
25. van der Willik EM, van Zwet EW, Hoekstra T, van Ittersum FJ, Hemmelder MH, Zoccali C, et al. Funnel plots of patient-reported outcomes to evaluate health-care quality: Basic principles, pitfalls and considerations. Nephrology (Carlton). 2021;26(2):95-104.
26. Ministry of Public Health Welfare and Sport (VWS). Voortgangsrapportage van het Programma Uitkomstgerichte zorg 2023-2024 [Progress report of Outcome-Oriented Care Program 2023-2024] [Available from: https://www.uitkomstgerichtezorg.nl/].
27. van Linschoten RCA, Amini M, van Leeuwen N, Eijkenaar F, den Hartog SJ, Nederkoorn PJ, et al. Handling missing values in the analysis of between-hospital differences in ordinal and dichotomous outcomes: a simulation study. BMJ Qual Saf. 2023;32(12):742-9.
28. Mack C, Su Z, Westreich D. Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition. 2018.
29. Van Calster B, van Smeden, M., van Amsterdam, W., Coemans, M., Wynants, L., & Steyerberg, E. W. The Enemies of Reliable and Useful Clinical Prediction Models: A Review of Statistical and Scientific Challenges. Annual Review of Statistics and Its Application. 2025.
30. Hueting TA, van Maaren MC, Hendriks MP, Koffijberg H, Siesling S. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. J Clin Epidemiol. 2022;152:238-47.
31. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
32. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323(7321):1123-4.
33. van der Heijden TGW, de Ligt KM, Hubel NJ, van der Mierden S, Holzner B, van de Poll-Franse LV, et al. Exploring the role of health-related quality of life measures in predictive modelling for oncology: a systematic review. Qual Life Res. 2025;34(2):305-23.

34. Stephens AR, Bender NR, El-Hassan R, Patel RK. Evidence of non-response bias in patient reported outcome measurement information system surveys. Interv Pain Med. 2025;4(2):100588.
35. Gebert P, Hage AM, Blohmer JU, Roehle R, Karsten MM. Longitudinal assessment of real-world patient adherence: a 12-month electronic patient-reported outcomes follow-up of women with early breast cancer undergoing treatment. Support Care Cancer. 2024;32(6):344.
36. Wintraecken VM, Vulik S, de Wild S, Dirksen C, Koppert LB, de Vries J, Smidt ML. A descriptive systematic review of the relationship between personality traits and quality of life of women with non-metastatic breast cancer. BMC Cancer. 2022;22(1):426.
37. Bruinsma FFE, Nienhuijs SW, Liem RSL, Greve JWM, Marang-van de Mheen PJ. Refining the Quality Indicator Set from the Dutch Audit for Treatment of Obesity to Ensure Usefulness in Clinical Practice. Obes Surg. 2025;35(7):2616-25.
38. van Klaveren D, Serruys PW, Kent DM. Estimating the benefits of personalised glycaemic control. Lancet. 2025;406(10506):913-4.
39. Dennis JM, Young KG, Cardoso P, Gudemann LM, McGovern AP, Farmer A, et al. A five-drug class model using routinely available clinical features to optimise prescribing in type 2 diabetes: a prediction model development and validation study. Lancet. 2025;405(10480):701-14.



Addendum

Summary (English)
Samenvatting (Dutch)
List of publications
PhD portfolio
About the author
Dankwoord

SUMMARY (ENGLISH)

Quality of care is a central yet complex concept within healthcare. According to the widely adopted framework by Donabedian, quality can be conceptualized across three core domains: structure, process, and outcomes. These domains form the foundation for developing **quality indicators** that enable healthcare to be systematically measured and compared. While all three domains are important, modern health systems increasingly prioritize outcomes. These outcomes include both **clinical endpoints**, such as complications and survival, and **patient-reported outcomes (PROs)** that measure experienced quality of life.

The systematic measurement and application of outcome information provide opportunities to improve healthcare systems and to inform individual treatment decisions. At the same time, this requires high standards for the **validity**, **reliability**, and interpretation of outcome information. Differences in patient populations between hospitals (case mix), statistical uncertainty (reliability), and data quality can all influence results. A rigorous methodological approach is therefore essential to apply outcome information in a meaningful way.

This thesis investigates how outcome information can be utilized to ensure both reliability and meaningful interpretation. Part I focuses on the use of valid and reliable outcome information to stimulate continuous quality improvement through hospital comparisons, whereas Part II demonstrates how prediction models can translate outcome information into personalised decision-making and expectation management in clinical practice.

Part I: Using outcome information to stimulate continuous quality improvement

Chapter 2 presents a literature review that assesses which cancer outcomes are used for hospital comparisons (i.e. benchmarking) in Europe and how case-mix adjustment is applied. The study shows that benchmarking is conducted in only a few countries, often focusing on short-term outcomes (≤90 days) and lacking standardized or validated case-mix models. Moreover, there is limited transparency regarding the applied case-mix methodologies, which restricts comparability both within and across countries. These findings highlight that benchmarking of cancer outcomes is currently fragmented and methodologically underdeveloped. There is an urgent need for harmonized definitions, standardized approaches to case-mix modelling, and transparent reporting to enable outcome-based care to be compared and improved at national and international levels.

Chapter 3 addresses the validity of quality indicators and describes the development of a case-mix model that adjusts for differences in patient and tumour characteristics between hospitals. Such adjustment helps to ensures that observed differences in hospital

performance reflect true differences in quality of care rather than variations in patient case mix. This chapter specifically focuses on the quality indicator *complications after breast cancer surgery*. The results show that, after stratification by type of surgery, additional case-mix adjustment had negligible effects on hospital differences. The patient and tumour characteristics determining complication risk were evenly distributed across hospitals. This suggests that, for this indicator, stratification alone is sufficient and further adjustment is unnecessary. However, such evaluations should be performed for each indicator individually, particularly for outcome indicators where case-mix adjustment typically has a greater impact.

Chapter 4 introduces a framework to systematically assess quality indicators based on four measurable criteria: feasibility (data availability), discriminatory ability, validity (case-mix), and reliability. This framework was applied to the Dutch Breast Cancer Audit (NBCA: NABON Breast Cancer Audit). Most NBCA indicators demonstrated high data availability and moderate to good discriminatory ability. For a few indicators, case-mix adjustment is necessary. However, reliability was often a concern, as many observed hospital differences were largely driven by random variation, requiring caution in public reporting. his chapter demonstrates that many existing quality indicators are well suited for internal quality improvement but not yet for external public comparison. Reliability should therefore play a central role in decisions about public disclosure. Furthermore, combining data across multiple years substantially improves reliability, offering a practical solution. The proposed framework thus provides a concrete tool to systematically evaluate quality indicators and enhance the quality and transparency of indicator sets.

Chapter 5 applies the framework from **chapter 4** to a low-volume disease (oral cavity cancer). Sample size calculations were performed to determine how many patients, and consequently how many years of data, are required for reliable reporting. This provides practical guidance on the minimum reporting period needed to detect quality differences with sufficient reliability.

Part II: Using outcome predictions to support personalised care

While the first half of this thesis uses outcome information to understand differences between hospitals, the second half applies the same type of information to personalise care for individual patients. We focus on prediction models that can either support *shared decision-making* (**Chapters 6 and 7**) or help manage patient expectations (**Chapter 8**). **Chapter 6** externally validates the latest version of the prediction model *PREDICT Breast (v3.1)* in Dutch and Swedish cohorts, with particular attention to young patients (≤40 years) and those with lobular breast cancer. External validation means evaluating an existing model in a different population than the one in which it was developed, to assess whether it performs accurately and reliably in new settings. *PREDICT Breast v3.1* performed well in the general population and in patients with lobular breast cancer but

was less accurate for younger patients. These findings show that *PREDICT* can be broadly applied in clinical practice, but that model performance is not uniform across all subgroups. This underscores the importance of ongoing validation and subgroup-specific recalibration to ensure the reliability of personalised survival predictions.

Chapter 7 compares the performance of PREDICT Breast v3.1 with that of the previous version (v2.2) in the Dutch population. Both models demonstrated accurate predictive performance, with only minor differences between subgroups. These findings, together with those from **Chapter 6**, show that external validation is an efficient and valuable strategy: it strengthens clinical confidence and prevents the fragmentation that arises from continuously developing new models from scratch. Systematic validation and recalibration of existing (and, where possible, already implemented) models is therefore often more effective and sustainable than developing multiple overlapping models.

Chapter 8 develops and externally validates fifteen prediction models for various domains of health-related quality of life (HRQoL) after breast cancer surgery. In this case, new model development was justified, as no previous models existed for HRQoL outcomes and a novel methodological approach was applied that analyses trends over time. The preoperative PRO score (baseline measurement) of a specific domain was the strongest predictor of the postoperative trajectory in that domain. The models performed acceptable in external validation and offer potential for use in personalised expectation management. The reason these models are not suitable for *shared decision-making* about treatment choice is that they were developed using real-world data, in which patients were not randomised to a treatment. As a result, there is selection bias: the patient groups undergoing each type of surgery are not comparable. Therefore, the model can only be applied once the choice of surgical procedure has been made. Nevertheless, this chapter represents an important step towards the use of patient-reported outcomes in prediction care, where managing patient expectations plays a crucial role. However, further refinement and validation are needed before integration into electronic care pathways can be considered.

Conclusion

This thesis highlights both the potential and the challenges of using outcome information to improve the quality of cancer care. It emphasizes the importance of robust methodological approaches and transparent reporting to enable valid comparisons of hospital performance. The reliability of quality indicators remains an important consideration, and combining outcome data across multiple years provides a practical way to strengthen this reliability.

Furthermore, validity is crucial for prediction models: external validation ensures the accuracy and generalizability of models across different patient groups and healthcare

settings. This should take priority over developing new models, as it is more efficient and, when performance is good (as with *PREDICT Breast*), it strengthens confidence in their clinical application. Finally, predicting health-related quality of life (HRQoL) represents an important yet challenging next step toward further personalising cancer care.

Recommendations

1. Healthcare systems and quality registries should adopt a methodological framework for developing and validating quality indicators, with explicit attention to validity, reliability, and transparent reporting of methodology.
2. Public reporting of hospital-level quality indicators should be based on multi-year data to minimize random variation (statistical uncertainty) and improve reliability.
3. Priority should be given to the external validation of existing prediction models over the continuous development of new ones.
4. Prediction models for health-related quality of life (HRQoL) offer a new opportunity to further personalise care. Continued refinement, validation, and integration into clinical practice are needed to fully realize this potential.

SAMENVATTING (DUTCH)

Kwaliteit van zorg is een centraal maar complex begrip binnen de gezondheidszorg. Volgens het klassieke raamwerk van Donabedian kan kwaliteit worden beoordeeld aan de hand van drie domeinen: structuren, processen en uitkomsten. Deze drie domeinen vormen de basis voor het ontwikkelen van **kwaliteitsindicatoren** waarmee zorg systematisch kan worden gemeten en vergeleken. Hoewel alle drie de domeinen belangrijk zijn, leggen moderne zorgsystemen in toenemende mate de focus op uitkomsten. Tot die uitkomsten behoren zowel **klassieke klinische uitkomsten**, zoals complicaties en overleving, als **patiënt gerapporteerde uitkomsten (PROs)** die de ervaren kwaliteit van leven meten.

Het systematisch meten en gebruiken van deze uitkomsten biedt kansen om zowel zorgsystemen als individuele behandelkeuzes te verbeteren en te ondersteunen. Tegelijkertijd stelt dit hoge eisen aan de **validiteit**, **betrouwbaarheid** en interpretatie van uitkomstinformatie. Verschillen in patiëntpopulaties tussen ziekenhuizen (case-mix), statistische onzekerheid (betrouwbaarheid) en datakwaliteit kunnen bijvoorbeeld de resultaten beïnvloeden. Een zorgvuldige methodologische aanpak is daarom essentieel om uitkomstinformatie op een betekenisvolle manier toe te passen.

Dit proefschrift onderzoekt hoe uitkomstinformatie betrouwbaar en betekenisvol kan worden gebruikt. **Deel I** richt zich op het gebruik van uitkomstinformatie voor continue kwaliteitsverbetering via ziekenhuisvergelijkingen, terwijl **deel II** laat zien hoe predictiemodellen uitkomstinformatie kunnen vertalen naar gepersonaliseerde besluitvorming of persoonlijk verwachtingsmanagement in de klinische praktijk.

Deel I: Gebruik van uitkomstinformatie om continue kwaliteitsverbetering te stimuleren

In **hoofdstuk 2** wordt door middel van een literatuurstudie onderzocht welke kankeruitkomsten worden gebruikt voor ziekenhuisvergelijkingen (benchmarking) in Europa en hoe case-mixcorrectie wordt toegepast. De studie laat zien dat benchmarking slechts in enkele landen plaatsvindt, vaak met nadruk op korte termijnuitkomsten (≤90 dagen) en zonder gestandaardiseerde en gevalideerde case-mixmodellen. Er bestaat bovendien weinig transparantie over de gehanteerde methodologie van de case-mix modellen als deze wel gebruikt zijn, wat de vergelijkbaarheid in en tussen landen beperkt. Deze bevindingen benadrukken dat benchmarking van kankeruitkomsten momenteel gefragmenteerd en methodologisch onvoldoende onderbouwd is. Er is dringend behoefte aan geharmoniseerde definities, uniforme benaderingen voor case-mix modellen en transparante rapportage, zodat uitkomstgerichte zorg nationaal en internationaal kan worden vergeleken en verbeterd.

Hoofdstuk 3 richt zich op de validiteit van kwaliteitsindicatoren en beschrijft de ontwikkeling van een case-mixmodel. Zo'n model corrigeert voor patiënt- en tumorkarakteristieken, zodat ziekenhuizen op basis van vergelijkbare patiëntpopulaties met elkaar kunnen worden vergeleken. Het is immers niet wenselijk dat verschillen in ziekenhuisprestaties in werkelijkheid slechts de samenstelling van patiëntengroepen weerspiegelen. **Hoofdstuk 3** richt zich specifiek op de kwaliteitsindicator 'complicaties na borstkankerchirurgie'. De resultaten laten zien dat, na stratificatie naar type chirurgie, verdere case-mixcorrectie nauwelijks invloed heeft op de verschillen tussen ziekenhuizen. De patiënt- en tumorfactoren die het risico op complicaties beïnvloeden, bleken gelijkmatig over de ziekenhuizen verdeeld. Dit betekent dat binnen deze indicator stratificatie volstaat en aanvullende correctie niet nodig is. Wel blijft het belangrijk om dit per indicator te evalueren, vooral bij uitkomstindicatoren, waarbij case-mixcorrectie doorgaans een grotere invloed heeft.

Hoofdstuk 4 introduceert een raamwerk om kwaliteitsindicatoren systematisch te beoordelen aan de hand van vier meetbare criteria: databeschikbaarheid, discriminatievermogen, validiteit (case-mix) en betrouwbaarheid. Dit raamwerk is toegepast op de Nederlandse borstkankerregistratie (NBCA: NABON Breast Cancer Audit). De meeste NBCA-indicatoren bleken een hoge databeschikbaarheid en een goed discriminerend vermogen te hebben. Voor enkele indicatoren bleek case-mixcorrectie noodzakelijk. De betrouwbaarheid was daarentegen vaak beperkt. Veel ziekenhuisverschillen bleken gerelateerd aan toeval, wat voorzichtigheid vereist bij publieke rapportage. Dit hoofdstuk laat zien dat veel bestaande kwaliteitsindicatoren wél geschikt zijn voor interne kwaliteitsverbetering, maar nog niet voor publieke vergelijking. Betrouwbaarheid moet daarom een centrale rol spelen bij de beslissing om indicatoren extern te publiceren. Daarnaast laten we zien dat het samenvoegen van gegevens over meerdere jaren de betrouwbaarheid aanzienlijk vergroot en daarmee een praktische oplossing biedt. Het gepresenteerde raamwerk vormt zo een concreet instrument om kwaliteitsindicatoren systematisch te toetsen en de kwaliteit en transparantie van indicatorensets te versterken.

Hoofdstuk 5 past het raamwerk uit **hoofdstuk 4** toe op een laagvolumeziekte (mondholtekanker). Door middel van steekproefgrootteberekeningen wordt vastgesteld hoeveel patiënten of jaren nodig zijn voor betrouwbare rapportage. Dit biedt concrete handvatten voor het bepalen van de minimale rapportageperiode die vereist is om kwaliteitsverschillen met voldoende betrouwbaarheid te kunnen identificeren.

Deel II: Gebruik van uitkomsten voor voorspellingen ter ondersteuning van gepersonaliseerde zorg

Waar de eerste helft van dit proefschrift uitkomstinformatie inzet om verschillen tussen ziekenhuizen te begrijpen, gebruikt de tweede helft diezelfde soort informatie om zorg te personaliseren voor de individuele patiënt. Hiervoor focussen we ons op voorspellingsmodellen, die zowel kunnen ondersteunen bij het "samen beslissen" (**hoofdstuk**

6,7) of kunnen helpen het verwachtingsmanagement voor een patiënt (**hoofdstuk 8**). In **hoofdstuk 6** wordt de nieuwste versie van het predictiemodel *PREDICT Breast* (v3.1) extern gevalideerd in Nederlandse en Zweedse cohorten, met bijzondere aandacht voor jonge patiënten (≤40 jaar) en patiënten met lobulaire borstkanker. Externe validatie houdt in dat een bestaand model wordt getest in een andere populatie dan waarin het is ontwikkeld, om te beoordelen of het daar eveneens accuraat en betrouwbaar presteert. *PREDICT Breast* v3.1 presteerde goed in de algemene populatie en bij patiënten met lobulaire borstkanker, maar minder nauwkeurig bij jonge patiënten. Deze resultaten tonen aan dat *PREDICT* breed toepasbaar is in de klinische praktijk, maar dat de modelprestaties niet uniform zijn voor alle subgroepen. Dit benadrukt de noodzaak van voortdurende validatie en subgroep-specifieke recalibratie om de betrouwbaarheid van gepersonaliseerde overlevingsvoorspellingen te waarborgen.

Hoofdstuk 7 vergelijkt de prestaties van *PREDICT Breast* v3.1 met de eerdere versie (v2.2) binnen de Nederlandse populatie. Beide modellen voorspelden de overleving accuraat, met slechts kleine verschillen tussen subgroepen. Deze bevindingen, samen met die uit hoofdstuk 6, tonen aan dat herhaalde externe validatie een efficiënte en waardevolle strategie is: het versterkt het klinisch vertrouwen en voorkomt de fragmentatie die ontstaat wanneer voortdurend nieuwe modellen van nul worden ontwikkeld. Systematische validatie en recalibratie van bestaande (en waar mogelijk reeds geïmplementeerde) modellen is daarmee vaak effectiever en duurzamer dan het ontwikkelen van steeds nieuwe, overlappende modellen.

In **hoofdstuk 8** worden vijftien voorspellende modellen ontwikkeld en extern gevalideerd voor verschillende domeinen van gezondheid gerelateerde kwaliteit van leven (KvL) na borstkankerchirurgie. In dit geval was nieuwe modelontwikkeling wél gerechtvaardigd, omdat er nog geen bestaande modellen voor KvL-uitkomsten beschikbaar waren en een nieuwe methodologische aanpak werd toegepast die trends over tijd analyseert. De preoperatieve PRO-score (baseline meting) van een specifiek domein bleek de sterkste voorspeller van het postoperatieve traject van hetzelfde domein. De modellen presteerden redelijk in externe validatie en bieden perspectief voor gebruik in gepersonaliseerd verwachtingsmanagement. De reden dat deze modellen niet geschikt zijn om te gebruiken tijdens het "samen beslissen" over de behandelkeuze, is dat ze zijn ontwikkeld op basis van *real-world* data, waarbij patiënten niet gerandomiseerd zijn toegewezen aan een behandeling. Hierdoor is er sprake van selectiebias: de patiëntengroepen die elk type chirurgie hebben ondergaan, zijn niet vergelijkbaar. Het model kan daarom pas worden toegepast nadat de keuze voor het type chirurgie is gemaakt. Desalniettemin vormt dit hoofdstuk een belangrijke stap richting de toepassing van patiënt gerapporteerde uitkomsten in voorspellende zorg, aangezien het managen van patiëntverwachtingen een cruciale rol speelt. Verdere verfijning en validatie zijn echter noodzakelijk voordat integratie in elektronische zorgpaden kan worden overwogen.

Conclusie

Dit proefschrift belicht zowel de mogelijkheden als de uitdagingen van het gebruik van uitkomstinformatie om de kwaliteit van de kankerzorg te verbeteren. Het benadrukt het belang van een robuuste methodologische aanpak en de noodzaak van transparante rapportage, om valide vergelijkingen van ziekenhuisprestaties mogelijk te maken. De betrouwbaarheid van kwaliteitsindicatoren blijft daarbij een belangrijk aandachtspunt. Het combineren van uitkomsten over meerdere jaren vormt een praktische oplossing om deze betrouwbaarheid te vergroten.

Daarnaast is validiteit cruciaal voor voorspellende modellen: externe validatie waarborgt de nauwkeurigheid en generaliseerbaarheid van modellen over verschillende patiëntengroepen en zorgcontexten heen. Dit verdient prioriteit boven de ontwikkeling van steeds nieuwe modellen, omdat het efficiënter is en, wanneer de prestaties goed zijn (zoals bij *PREDICT Breast*) het vertrouwen in klinische toepassing versterkt. Het voorspellen van gezondheid gerelateerde KvL vormt ten slotte een belangrijke maar uitdagende volgende stap in het verder personaliseren van de zorg.

Aanbevelingen

1. Zorgsystemen en kwaliteitsregistraties zouden een methodologisch raamwerk moeten hanteren voor de ontwikkeling en validatie van kwaliteitsindicatoren, met expliciete aandacht voor validiteit, betrouwbaarheid en transparante rapportage van de gebruikte methodologie.
2. Publieke rapportage van ziekenhuisindicatoren zou gebaseerd moeten zijn op meerjarige data om toevallige variatie (statistische onzekerheid) te verminderen en de betrouwbaarheid te vergroten.
3. Prioriteit moet worden gegeven aan de externe validatie van bestaande predictiemodellen, boven het ontwikkelen van steeds nieuwe modellen.
4. Voorspellende modellen voor gezondheid gerelateerde KvL bieden een nieuwe kans om zorg verder te personaliseren. Verdere verfijning, validatie en integratie in klinische praktijk zijn noodzakelijk om dit potentieel volledig te benutten.

LIST OF PUBLICATIONS

This thesis

Thurell J, Doppelbauer L, **Verheul EM**, Petrov I, Karsten MM, Koppert LB, Bergh J, Fredriksson I, Lindgren P, Kiani N, Hedayati E. Benchmarking cancer outcomes in Europe: a scoping review of methodologies and case-mix adjustments. *ESMO Real-World Data and Digital Oncology*. 2025.

Verheul EM, van Klaveren D, Lingsma HF, Vos E, Hoornweg MJ, Siesling S, Koppert LB; NBCA Consortium. High-impact complications after breast cancer surgery in the Dutch National Quality Registry: evaluating case-mix adjustment for hospital comparisons. *BJS Open*. 2024.

Verheul EM*, van der Linde M*, Lingsma HF, Vos E, Siesling S, Koppert LB; NBCA Consortium. Comprehensive evaluation of quality indicators: analyzing the Dutch Breast Cancer Audit. *Int J Health Policy Manag*. 2025. *Shared first authorship*.

van Oorschot HD, **Verheul EM**, Hardillo JA, van Es RJJ, van den Broek GB, de Graaff MR, Lingsma HF, Baatenburg de Jong RJ; Dutch Head and Neck Audit – Oral cavity collaborator group. Public performance reporting of oral cancer surgery: detecting meaningful between-hospital variation. *Submitted*.

Verheul EM, Doornkamp F, Petrov I, Siesling S, Lingsma HF, Koppert LB, Vreven LWA, Voogd AC, Karsten MM, Doppelbauer L, Gebert P, Kiani N, Borstnar S, Hedayati E, Steyerberg EW, van Klaveren D, on behalf of the 4D PICTURE Consortium; External validation of PREDICT Breast v3.1 for overall survival in international cohorts, including young and invasive lobular subgroups. *Submitted.*

Vreven LWA, **Verheul EM**, van Maaren MC, Doornkamp F, Schipper RJ, Siesling S, Pharoah PDP, Tjan-Heijnen VCG, Voogd AC. Comparative validation of PREDICT versions 3.1 and 2.2 for overall survival in the Dutch breast cancer population. *The Breast*. 2025.

Verheul EM, Karsten MM, Gebert P, Doppelbauer L, Borstnar S, Siesling S, Lingsma HF, Vrancken-Peeters NJMC, Stiggelbout AM, Mureau MAM, Koppert LB, van Klaveren D; 4D PICTURE Consortium. Development and validation of prediction models for health-related quality of life outcomes after breast cancer surgery and reconstruction. *European Journal of Surgical Oncology (ESJO), 2026*.

Other publications

Doornkamp F, de Wreede LC, Jager A, Koppert LB, **Verheul EM**, Steyerberg EW. Guiding treatment decisions in early breast cancer: a comparative analysis of the OncotypeDX and MammaPrint tests. *The Breast*. 2026.

List EB, Hahn BA, Qiu SS, de Jong T, Rakhorst HA, **Verheul EM**, Maarse W, Coert JH, Krijgh DD. Free fasciocutaneous versus muscle flaps in lower extremity reconstruction: implications for functionality and quality of life. *J Reconstr Microsurg*. 2025;41(8):641–8. doi:10.1055/a-2483-5388. Epub 2024 Nov 25.

Krijgh DD, List EB, Beljaars B, Qiu SS, de Jong T, Rakhorst HA, **Verheul EM**, Maarse W, Coert JH. Patient-reported esthetic outcomes following lower extremity free flap reconstruction: a cross-sectional multicenter study. *J Plast Reconstr Aesthet Surg*. 2024;93:1–8. doi:10.1016/j.bjps.2024.03.007. Epub 2024 Mar 21.

Oflazoglu K, **Verheul EM**, Pong TM, Ritt MJFP, Rakhorst H, Chen NC. A threshold QuickDASH score for estimating a diagnosis of major depression in patients with fingertip injuries in the American and Dutch population. *Hand (N Y)*. 2023;18(4):692–700. doi:10.1177/15589447211060456. Epub 2021 Dec 30.

Verheul EM, Berner JE, Oflazoglu K, Troisi L, Arnež Z, Ortega-Briones A, Nanchahal J, Rakhorst H. Development of guidelines for the management of patients with open fractures: the potential cost-savings of international collaboration. *J Plast Reconstr Aesthet Surg*. 2022;75(1):439–88. doi:10.1016/j.bjps.2021.08.047. Epub 2021 Oct 9.

PHD PORTFOLIO

PhD candidate: Elfi Malou Verheul
Department: Medical decision making, Public health, Erasmus MC Rotterdam
PhD period: 01/02/2023 – 30/12/2025
Promotors: Prof. dr. H.F. Lingsma, Prof. dr. S. Siesling & Dr. L. B. Koppert
Co-promotors: Dr. D. van Klaveren

	Year(s)	ECTS
PhD training		
Courses		
PhD Introduction session	2023	0.2
Scientific integrity	2023	0.3
Datacamp R courses (introduction to R, introduction to statistics in R, introduction to the Tidyverse, Data manipulation with dplyr, Intermediate R)	2023	1
R course Erasmus Graduate school	2023	0.8
Diversity & Inclusivity	2023	0.1
Survival Analysis (Module III)	2023	1.2
Advanced Analysis of Prognosis Studies	2023	0.9
Introduction to Data-analysis	2023	0.7
Fundamentals of Medical Decision Making	2023	0.7
Biomedical writing for PhD candidates	2023-2024	1.5
Review of Mathematics and Introduction to Statistics	2024	1.0
This Thing Called Science	2024-2025	1.3
Intercultural Communication	2024	0.2
Study Design	2024	4.0
Biostatistics I	2024	4.5
Biostatistics II	2024	4.5
Clinical Epidemiology	2024	3.0
Principles of Public Health	2024	3.0
Core Competences Exam	2025	1.0
Selected Topics in Epidemiology	2025	3.0
Topics in Medical Decision-making	2025	1.4
Qualitative Research Methods in Medicine	2025	1.1
Advanced Clinical Trials	2025	1.9
Quality of Life Measurement	2025	0.9
Causal Mediation Analysis	2025	1.4
Methods to Evaluate Public Health Interventions	2025	0.7
Joint Models for Longitudinal and Survival Data	2025	0.7
Machine Learning for Health Research	2025	0.7

	Year(s)	ECTS
Personal development		
Intervision for PhD candidates	2023-2025	1
Design and Pitch a good poster	2024	0.2
Presenting skills for PhD candidates	2023	0.8
Intercultural Communication	2024	0.2
Match coachingstraject	2025	1
Seminars and workshops		
Research seminars, medical decision making, department of public health, Erasmus MC, Rotterdam, The Netherlands	2023-2025	1.0
Research seminars, Dutch institute of Clinical auditing, Leiden, The Netherlands	2023-2025	3.0
Lectures, Dutch institute of Clinical auditing, Leiden, The Netherlands	2023-2025	1.0
4D picture seminars and methodological meetings	2023-2025	0.5
Seminar Bias and Fairness in the use of Clinical Prediction Models (by prof. David Kent)	2023	0.05
Seminar AI and machine learning to improve medical care? Pitfalls and challenges (by prof. Ewout Steyerberg)	2023	0.05
ICHOM webinar: Introducing ICHOM: A Value-based Healthcare Model	2024	0.05
ICHOM webinar: The Central Role of Patients in Defining Health Outcomes: How Co-creation Benefits all Stakeholders	2024	0.05
ICHOM webinar: Overcoming Common Implementation Barriers - Data Collection	2024	0.05
Short course: The Potential and Pitfalls of Predicting Treatment Effects, Rotterdam, the Netherlands	2025	0.1
Webinar NVGE PhD Netwerk: Medisch Onderzoek in de media	2025	0.05
Presentations at conferences		
4D picture consortium meeting, Berlin, Germany	2023	1
Clinical & Health Sciences PhD Day Erasmus, Rotterdam, the Netherlands (Pitch & best poster award)	2024	0.5
Borstkanker Behandeling Beter Symposium (BBB), Rotterdam, the Netherlands	2024	0.5
ICHOM, Amsterdam, the Netherlands (poster)	2024	1
Society of Medical Decision Making conference (SMDM), Boston, USA (poster 2x)	2024	1.5
4D picture consortium meeting, Kolding, Denmark	2025	1
NVPC Wetenschapsdag (Nederlandse Vereniging voor Plastische Chirurgie) 2025, AFAS theater Leusden, the Netherlands	2025	0.5
Attendance at conferences and symposia		
PhD Day Erasmus MC: Shaping your future, Rotterdam, the Netherlands	2023	0.5
Metromapping Symposium, Utrecht, the Netherlands	2023	0.2

	Year(s)	ECTS
Symposium Dure Geneesmiddelen organised by IKNL, Utrecht, the Netherlands	2023	0.2
Healthcare in Shape Symposium, TU Delft, the Netherlands	2023	0.5
Symposium Impact COVID-19-pandemie bij borstkanker, UT Twente, the Netherlands	2023	0.2
Van meten naar leren en verbeteren: kansen en uitdagingen van uitkomstgerichte zorg, Rotterdam, the Netherlands	2024	0.2
Value Based Health Care Summer course, Rotterdam, the Netherlands	2024	1.0
Lof der Geneeskunst Symposium, Rotterdam, the Netherlands	2024	0.2
International Lobular Breast Cancer Symposium, Utrecht, the Netherlands	2024	0.2
SES & kanker Symposium organised by IKNL, Utrecht, the Netherlands	2024	0.2
Teaching		
Supervising NIHES master student, Erasmus MC	2023-2024	1.5
Workgroup teacher Value Based Healthcare at Erasmus School of Health Policy & Management	2023	0.6
Invited expert, Quality Management in Healthcare, University of Twente	2025	0.1
Medical advisor, Service Design Course, TU Delft	2025	0.1
Other activities		
Presentation at the Surgery Department, Erasmus MC	2024	0.05
Presentation at 'refereermiddag IKNL' and 'IKNL datamanagers meeting'	2024	0.1
Presentation at DICA's 'maandoverleg', 'voorzittersoverleg' & 'ontwikkeldag'	2024-2025	0.15
Member of the project group "Gene Expression Profiles (GEPs)" for the ZonMw grant application "Kaderprogramma Passende Zorg – Voorbereidende implementatieactiviteiten"	2025	0.1

ABOUT THE AUTHOR

Elfi M. Verheul was born on July 21, 1995, and grew up in Houten, the Netherlands. In 2013, she graduated from Cals College in Nieuwegein and moved to Groningen to study Medicine. During her bachelor's program, Elfi was actively involved in various committees and initiatives, including one year on the board of the medical faculty's student bar *Villa Volonté* and a later board position for the national SurgEOn Day. After completing her bachelor's degree and prior to starting her clinical rotations, she pursued minors in neuroscience, behavioural neuroscience, and sport science, held various side jobs, and travelled through Asia for three months.

In November 2017, Elfi started her Master's in Medicine and began her clinical rotations at the University Medical Center Groningen (UMCG). In December 2018, she moved to Enschede to complete the final two years of her master's training at Medisch Spectrum Twente (MST), where she graduated within the departments of surgery and plastic surgery.

During the COVID-19 pandemic, Elfi worked for one year at the department of surgery at Isala Hospital in Zwolle, followed by one year at the plastic surgery department at Ziekenhuisgroep Twente (ZGT). In addition to her work at ZGT, she regularly covered evening and weekend shifts for the regional plastic surgery service and worked occasional days at the Expert Clinic Enschede.

In February 2023, she started her PhD at the Erasmus Medical Center Rotterdam (EMC) in collaboration with the Dutch Institute for Clinical Auditing (DICA). Alongside her PhD, Elfi completed a second master's degree in Health Sciences with a specialization in Clinical Epidemiology. During the final year of her PhD, she occasionally worked night shifts at the Alexander Monro Hospital for breast care and took Spanish lessons in preparation for future travels in South America with her boyfriend. As outside of work and research, Elfi enjoys sports, spending time with friends, and traveling.



Vanuit mijn rol zie ik de impact van onze samenwerking in het onderzoek naar het verbeteren van de kwaliteit en personalisatie van de (borst)kankerzorg.

Elfi Verheul
arts-onderzoeker Dutch Institute for Clinical Auditing (DICA),
NABON Breast Cancer Audit (NBCA), Medical Decision-Making Erasmus MC

Uit: IKNL jaarplan 2025

DANKWOORD

Dit wilde ik kort houden, maar dat is misschien toch niet helemaal gelukt. Allereerst mijn fantastische promotieteam:

David van Klaveren, dank voor je eindeloze geduld (terwijl ik dat zelf soms niet had), de ruimte om (ook domme) vragen te stellen, je begrip voor de soms wat gekke medische wereld, het vertrouwen dat ik veel dingen tegelijk mocht doen en de oplossing voor de partial R2 struggle. Betere begeleiding had ik me niet kunnen wensen! Door onze gezamenlijke liefde voor bier, de vele borrels, etentjes, Berlijn en Boston heb ik je ook op persoonlijk vlak goed mogen leren kennen. Hopelijk komt er nog een keer een feestje met open bar en liveband (als jij prof wordt?), of met uitzicht op een baseballveld inclusief een "lonely man" die onze biertjes sponsort.

Linetta Koppert, jouw positiviteit werkt aanstekelijk! Dank voor je begeleiding binnen de NBCA, de internationale connecties, de mogelijkheid voor de BBB-presentatie, de gesprekken over mijn vervolgstappen, en dat je op je vrije vrijdagmiddag je pannenkoekenlunch soms een paar minuten later startte om mij te helpen. En vergeet niet je heerlijke kookkunsten die ik heb mogen ervaren bij Noëlle's promotiediner bij jouw thuis!

Hester Lingsma, jouw snelheid in het vormen van een scherpe visie over stukken en je kritische blik op de inhoud waren ontzettend waardevol. Ik heb genoten van onze carrièregesprekken, de biertjes in Berlijn en de vele (nu niet meer bestaande) CMB-borrels (gelukkig heb je de stressbal nog!).

Sabine Siesling, dank voor de kansen die je me bood! Zoals het meeschrijven aan een boek en het deelnemen aan de Board of Directors Quality of Care op de UT in Twente. Maar de mooiste herinnering samen is toch wel het fantastische congres in Boston (inclusief kreeft-sleutelhanger) en de karaokebar met Axel en Jens (naar links tudududududuuuuuu).

En shoutout naar **Judith**, **Manon**, **Marlou** en **Esmee**! Daar mogen jullie je gelukkig mee prijzen Hester, David, Sabine en Linetta!

Dank aan alle mooie samenwerkingen met coauteurs, want teamwork is dreamwork, en multidisciplinariteit maakt de onderzoeken echt beter! Natuurlijk mijn **4D Picture-collega's**, waaronder Team Charité (**Maria Margarete Karsten, Lea Doppelbauer, Pimrapat Gebert**) Team Karolinska (**Elham Hedayati, Iurii Petrov, Narsis Kiani**) en Team Lublijana (**Simona Borstnar**). En dan niet te vergeten al onze **Nederlandse toppers**, met in het bijzonder **Frank**, voor de samenwerking aan het validatiepaper – jij de statistiek, ik de medische inhoud – en de herinneringen in Berlijn en Denemarken. Het QoL predictie

paper, dank **Marc Mureau** voor de scherpe inhoudelijke input en **Ewout Steyerberg** voor de kritische vragen t.a.v de statistiek. **Jacob Thurell**, om mij te betrekken bij onze Europese '*systematic review*'.

Mijn **CMB-collega's en secretariaat**, waaronder **Carolien** en **Rana**, dank voor jullie tips en steun als doorgewinterde PhD'ers. **Diaz**, Boston was leuker dankzij jou! Daar heb ik je echt beter mogen leren kennen. **Jente**, dank voor onze fijne en persoonlijke gesprekken.

Margrietha, Bij Hesters oratie liepen we elkaar voor het eerst tegen het lijf, en het klikte meteen! Ik kan over jou (en Clau) wel twee pagina's volschrijven, maar ik houd het kort: wij zijn een topteam en hebben elkaar echt versterkt, met onze brainstorm- en feedbacksessies en jouw kritische blik. Daarnaast onze huizenliefde (Funda), sportschoolsessies (al zijn die er de laatste tijd wel een beetje ingeslopen), koffietjes, (Westkop)borrels, gesprekken over het leven, de kapperssessies en rode wijntjes. Gelukkig zijn onze zelfgemaakte kussens in mijn erkerbank een permanente herinnering aan onze gezelligheid!

Claudia: ik doe het in het Nederlands, want lezen is makkelijker voor jou dan mijn veel te snelle praatjes (oeps). Maar ondanks dat het vaak te snel gaat, hou jij je wel staande, bijvoorbeeld op mijn 29e verjaardag met alleen maar Dutchies! Trots! Wij begonnen exact tegelijk aan ons PhD traject, en op hetzelfde project (zo leuk!). Jij op prostaat, ik op borst en samen met Diaz vormden we de eerste CMB dag orga. Daarna maakten we nog vele herinneringen, maar als hoogtepunten de tripjes: Berlijn, Lissabon, Boston en Denemarken, need i say moreeee? En dankzij jouw initiatief en organisatie skills, ook vele borrels, bbq's en zelfs een lesje aerial silk. Ook heb ik van jou geleerd dat niet alles direct resultaat hoeft te hebben, en dat brainstormen zonder conclusie soms ook oke is.

Het **NBCA-bestuur (Linetta, Janneke, Sabine, Enja, Rhode, Marije)** en **wetenschappelijke commissie**, jullie ambitie en medisch inhoudelijke expertise is heerlijk om mee samen te werken! Alle **IKNL en DICA collega's**, en in het speciaal **Marc:** na veel cappuccino's op de valreep nog onze gezamenlijke liefde ontdekt voor kokoscappu bij Lebkov en kokosmakronen (naast mosterd, jum)! Sparren over de inhoud, maar ook over ons persoonlijke leven, reisjes en toekomst. Dank! En dan **Hanneke**, met jou kom je verder! Je persoonlijke touch en scherpe kritische blik, waarbij je het ook nog op de goede manier weet te voorwoorden, zijn super waardevol! De samenwerking met ons low-incidence paper verliep dat ook als een speer (behoudens de submission ☺). En dankzij **Alex** worden mijn aanbevelingen ook nog eens geimplementeerd (go case-mix, go missing values in outcomes).

Hinne Rakhorst, onze wegen kruiste al in 2019. Van jou leerde ik prioriteren, efficiënt werken, dingen echt afmaken, mensen enthousiasmeren en verbinden en visie ontwikkelen. Allemaal onmisbare vaardigheden die ervoor gezorgd hebben dat ik mijn PhD

succesvol kon afronden. Ik heb menig moment gedacht, “Hinne zei altijd…”. En, niet te vergeten, je support en het meedenken over mijn carrièrepad.

Lieve vrienden (ja jullie allemaal, ook al staat je naam niet hier!): dank voor de ontspanning, gezelligheid, vele tripjes en afleiding & **Merel** en **Nienke**, dank voor de fantastische champagneballon!!!!

Papa en mama, die niet alles altijd begrijpen, maar wel trots op mij zijn! Mn **zusje**, die er vooral voor waakt dat ik niet teveel werk, en genoeg leuke dingen doe. **Hilbert**, als “oefenpubliek” voor mijn presentaties, en briljante ideeën voor workations (“want nu kan het nog”), waar ik dan toch altijd weer een manuscriptje wist te typen. En niet te vergeten, het begrip tijdens december 2024, waar de combi PhD, master (core competence exam) en sociaal leven toch een beetje optimistisch was. **Bas en Renny**, waardoor ik steeds beter werd mijn onderzoek in begrijpelijke taal uit te leggen (hoop ik?)

Tot slot: **voor elk succesje….**! Dank aan iedereen die alle mijlpalen mee gevierd heeft!