European Journal of Surgical Oncology 52 (2026) 111466

Contents lists available at ScienceDirect

European Journal of Surgical Oncology

journal homepage: www.ejso.com

# Development and validation of prediction models for health-related quality of life outcomes after breast cancer surgery and reconstruction

Elfi M. Verheul [a,b,*], Maria Margarete Karsten [c], Pimrapat Gebert [c,d], Lea Doppelbauer [c], Simona Borstnar [e], Sabine Siesling [f,g], Hester F. Lingsma [a], Noëlle J.M.C. Vrancken-Peeters [h], Anne M. Stiggelbout [i], Marc A.M. Mureau [j], Linetta B. Koppert [h], David van Klaveren [a], on behalf of the4D PICTURE Consortium [1]

[a] Center for Medical Decision Making, Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands
[b] Dutch Institute for Clinical Auditing, Leiden, the Netherlands
[c] Department of Gynecology with Breast Center, Charité - Universitätsmedizin Berlin, Berlin, Germany
[d] Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Berlin, Germany
[e] Department of Medical Oncology, Institute of Oncology Ljubljana, Ljubljana, Slovenia
[f] Department of Research, Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, the Netherlands
[g] Department of Health Technology and Services Research, Technical Medical Centre, University of Twente, Enschede, the Netherlands
[h] Department of Surgery, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands
[i] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands
[j] Department of Plastic and Reconstructive Surgery, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands

ARTICLE INFO

*Keywords:*
Breast cancer surgery
Health-related quality of life (HRQoL)
Patient-centered care
Prediction model
Patient-reported outcome measures (PROMs)

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.

* Corresponding author. Center for Medical Decision Making, Department of Public Health, Erasmus University Medical Center, Rotterdam, the Netherlands.
*E-mail address:* e.verheul@erasmusmc.nl (E.M. Verheul).
[1] **4D PICTURE Collaborator authorship:** the scientific members of the 4D PICTURE consortium who are not authors on this manuscript.

https://doi.org/10.1016/j.ejso.2026.111466
Received 25 September 2025; Received in revised form 12 January 2026; Accepted 9 February 2026
Available online 10 February 2026

## 1. 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.

## 2. Methods

### 2.1. 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; Fig. S1).

Ethics approval and consent to participate: The study is approved by the ethical committee of Erasmus MC (MEC-2023-0533).

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.

### 2.2. 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.

### 2.3. 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.

### 2.4. 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 $\chi^2$ 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.

## 3. Results

### 3.1. Patient characteristics and demographics

In total, 1835 patients were included (Berlin N = 795; Ljubljana N = 623; Rotterdam N = 417, Table 1; Fig. 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; Fig. S2A and S2B).

### 3.2. Model development and performance

The baseline score was by far the strongest predictor, followed by type of surgery (Fig. 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 (Fig. 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 (Fig. 2, S4, S5, S6, S7).

### 3.3. 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

**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 | 1328 (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 | 1259 (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 |
| BCS/ oncoplastic | 1199 (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 | 1376 (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 |

(continued on next page)

Table 1 (*continued*)

| Characteristic | All patients N = 1,835[1] | Berlin N = 795[1] | Ljubljana N = 623[1] | Rotterdam N = 417[1] | p-value[2] |
|---|---|---|---|---|---|
| (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 | |
| **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.

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 (Fig. 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 (Fig. S8–S10). Validation at 6 months showed similar results to the 12-month predictions (Fig. S11–S13).

## 4. Discussion

### *4.1. 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.

### *4.2. 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 multi-dimensional 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 (Fig. 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

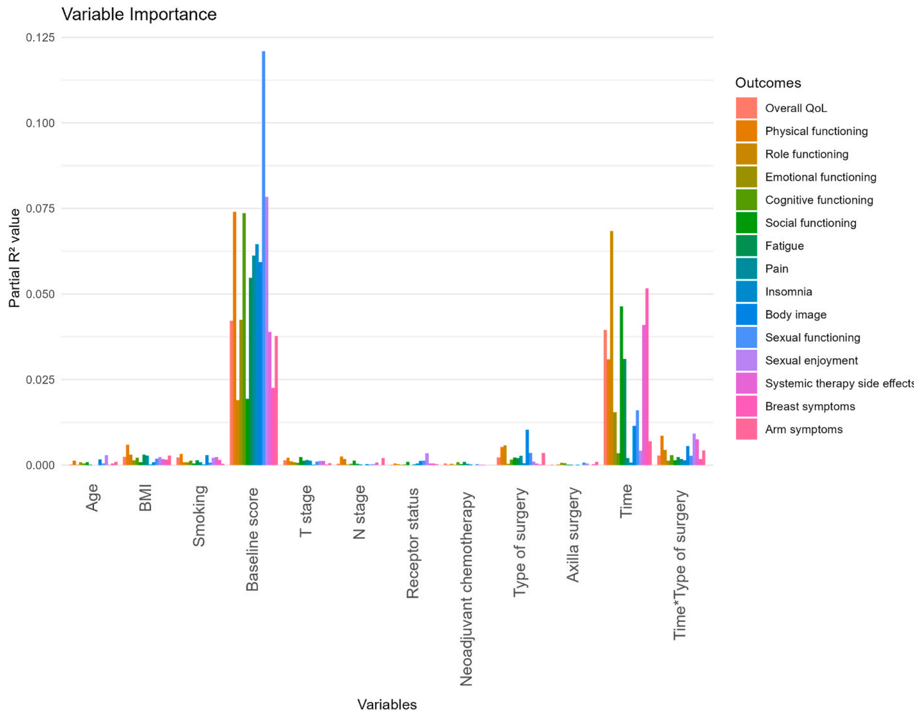


**Fig. 1.** The contribution of each variable to the model's performance measured by the partial $mR^2$. The partial $mR^2$ represents the proportion of variance in the outcome explained by each predictor. *QoL = Quality of life, Time*Type of surgery reflects the interaction between time and the type 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** | **$r^2$ T12** | **RMSE T12** | **$r^2$ T12** | **RMSE T12** | **$r^2$ T12** | **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 |

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.

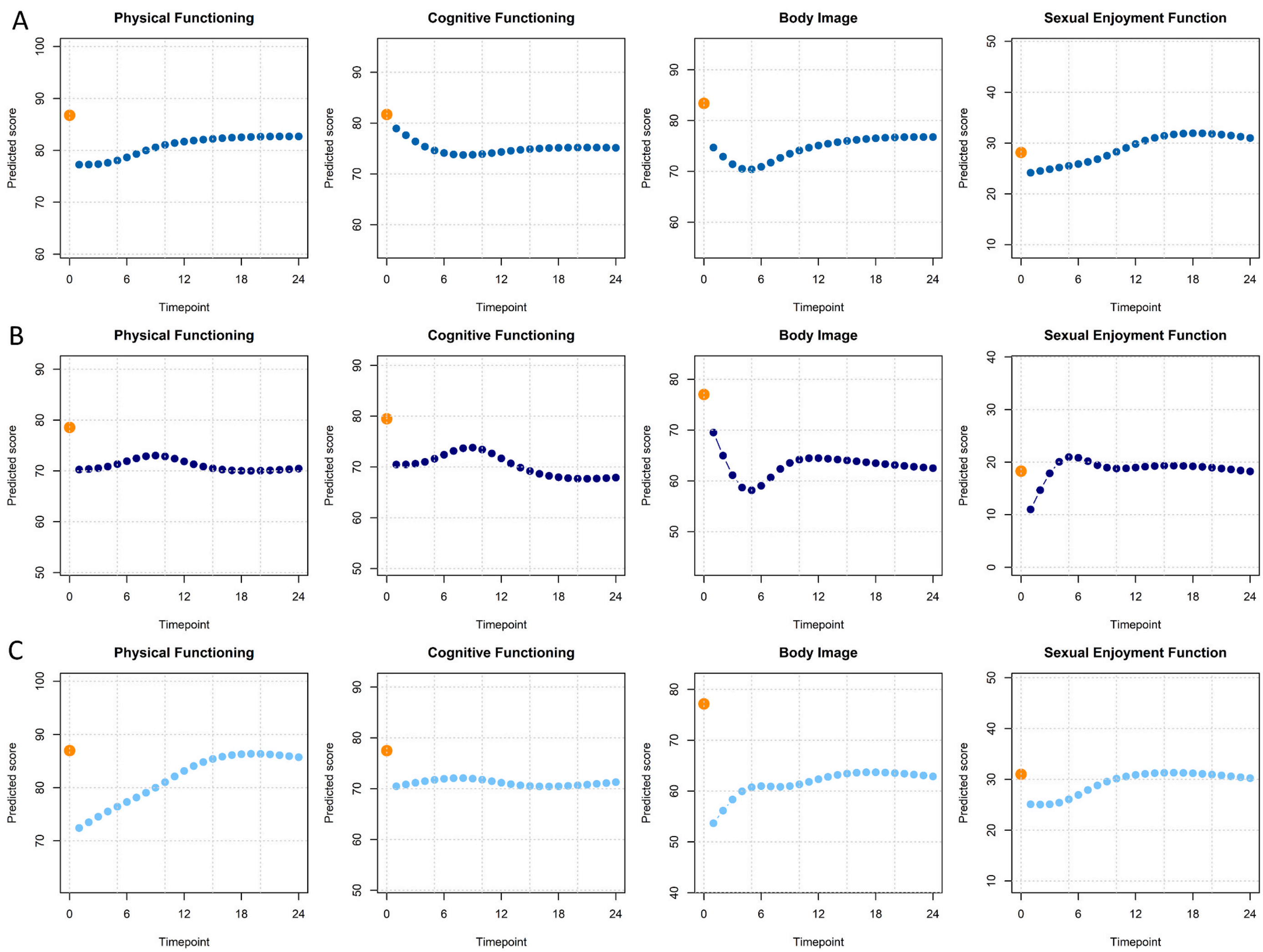


**Fig. 2.** Predictions from the core models, using baseline score and surgery type, for four Health-Related Quality of Life outcomes in an average patient undergoing (A) breast-conserving with or without oncoplastic surgery, (B) mastectomy, or (C) mastectomy with immediate implant-based reconstruction. *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.)*

### *4.3. 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.

### *4.4. 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

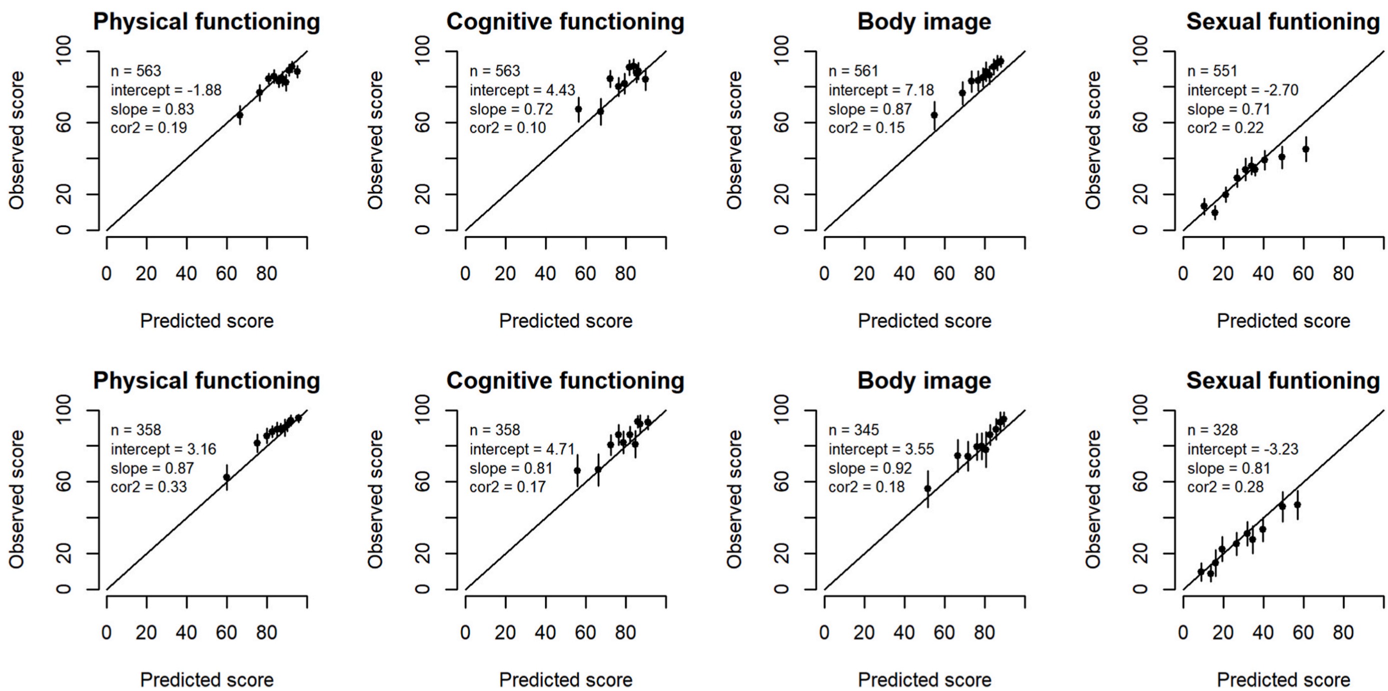


**Fig. 3.** Calibration plots at the 12-month timepoint for the Ljubljana validation cohort (top row) and the Rotterdam validation cohort (bottom row). The x-axis represents the predicted scores, and the y-axis represents the observed scores. The diagonal line indicates perfect calibration. Ideally, 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.

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.

### *4.5. 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.

## Data availability statement

We used data from clinical databases in three different countries. Data are owned by the contributing hospitals and are available upon reasonable request and subject to institutional approvals.

## Funding statement

Funded by the European Union under Horizon Europe Work Programme 101057332. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejso.2026.111466.

## References

[1] World Health Organization. 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: World Health Organization; 2023. Available from: https://www.who.int/publications/i/item/9789240067134.

[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 Outcome 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 Publ 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, et al. 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, et al. 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.