Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPiRED-001)[☆],[☆☆]

André Pfob [a], [b], Babak J. Mehrara [c], Jonas A. Nelson [c], Edwin G. Wilkins [d], Andrea L. Pusic [e], Chris Sidey-Gibbons [b], [f], [*]

[a] University Breast Unit, Department of Obstetrics & Gynecology, Heidelberg University Hospital, Heidelberg, Germany
[b] MD Anderson Center for INSPiRED Cancer Care (Integrated Systems for Patient-Reported Data), The University of Texas MD Anderson Cancer Center, Houston, USA
[c] Department of Plastic & Reconstructive Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[d] Department of Surgery, University of Michigan, Ann Arbor, MI, USA
[e] Patient-Reported Outcome Value & Experience (PROVE) Center, Department of Surgery, Harvard Medical School & Brigham and Women's Hospital, Boston, MA, USA
[f] Department of Symptom Research, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Article history:
Received 5 September 2021
Accepted 28 September 2021
Available online 29 September 2021

Keywords:
Decision-making
Breast reconstruction
Machine learning
Individualized treatment
Breast surgery
INSPiRED

ABSTRACT

Background: Women undergoing cancer-related mastectomy and reconstruction are facing multiple treatment choices where post-surgical satisfaction with breasts is a key outcome. We developed and validated machine learning algorithms to predict patient-reported satisfaction with breasts at 2-year follow-up to better inform the decision-making process for women with breast cancer.

Methods: We trained, tested, and validated three machine learning algorithms (logistic regression (LR) with elastic net penalty, Extreme Gradient Boosting (XGBoost) tree, and neural network) to predict clinically important differences in satisfaction with breasts at 2-year follow-up using the validated BREAST-Q. We used data from 1553 women undergoing cancer-related mastectomy and reconstruction who were followed-up for two years at eleven study sites in North America from 2011 to 2016. 10-fold cross-validation was used to train and test the algorithms on data from 10 of the 11 sites which were further validated using the additional site's data. Area-under-the-receiver-operating-characteristics-curve (AUC) was the primary outcome measure.

Results: Of 1553 women, 702 (45.2%) experienced an improved satisfaction with breasts and 422 (27.2%) a decreased satisfaction. In the validation set ($n = 221$), the algorithms showed equally high performance to predict improved or decreased satisfaction with breasts (all $P > 0.05$): For improved satisfaction AUCs were 0.86–0.87 and for decreased satisfaction AUCs were 0.84–0.85.

Conclusion: Long-term, individual patient-reported outcomes for women undergoing mastectomy and breast reconstruction can be accurately predicted using machine learning algorithms. Our algorithms may be used to better inform clinical treatment decisions for these patients by providing accurate estimates of expected quality of life.

[☆] Preliminary results presented at the International Society for Quality of Life Research Annual conference 2019, oral presentation "Cutting edge research plenary" on Monday, October 21, 2019, San Diego.

[☆☆] Preliminary results were selected for presentation as part of a Poster Discussion Session at the 2020 ASCO Annual Meeting, abstract number 520.

[*] Corresponding author. The University of Texas MD Anderson Cancer Center, Department of Symptom Research, 1515 Holcombe Boulevard, Unit 1450, Houston, TX, 77030, USA.

E-mail address: cgibbons@mdanderson.org (C. Sidey-Gibbons).

https://doi.org/10.1016/j.breast.2021.09.009

Abbreviations

CI	confidence interval
LR	logistic regression
LIME	local interpretable model agnostic explanation
PRO	patient-reported outcome
PROM	patient-reported outcome measurement
SHAP	Shapley Additive Explanations
MCID	minimal clinically important difference

1. Background

Breast cancer surgery has undergone many changes over the past decades. Foremost among there is the trend towards de-escalation of surgical severity; (radical) mastectomy was once standard of care but many women now receive breast conserving therapy [1]. In spite of de-escalation trends still about 35% of all breast cancer patients have to undergo primary, secondary, or prophylactic mastectomy [2]. The negative impact of mastectomy on quality of life is well documented and reconstruction is often offered to minimize such impacts [3,4]. However, women undergoing cancer-related mastectomy and reconstruction face difficult treatment choices like implant-based vs. autologous reconstruction or timing of radiotherapy. While several trials and observational studies have provided evidence to inform these treatment decisions, tailoring multi-disciplinary care to the unique needs and preferences of individuals undergoing breast reconstruction remains complex with many knowledge gaps [5–7].

Modern predictive modeling using machine learning techniques may help reduce complexity, overcome knowledge gaps, and improve clinical care and outcomes for patients undergoing cancer-related breast reconstruction. Machine learning algorithms identify complex patterns in data to make accurate outcome predictions of future events at an individual level [8–10]. Such algorithms have shown great performance in other areas of breast cancer treatment like identifying exceptional responders to neoadjuvant treatment or patients at risk of experiencing financial toxicity related to their cancer treatment [11,12]. As post-surgical satisfaction with breasts is a recommended key outcome for women undergoing cancer-related mastectomy and breast reconstruction [13], we hypothesized that machine learning algorithms may allow accurate, individualized predictions of long-term satisfaction with reconstructed breasts prior to the initiation of the breast reconstruction process to better inform the decision-making process for these women.

In this study, we developed and validated machine learning algorithms to predict individual, patient-reported satisfaction with breasts at 2-year follow-up to better inform decision-making for women undergoing cancer-related mastectomy and subsequent implant-based or autologous breast reconstruction.

2. Methods

2.1. Patient recruitment and selection

Patients were recruited as part of an international multicenter trial (Mastectomy Reconstruction Outcomes Consortium (MROC) study, NCT01723423). MROC patient recruitment was from 2012 to 2015, while data collection went from 2012 to 2017. The trial was conducted at 11 study sites across the US and Canada and included women who underwent cancer-related mastectomy and subsequent breast reconstruction. Study participants were followed-up for two years after surgery. Choice of breast reconstructive procedure was not randomly assigned but based on surgeon and patient preferences. Further details are published elsewhere [5,14]. Ethics approval was obtained from all study sites and all patients gave their written informed consent.

Inclusion criteria of the MROC study were patient age 18 years or older, prophylactic or therapeutic, bilateral or unilateral, immediate or delayed, implant-based or autologous breast reconstruction. Exclusion criteria included previously failed breast reconstruction attempts.

2.2. Design and definitions

In this analysis, we developed and validated machine learning algorithms to predict changes in individual patient-reported satisfaction with reconstructed breasts at 2-year follow-up prior to the initiation of the breast reconstruction process.

Patient-reported satisfaction with breasts was measured using the BREAST-Q ‘Satisfaction with Breasts’ subscale, a gold-standard measure for assessing patient-reported outcomes for women with breast cancer [15,16]. As described previously, we defined three types of outcome: improved, worsened, or stable satisfaction [17]. Changes greater than or equal to the minimal clinically important difference (MCID) were used as the outcome for both worsened (negative clinically-meaningful change in post-operative scores compared to baseline) and improved satisfaction with breasts (positive change compared to pre-operative score) [18].

For our predictive models, we used known clinical, patient, and patient-reported variables influencing the post-operative outcome after breast reconstruction [19,20]. A full list of co-variables and the outcome variable is shown in Table 1. In line with recent recommendations to avoid racial bias in predictive models we did not include socio-economic and ethnic information into the models but rather compared the predictive performance among different ethnic groups to ensure a fair algorithm performance [21,22]. The surgeon was not included as a predictive factor due to lacking objectivity and consensus with regard to suitable metrics for assessing a surgeon's performance (years of practice vs. caseload vs. patient satisfaction) and thus limited generalizability for future model iterations.

2.3. Algorithm development

Choice of algorithms, algorithm development, and reporting on them was informed by recent guidelines on how to use machine learning in medicine [23], how to report findings of diagnostic tests [24] and multivariate prediction models [25], and previously published research by our group [9,11,12,17]. We developed and validated three algorithms to predict clinically meaningful changes in satisfaction with reconstructed breasts:

1) Logistic Regression (LR) with Elastic Net Penalty: We chose this algorithm because of its known ability to attenuate the influence of certain predictors on the model, leading to greater generalizability to new datasets [26,27].
2) Extreme Gradient Boosting (XGBoost) Tree: Decision trees are commonly used because of their ability to identify more complex, non-linear relations between variables while still being readily interpretable. Gradient boosting, a machine learning technique where the final prediction model consists of an ensemble of several stepwise built models, has been shown to further improve the predictive performance of decision trees [28]. We used Shappley Additive Explanation (SHAP) values to provide insights into the model predictions [29].
3) Neural network: Neural networks are state-of-the-art algorithms that mimic the structure of the mammalian cortex and

Table 1
Baseline demographic and clinical characteristics of participating women.

	Whole cohort (n = 1553)	Development set (n = 1332)	Validation set (n = 221)	*P* value[c]
Patient variables				
Age[e], mean (SD), years	50.19 (9.98)	50.22 (10.10)	49.99 (9.22)	0.732[#]
BMI[e], mean (SD), kg/m^2	26.49 (5.39)	26.61 (5.48)	25.72 (4.76)	**0.012**[#]
Diabetes[e], no (%)				0.577[a]
No, no. (%)	1482 (95.4)	1269 (95.3)	213 (96.4)	
Yes, no. (%)	71 (4.6)	63 (4.7)	8 (3.6)	
Smoker[e]				
Never, no. (%)	1032 (66.5)	888 (66.7)	144 (65.2)	0.726[a]
Previous, no. (%)	482 (31.0)	409 (30.7)	73 (33.0)	0.534[a]
Current, no. (%)	26 (1.7)	24 (1.8)	2 (0.9)	0.498[a]
Unknown, no. (%)	13 (0.8)	11 (0.8)	2 (0.9)	1
Pre-operative patient-reported outcome data				
BREAST-Q satisfaction with breast[e], mean (SD), 0–100	59.96 (22.12)	59.76 (22.10)	61.33 (22.22)	0.330[b]
BREAST-Q psychosocial well-being[e], mean (SD), 0–100	69.57 (18.08)	69.49 (18.13)	70.04 (17.81)	0.673[b]
BREAST-Q physical well-being chest and upper body[e], mean (SD), 0–100	78.87 (14.55)	78.50 (14.56)	81.10 (14.31)	**0.013**[b]
BREAST-Q physical well-being abdomen, mean (SD)[e], 0–100	89.53 (13.46)	89.27 (14.56)	91.11 (11.37)	**0.033**[b]
BREAST-Q sexual well-being[e], mean (SD), 0–100	54.90 (20.73)	55.03 (20.59)	54.11 (21.55)	0.562[b]
Clinical variables				
Radiation[e]				
After reconstruction, no. (%)	293 (18.9)	252 (18.9)	41 (18.6)	0.971[a]
Before reconstruction, no. (%)	220 (14.2)	186 (14.0)	34 (15.4)	0.648[a]
None, no. (%)	1040 (67.0)	894 (67.1)	146 (66.1)	0.817[a]
Mastectomy[e]				
Nipple-sparing, no. (%)	170 (10.9)	149 (11.2)	21 (9.5)	0.531[a]
Simple, no. (%)	1377 (88.7)	1178 (88.4)	199 (90.0)	0.560[a]
Other, no. (%)	6 (0.4)	5 (0.4)	1 (0.5)	1
Reconstruction technique[e]				
Tissue expander (TE), no. (%)	831 (53.5)	699 (52.5)	132 (59.7)	0.054[a]
Direct-to-implant (DTI), no. (%)	71 (4.6)	62 (4.7)	9 (4.1)	0.833[a]
Transverse rectus abdominis (TRAM) flap, no. (%)	121 (7.8)	100 (7.5)	21 (9.5)	0.374[a]
Deep inferior epigastric perforator (DIEP) flap, no. (%)	291 (18.7)	251 (18.8)	40 (18.1)	0.865[a]
Latissimus dorsi (LD) flap, no. (%)	49 (3.2)	46 (3.5)	3 (1.4)	0.149[a]
Gluteal artery perforator (GAP) flap, no. (%)	8 (0.5)	7 (0.5)	1 (0.5)	1[a]
Superficial inferior epigastric artery (SIEA) flap, no. (%)	48 (3.1)	48 (3.1)	0 (0.0)	**0.008**[a]
Crossover flap, no. (%)	60 (3.9)	52 (3.9)	8 (3.6)	0.988[a]
Mixed flaps, no. (%)	46 (3.0)	39 (2.9)	7 (3.2)	1[a]
Mixed implant and autologous, no. (%)	28 (1.8)	28 (2.1)	0 (0.0)	0.057[a]
Chemotherapy[e]				0.584
Received, no. (%)	442 (28.5)	383 (28.8)	59 (26.7)	
Not received, no. (%)	1111 (71.5)	949 (71.2)	162 (73.3)	
Reconstruction laterality[e]				0.390[a]
Unilateral, no. (%)	700 (45.1)	594 (44.6)	106 (48.0)	
Bilateral, no. (%)	853 (54.9)	738 (55.4)	115 (52.0)	
Mastectomy indication[e]				0.706[a]
Therapeutic, no. (%)	1398 (90.0)	1197 (89.9)	201 (91.0)	
Prophylactic, no. (%)	155 (10.0)	135 (10.1)	20 (9.0)	
Axillary intervention[e]				
Axillary lymph node dissection (ALND), no. (%)	402 (25.9)	358 (26.9)	44 (19.9)	**0.035**[a]
Sentinel lymph node biopsy (SLNB), no. (%)	698 (44.9)	584 (43.8)	114 (51.6)	**0.039**[a]
None, no. (%)	453 (29.2)	390 (29.3)	63 (28.5)	0.878[a]
Socioeconomic and ethnic data				
Marital status				
Single, no. (%)	109 (7.1)	100 (7.5)	9 (4.1)	0.084[a]
Living with significant other, no. (%)	67 (4.3)	59 (4.5)	8 (3.6)	0.701[a]
Married, no. (%)	1176 (76.1)	1004 (75.8)	172 (77.8)	0.564[a]
Separated, no. (%)	27 (1.7)	23 (1.7)	4 (1.8)	1[a]
Divorced, no. (%)	125 (8.1)	104 (7.8)	21 (9.5)	0.483[a]
Widowed, no. (%)	42 (2.7)	35 (2.6)	7 (3.2)	0.825[a]
Education level				
Some high school, no. (%)	31 (2.0)	29 (2.2)	2 (0.9)	0.319[a]
High school degree, no. (%)	121 (7.8)	112 (8.4)	9 (4.1)	**0.036**[a]
Some college/trade school, no. (%)	255 (16.5)	226 (17.0)	29 (13.1)	0.179[a]
College/trade school degree, no. (%)	602 (38.8)	517 (38.9)	85 (38.5)	0.960[a]
Some masters/doctoral, no. (%)	61 (3.9)	54 (4.1)	7 (3.2)	0.655[a]
Masters/doctoral degree, no. (%)	480 (31.0)	391 (29.4)	89 (40.3)	**0.002**[a]
Working status				
Unable to work, no. (%)	37 (2.4)	29 (2.2)	8 (3.6)	0.295[a]
Unemployed, no. (%)	31 (2.0)	28 (2.1)	3 (1.4)	0.627[a]
Student, no. (%)	10 (0.7)	9 (0.7)	1 (0.5)	1[a]
Volunteer, no. (%)	8 (0.5)	7 (0.5)	1 (0.5)	1[a]
Retired, no. (%)	141 (9.2)	130 (9.9)	11 (5.0)	**0.028**[a]
Homemaker, no. (%)	179 (11.6)	152 (11.5)	27 (12.3)	0.842[a]

(*continued on next page*)

Table 1 (*continued*)

	Whole cohort (n = 1553)	Development set (n = 1332)	Validation set (n = 221)	*P* value[c]
Part time employed, no. (%)	216 (14.1)	175 (13.3)	41 (18.6)	**0.044**[a]
Full time employed, no. (%)	863 (56.1)	746 (56.6)	117 (53.2)	0.376[a]
Other, no. (%)	52 (3.4)	41 (3.1)	11 (5.0)	0.218[a]
Household income per year				
<25,000$, no. (%)	81 (5.4)	74 (5.8)	7 (3.2)	0.171[a]
25,000$ to 49,999$, no. (%)	163 (10.9)	149 (11.6)	14 (6.5)	**0.032**[a]
50,000$ to 74,999$, no. (%)	269 (17.9)	232 (18.1)	37 (17.1)	0.787[a]
75,000$ to 99,999$, no. (%)	233 (15.5)	207 (16.1)	26 (12.0)	0.144[a]
>100,000$, no. (%)	754 (50.3)	621 (48.4)	133 (61.3)	**<0.001**[a]
Ethnical background				
Caucasian, no. (%)	1398 (90.9)	1199 (90.8)	199 (91.7)	0.750[a]
African American, no. (%)	69 (4.5)	61 (4.6)	8 (3.7)	0.662[a]
Asian, no. (%)	63 (4.1)	54 (4.1)	9 (4.1)	1[a]
American Indian, no. (%)	8 (0.5)	7 (0.5)	1 (0.5)	1[a]
Outcome – patient-reported satisfaction with breasts at 2-year follow-up compared to baseline				
Improved[d], no. (%)	702 (45.2)	602 (45.2)	100 (45.2)	1[a]
Decreased[d], no. (%)	422 (27.2)	357 (26.8)	65 (29.4)	0.468[a]
Stable, no. (%)	429 (27.6)	373 (28.0)	56 (25.3)	0.460[a]

ALND = axillary lymph node dissection; SLNB = sentinel lymph node biopsy.
P values < 0.05 highlighted in bold.

[a] *P* values refer to Chi-square tests for binary feature evaluation (feature true vs. feature not true).

[b] *P* values refer to t-tests to evaluate mean differences of continuous data.

[c] *P* values refer to differences in the development and validation set.

[d] Increase or decrease equal or larger to minimal clinically important difference.

[e] Variable used in the predictive models.

that have shown great ability to detect even the most complex patterns in data. Neural networks are, however, prone to overfitting and can be difficult to interpret. We used Local Interpretable Model Interpretation (LIME) to provide insights into the model predictions [30,31].

According to guidelines for multivariable risk prediction models, validation of such a model is recommended in a dataset of at least 100 events [25]. For improved satisfaction with breasts, this requirement was satisfied by three trial sites of this large, international, multicenter trial of which the one with a maximum amount of events for decreased satisfaction was chosen as an independent validation set on which the final model was (externally) validated. The other 10 trial sites were used as development set (Fig. 1).

Data preparation steps (separately applied to every fold of the cross-validation process) included imputation of missing data using k-nearest neighbors (5 neighbors), removal of zero-variance variables, one hot-encoding for categorical variables, as well as feature scaling and centering for continuous variables. We used 10-fold cross-validation to train and tune the algorithms on the development set; we used a hypergrid-search for hyperparameter tuning (see Supplementary Appendix for optimal hyperparameters). To address possible class-imbalances we used the Kappa performance metric (mean value over the 10 folds) to select the final model. To avoid overfitting and to improve generalizability of our models we applied a tolerance threshold of 3% meaning that the simplest model within a 3% tolerance of the empirically optimal model was chosen as the final model [32]. A more detailed description of the algorithm development and compliance with the above-mentioned guidelines [23–25] can be found in the online Supplementary Appendix.

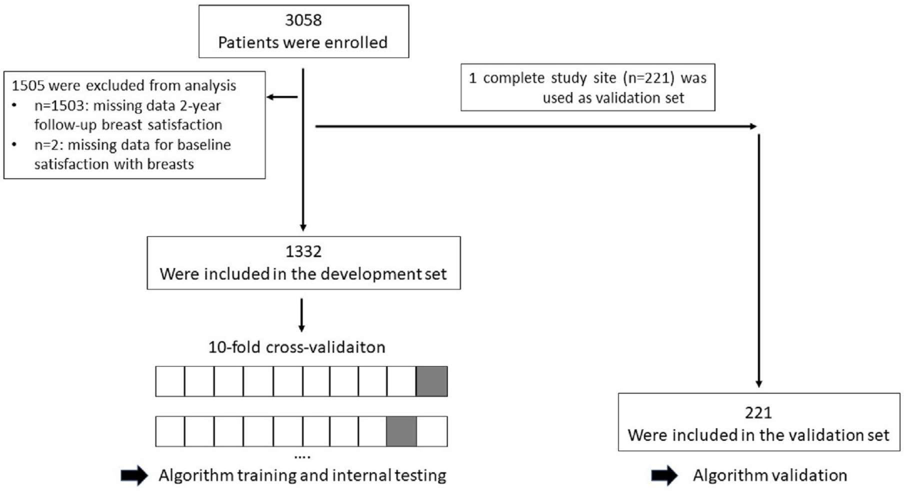
3058
Patients were enrolled
1505 were excluded from analysis
• n=1503: missing data 2-year follow-up breast satisfaction
• n=2: missing data for baseline satisfaction with breasts
1 complete study site (n=221) was used as validation set
1332
Were included in the development set
10-fold cross-validaiton
....
Algorithm training and internal testing
221
Were included in the validation set
Algorithm validation

Fig. 1. Flow of participants.

2.4. Statistical analysis

We used descriptive statistics including absolute and relative frequencies as well as chi-square tests for categorical data and mean and standard deviation alongside t-tests for continuous data to compare the distribution of baseline and outcome variables in the development and validation set.

To assess the diagnostic performance of machine learning algorithms predicting clinically meaningful changes in patient-reported satisfaction with breasts, area under the receiver-operating characteristics curve (AUC) and accompanying 95% confidence intervals were calculated for every model using 2000 bootstrap replicates that were drawn from the validation dataset and stratified for the outcome variable (clinically meaningful increase/decrease).

Additionally, model overall accuracy was computed by comparing the model predictions to the actual patient-reported outcome at 2-year follow up; 95% Clopper-Pearson confidence intervals were computed.

We used calibration plots (observed vs. predicted probabilities [33]) and Spiegelhalter's Z statistic [34] to evaluate model calibration.

No multiplicity adjustments against type-I-error inflation were performed. All *P* values should be interpreted descriptively and have no confirmatory value. A *P* value smaller than 0.05 was considered statistically significant. Analysis was conducted using R software, Version 3.6.1; the “caret” package of R was used for the model development.

3. Results

3.1. Clinical and demographic characteristics

Of 3058 women enrolled, 1503 returned both baseline and 2-year follow-up information for satisfaction with breasts. Of those 1503 patients, 1332 patients from 10 trial sites were included in the development set for the algorithm training and initial testing and one complete trial site (n = 221) was set aside as validation set (Fig. 1).

Clinical and demographic characteristics in the whole cohort as well as in the development and validation datasets are listed in Table 1. A total of 702 (45.2%) women experienced a clinically meaningful improved satisfaction with breasts two years after surgery, 422 (27.2%) a clinically meaningful decreased satisfaction, and 429 (27.6%) a stable satisfaction with breasts. In the validation set, women had a significantly lower BMI (26.6 vs. 25.7), a higher pre-operative physical well-being for chest and upper body (78.5 points vs. 81.1) as well as abdomen (89.3 points vs. 91.1), they underwent autologous reconstructions with superficial inferior epigastric artery flaps less often (0% vs. 3.1%), and received SLNB (51.6% vs. 43.8%) instead of ALND (19.9% vs. 26.9%) more often.

3.2. Algorithm performance

Table 2 summarizes the performance of the LR with elastic net penalty, the XGBoost tree, and the neural network algorithm in the prediction of improved or worsened satisfaction with breasts at 2-year follow up.

For the prediction of worsened satisfaction in the validation set, the LR with elastic net penalty, the XGBoost tree, and the neural network showed an AUC of 0.84 (95%CI 0.78–0.90), 0.84 (95%CI 0.78–0.90), and 0.85 (95%CI 0.78–0.90), respectively. For the prediction of improved satisfaction in the validation set, the three algorithms showed an AUC of 0.87 (95%CI 0.82–0.91), 0.86 (95%CI 0.81–0.91), and 0.87 (95%CI 0.83–0.92) for improved satisfaction. When compared against each other, the performance of the different algorithms did not differ significantly (Fig. 2). Accompanying ROC curves are illustrated in Fig. 3.

Calibration plots of all algorithms in the validation set yielded good calibration (Fig. 4). Spiegelhalter's Z statistic confirmed a well-calibrated model for the XGBoost tree algorithm to predict worsened or improved satisfaction (z score 0.127 and −0.152, *P* value 0.449 and 0.440) but not for the other algorithms.

3.3. Predictive coefficients and insights into variable importance

The predictive coefficients of the LR with elastic net penalty (Table 3) illustrate that baseline satisfaction with breasts was most strongly associated with the outcomes at 2-year follow up followed by type of reconstruction, timing of radiotherapy, smoker status, mastectomy indication, and laterality. Specifically, a low baseline satisfaction with breasts, autologous reconstruction with TRAM or DIEP flaps, radiation before reconstruction, and never having smoked were associated with improved satisfaction at 2-year follow up. Contrary, a high baseline satisfaction with breasts, implant-based reconstruction using tissue expanders, autologous reconstruction using mixed flaps, radiation after reconstruction, a current smoker status, and unilateral reconstruction were associated with worsened outcomes at 2-year follow up.

Figs. 5 and 6 provide insights into the variable importance of the XGBoost tree and the neural network using local interpretable methods. Generally, the same effects as for the logistic regression with elastic net penalty could be observed.

For comparison, the results of a traditional multivariable regression with decreased satisfaction at 2-year follow up as outcome variable are listed in Table 4.

3.4. Subgroup analysis

To assess potential racial bias of the predictive models, we evaluated the algorithms' performance among different ethnic groups. The logistic regression with elastic net penalty, the XGboost tree, and the neural network showed significantly better performance among African American women compared to Caucasian women (AUC 1 vs. 0.84; $P < 0.001$). No significant differences could be observed when comparing the algorithms’ performance among Asian and Caucasian women or Asian and African American women (all $P > 0.05$).

4. Discussion

In this study, we demonstrate that machine learning algorithms can accurately predict individual, long-term patient-reported outcomes for women undergoing cancer-related mastectomy and subsequent implant-based or autologous breast reconstruction. The strength of these algorithms is that they may better inform clinical treatment decisions (e.g., autologous vs. implant-based reconstruction) for these patients by using contextualizing clinical, patient, and patient-reported variables to provide patient-relevant outcome predictions tailored to the individual patient's situation. Insights into the predictions made by the algorithms showed that baseline patient-reported variables were more important than clinical treatment decisions.

Identifying and recommending optimal treatment decisions for women undergoing cancer-related mastectomy and breast reconstruction has been a major focus of clinical research during the past years [7]. For example, previous clinical trials have found that autologous reconstruction is generally associated with better outcomes compared to implant-based reconstruction, that radiotherapy is generally associated with poorer quality of life outcomes

Table 2
Evaluation of algorithms trained to predict satisfaction with breasts at two-year follow-up.

	2-year follow-up satisfaction lower than baseline		2-year follow-up satisfaction higher than baseline	
	Accuracy (95% CI)	AUC (95% CI)	Accuracy (95% CI)	AUC (95% CI)
Logistic regression with elastic net penalty				
Test set (n = 1332)	0.84 (0.83–0.85)	0.85 (0.84–0.87)	0.77 (0.76–0.78)	0.85 (0.84–0.86)
Additional validation set (n = 221)	0.83 (0.78–0.88)	0.84 (0.78–0.90)	0.78 (0.72–0.83)	0.87 (0.82–0.91)
XGBoost Tree				
Test set (n = 1332)	0.84 (0.82–0.85)	0.85 (0.84–0.87)	0.76 (0.75–0.78)	0.85 (0.83–0.86)
Additional validation set (n = 221)	0.83 (0.77–0.88)	0.84 (0.78–0.90)	0.77 (0.71–0.83)	0.86 (0.81–0.91)
Neural network				
Test set (n = 1332)	0.83 (0.82–0.84)	0.86 (0.85–0.87)	0.76 (0.74–0.77)	0.84 (0.83–0.86)
Additional validation set (n = 221)	0.84 (0.78–0.88)	0.85 (0.78–0.90)	0.78 (0.72–0.84)	0.87 (0.83–0.92)

AUC = Area under the Receiver Operating Characteristic Curve.

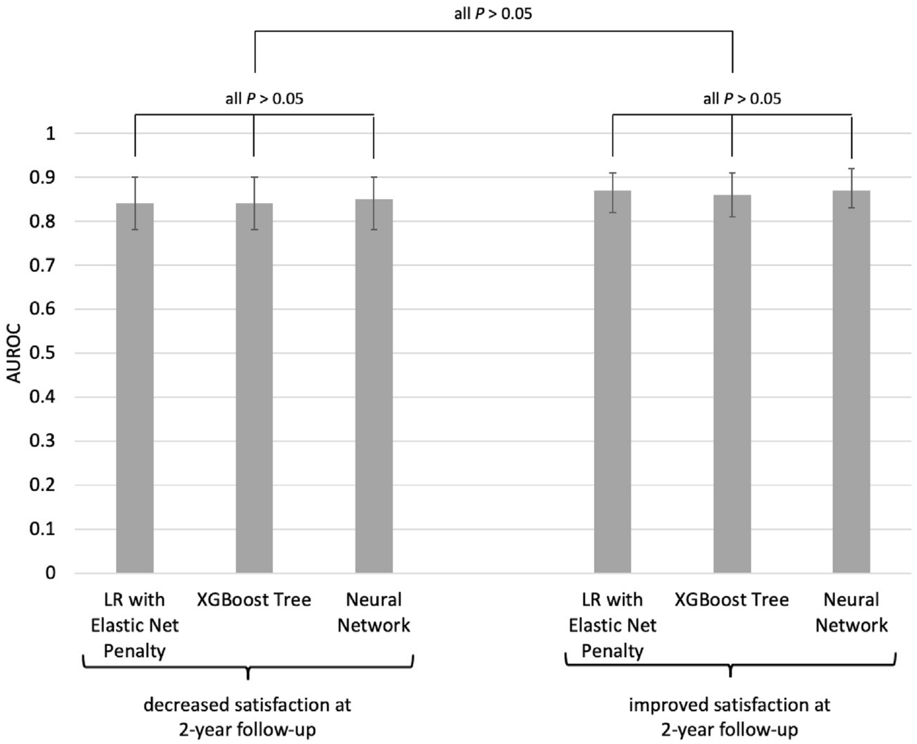

Fig. 2. Performance comparison between the models to predict improved and decreased satisfaction with breasts at 2-year follow-up.

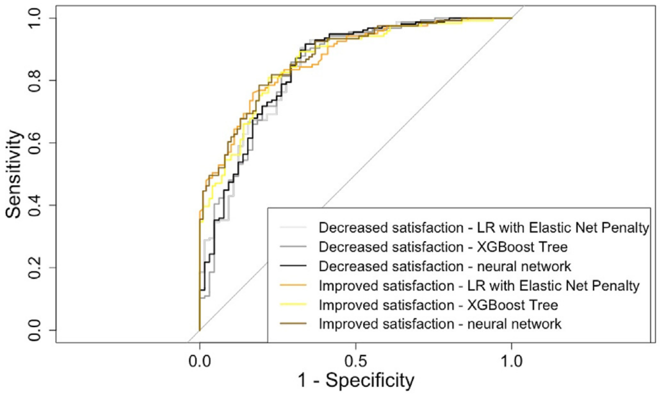

Fig. 3. Receiver Operating Characteristic Curves of the models to predict improved and decreased satisfaction with breasts at 2-year follow-up.

(but the optimal timing, before or after reconstruction, remains unclear), and that nipple-sparing mastectomy may offer no benefit over simple mastectomy [6,19,20,35,36]. However, clinical applicability of these findings may have been limited so far due to inherently limited group-level inferences of traditional statistics [37]. For example, while autologous reconstruction may have statistically better outcomes compared to implants in a prospective study such a finding does not mean that all women will have better outcomes if they receive an autologous reconstruction. Traditional primary endpoint evaluations are performed under the assumption that all other co-variables (e.g. radiation, type of mastectomy, baseline satisfaction etc.) are held constant to eliminate their influence on the outcome. Thus, while it is important to know that autologous reconstruction was generally associated with better outcomes in prospective studies if all other influences were statistically eliminated, this may not fully represent the complex interactions between different treatment choices and the individual patient's situation within the breast reconstruction process. Accurate predictive models like our algorithms may better inform the

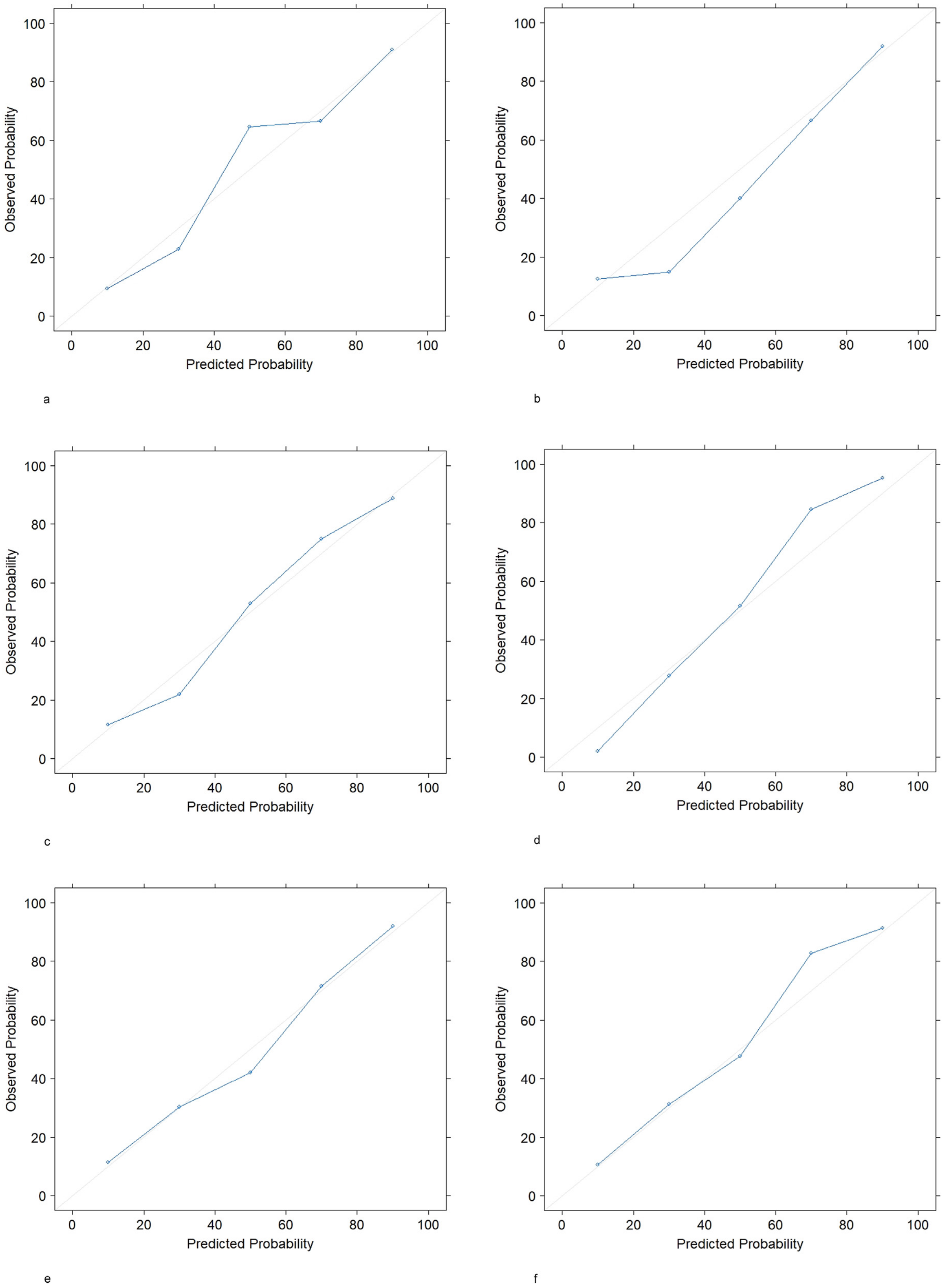

Fig. 4. Calibration Plots of the Machine Learning Models in the Validation Set. 4a. Decreased satisfaction – Logistic Regression with Elastic Net Penalty.4b. Decreased satisfaction – XGBoost Tree.4c. Decreased satisfaction – neural network.4d. Improved satisfaction - Logistic Regression with Elastic Net Penalty. 4e. Improved satisfaction – XGBoost Tree.4f. Improved satisfaction – neural network.

Table 3
Regularized coefficients from the logistic regression with elastic net penalty.

	Regularized coefficient for lower satisfaction at 2-year follow-up (positive values indicate a positive correlation with low satisfaction)	Regularized coefficient for higher satisfaction at 2-year follow-up (positive values indicate a positive correlation with high satisfaction)
Patient variables		
Age	0.01	0.0
BMI	0.05	0.0
Diabetes	0.0	0.0
Smoker		
Never	−0.13	0.14
Previous	0.0	0.0
Current	0.42	−0.37
Patient-reported outcomes at baseline		
Satisfaction with breasts	1.44	−1.26
Psychosocial well-being	0.0	0.0
Physical well-being chest and upper body	−0.04	0.0
Physical well-being abdomen	0.01	0.0
Sexual well-being	0.0	0.0
Clinical variables		
Radiation		
After reconstruction	0.52	−0.42
Before reconstruction	−0.22	0.02
None	0.0	0.0
Mastectomy		
Nipple-sparing	0.0	−0.01
Simple	0.0	0.0
Other	0.11	0.0
Reconstruction – Implant-based		
Tissue expander	0.40	−0.28
Direct-to-implant	0.0	0.0
Reconstruction – Autologous (flap)		
TRAM	−0.76	0.16
DIEP	−0.22	0.23
LD	0.0	0.0
GAP	0.0	0.0
SIEA	−1.00	0.0
Crossover	0.0	0.0
Mixed flaps	0.28	0.0
Reconstruction – Mixed implants and autologous	−0.11	0.01
Chemotherapy		
Received	−0.12	−0.03
Not received	0.12	0.03
Laterality		
Unilateral reconstruction	0.16	−0.05
Bilateral reconstruction	−0.16	0.05
Mastectomy indication		
Therapeutic	0.19	0.0
Prophylactic	−0.19	0.0
Axillary intervention		
Axillary lymph node dissection	0.0	−0.13
Sentinel lymph node biopsy	0.01	0.0
No axillary intervention	0.0	0.21

decision-making process for patients and clinicians. A clinical trial comparing the outcomes of “intelligent” (algorithm-supported) decision-making against traditional decision-making for women undergoing mastectomy and breast reconstruction seems warranted.

Our group has previously reported the results of predicting short-term patient-reported outcomes for women undergoing breast reconstruction [17]. Our present analysis does not only demonstrate that longer-term outcomes can be predicted with great accuracy but also that the drivers for predicting short- or long-term outcomes are different. For example, direct-to-implant reconstruction was strongly associated with worse short-term but not long-term outcomes (regularized $\beta = 0.34$ vs. 0.0) whereas the negative impact of implant-based reconstruction with tissue expanders on patient-reported satisfaction with breasts worsened in the long-term follow up (regularized $\beta = 0.08$ vs. 0.40). For autologous reconstruction, TRAM, DIEP, and SIEA flaps were associated with a decreased risk of a poorer short-term and long-term outcome. GAP and crossover flaps were previously found to be associated with worsened short-term follow up; this effect could not be observed for 2-year outcomes. Interestingly, also the importance of patient and patient-reported variables changed over

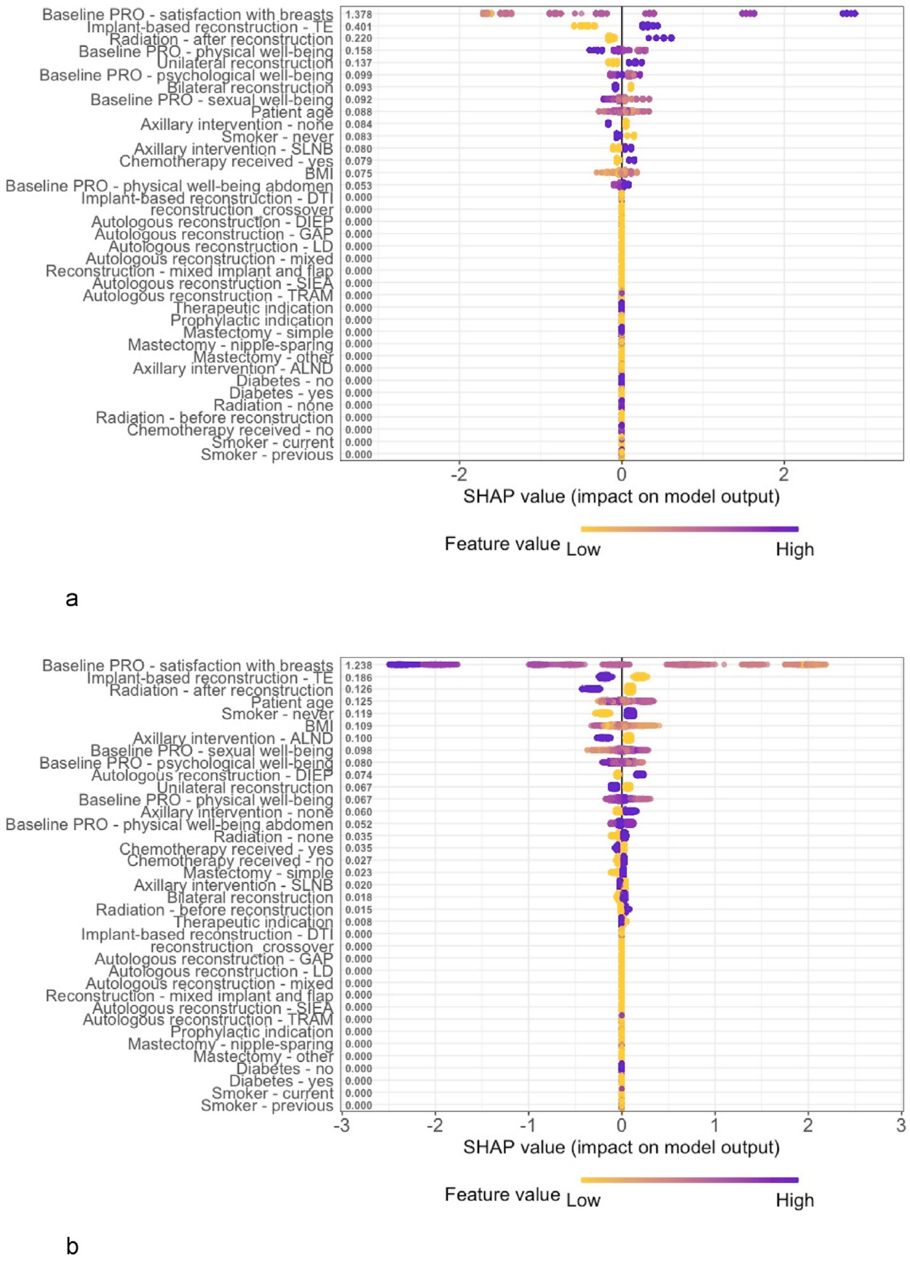

Fig. 5. Shapley Additive Explanations (SHAP) Value Summary Plot of the Extreme Gradient Boosting (XGBoost) Tree Model. 5a. Decreased satisfaction.5b. Improved satisfaction.

time: A higher patient age was strongly associated with poorer short-term outcomes but not with long-term outcomes (regularized $\beta = 0.66$ vs. 0.01) and patient-reported baseline variables other than satisfaction with breasts (psychosocial, physical, and sexual well-being) were less important in predicting long-term outcomes.

In interpreting our findings, some limitations of our study should be considered.

First, although we used the largest dataset of its kind to develop and validate our algorithms, some procedures showed a limited sample size in the development and validation set. Moreover, most of our study sites were high-volume academic centers located in North America and although the algorithms performed equally well among different ethnic groups, our study population showed limited ethnic diversity. Future research may verify our findings in more diverse populations.

Second, only about 50% of study participants completed long-term follow-up of the patient-reported outcome satisfaction with breasts. Possible bias arises from this loss to follow up but previous research in the field of patient-reported outcomes indicates similar completion rates [17,19,20,35]. The initiation of digitized patient-reported outcome assessments may allow improving engagement with those assessments in future trials [38,39].

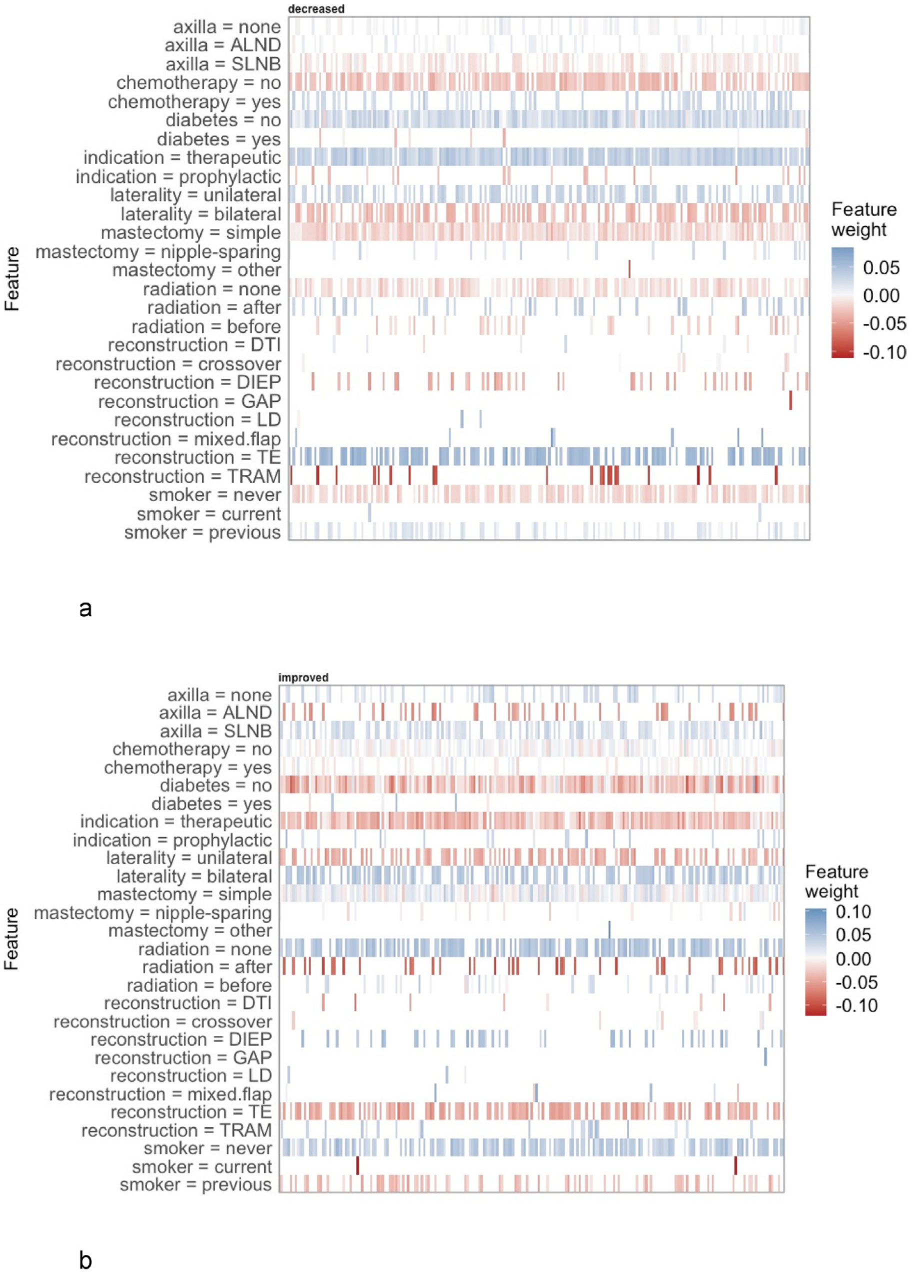

Fig. 6. Local interpretable model-agnostic explanations of the neural network. 6a. Decreased satisfaction.6b. Improved satisfaction.

Third, although satisfaction with reconstructed breasts is a key outcome for women undergoing cancer-related mastectomy and breast reconstruction it is not the only relevant outcome. Future research may look into developing additional prediction models for other patient-reported outcomes like psychological well-being, physical well-being, and sexual well-being after breast reconstruction to inform even more comprehensively about expected quality of life after breast reconstruction.

5. Conclusion

Long-term, individual patient-reported outcomes for women undergoing cancer-related mastectomy and breast reconstruction can be accurately predicted using machine learning algorithms. Our algorithms may be used to better inform clinical treatment decisions for these patients by providing accurate estimates of expected quality of life prior to the initiation of the breast reconstruction process.

Table 4
Multivariable logistic regression for decreased satisfaction with reconstructed breasts.

	Odds ratio (95% CI)	*P* value
Patient variables		
Age	1.01 (0.99–1.02)	0.550
BMI	1.02 (0.99–1.06)	0.202
Diabetes		
No	1 [reference]	–
Yes	1.74 (0.77–3.73)	0.166
Smoker		
Never	1 [reference]	–
Previous	1.41 (1.01–1.98)	**0.045**
Current	2.06 (0.56–6.98)	0.256
Patient-reported outcomes at baseline		
Satisfaction with breasts	1.11 (1.09–1.12)	**<0.001**
Psychosocial well-being	0.99 (0.98–1.00)	**0.033**
Physical well-being chest and upper body	0.99 (0.98–1.00)	0.067
Physical well-being abdomen	1.00 (0.99–1.01)	0.922
Sexual well-being	0.99 (0.98–1.00)	0.097
Clinical variables		
Radiation		
After reconstruction	2.62 (1.70–4.08)	**<0.001**
Before reconstruction	0.95 (0.49–1.78)	0.878
None	1 [reference]	–
Mastectomy		
Nipple-sparing	1.09 (0.65–1.81)	0.745
Simple	1 [reference]	–
other	0.33 (0.01–3.35)	0.416
Reconstruction		
Tissue expander	1 [reference]	–
Direct-to-implant	0.81 (0.40–1.60)	0.548
TRAM	0.23 (0.10–0.49)	**<0.001**
DIEP	0.35 (0.22–0.55)	**<0.001**
LD	0.43 (0.16–1.11)	0.093
GAP	0.40 (0.04–3.03)	0.417
SIEA	0.08 (0.02–0.25)	**<0.001**
Crossover	0.82 (0.38–1.73)	0.614
Mixed flaps	1.07 (0.40–2.60)	0.891
Mixed implants and autologous	0.23 (0.03–1.36)	0.137
Chemotherapy		
Not received	1 [reference]	–
Received	1.31 (0.92–1.87)	0.138
Laterality		
Unilateral reconstruction	1 [reference]	–
Bilateral reconstruction	0.72 (0.51–1.00)	0.052
Mastectomy indication		
Therapeutic	1 [reference]	–
Prophylactic	0.45 (0.21–0.95)	**0.042**
Axillary intervention		
Axillary lymph node dissection	0.69 (0.40–1.20)	0.192
Sentinel lymph node biopsy	0.90 (0.56–1.45)	0.662
No axillary intervention	1 [reference]	–

Role of the funding source

The founding party played no role in analyzing or interpreting the data.

Writing assistance

No writing assistance was provided.

Ethics committee approval

Appropriate institutional review board or research ethics board approval was obtained from all sites. Clinical trial information: NCT01723423.

Declaration of competing interest

André Pfob

No relationship to disclose.

Babak J. Mehrara

No relationship to disclose.

Jonas A. Nelson

No relationship to disclose.

Edwin G. Wilkins

No relationship to disclose.

Andrea L. Pusic

Patents, Royalties, Other Intellectual Property: I am co-developer of BREAST-Q and receive royalty payments when it is used in industry-sponsored trials.

Chris Sidey-Gibbons

No relationship to disclose.

Acknowledgments

Supported by National Cancer Institute Grant No. R01 CA152192 and in part by National Cancer Institute Support Grant No. P30 CA008748 to A.L.P. and E.M.

The BREAST-Q is owned by Memorial Sloan Kettering Cancer Center and the University of British Columbia. A.L.P. is a co-developer of the BREAST-Q and receives royalties when it is used in for-profit industry-sponsored clinical trials.

Authors contributions

Conception and design: Chris Sidey-Gibbons, André Pfob, Andrea L. Pusic, Edwin G. Wilkins, Collection and assembly of data: Chris Sidey-Gibbons, André Pfob, Andrea L. Pusic, Edwin G. Wilkins, Data analysis and interpretation: Chris Sidey-Gibbons, André Pfob, Manuscript writing: All authors, Final approval of manuscript: All authors, Accountable for all aspects of the work: All authors.

Data sharing statement

Will individual participant data be available (including data dictionaries)?.

Yes.

What data in particular will be shared?.

Individual participant data that underlie the results reported in this article, after deidentification (text, tables, figures, and appendices).

What other documents will be available?.

Study Protocol, Statistical Analysis Plan.

When will data be available (start and end dates)?

Immediately following publication. No end date.

With whom?.

Researchers who provide a methodologically sound proposal.

For what types of analyses?.

To achieve aims in the approved proposal.

By what mechanism will data be made available?.

Proposals should be directed to cgibbons@mdanderson.org

To gain access, data requestors will need to sign a data access agreement.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.breast.2021.09.009.

References

[1] Heil J, Kuerer HM, Pfob A, et al. Eliminating the breast cancer surgery paradigm after neoadjuvant systemic therapy: current evidence and future challenges. Ann Oncol 2020;31:61–71.

[2] Kummerow KL, Du L, Penson DF, Shyr Y, Hooks MA. Nationwide trends in mastectomy for early-stage breast cancer. JAMA Surg 2015;150:9–16.

[3] Parker PA, Peterson SK, Bedrosian I, et al. Prospective study of surgical

decision-making processes for contralateral prophylactic mastectomy in women with breast cancer. Ann Surg 2016;263:178–83.

[4] Frost MH, Hoskin TL, Hartmann LC, Degnim AC, Johnson JL, Boughey JC. Contralateral prophylactic mastectomy: long-term consistency of satisfaction and adverse effects and the significance of informed decision-making, quality of life, and personality traits. Ann Surg Oncol 2011;18:3110–6.

[5] Pusic AL, Matros E, Fine N, et al. Patient-reported outcomes 1 year after immediate breast reconstruction: results of the mastectomy reconstruction outcomes Consortium study. J Clin Oncol 2017;35:2499–506.

[6] Ho AY, Hu ZI, Mehrara BJ, Wilkins EG. Radiotherapy in the setting of breast reconstruction: types, techniques, and timing. Lancet Oncol 2017;18: e742–53.

[7] Weber WP, Morrow M, Boniface J de, et al. Knowledge gaps in oncoplastic breast surgery. Lancet Oncol 2020;21:e375–85.

[8] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719–31.

[9] Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019 191 2019;19:1–18.

[10] Alpaydin E. In: Introduction to machine learning. fourth ed. Cambridge, United States: The MIT Press; 2020.

[11] Pfob A, Sidey-Gibbons C, Lee H-B, et al. Identification of breast cancer patients with pathologic complete response in the breast after neoadjuvant systemic treatment by an intelligent vacuum-assisted biopsy. Eur J Canc 2021;143: 134–46.

[12] Sidey-Gibbons C, Pfob A, Asaad M, et al. Development of machine learning algorithms for the prediction of financial toxicity in localized breast cancer following surgical treatment. JCO Clin Canc Inf 2021;5:338–47.

[13] Ong WL, Schouwenburg MG, Van Bommel ACM, 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;3:677–85.

[14] Santosa KB, Qi J, Kim HM, Hamill JB, Wilkins EG, Pusic AL. Long-term patient-reported outcomes in postmastectomy breast reconstruction. JAMA Surg 2018;153:891–9.

[15] Pusic AL, Klassen AF, Scott AM, Klok JA, Cordeiro PG, Cano SJ. Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. Plast Reconstr Surg 2009;124:345–53.

[16] Cano SJ, Klassen AF, Scott AM, Cordeiro PG, Pusic AL. The BREAST-Q: further validation in independent clinical samples. Plast Reconstr Surg 2012;129: 293–302.

[17] Pfob A, Mehrara BJ, Nelson JA, Wilkins EG, Pusic AL, Sidey-Gibbons C. Towards patient-centered decision-making in breast cancer surgery. Ann Surg 2021. https://doi.org/10.1097/SLA.0000000000004862.

[18] Voineskos SH, Klassen AF, Cano SJ, Pusic AL, Gibbons CJ. Giving meaning to differences in BREAST-Q scores: minimal important difference for breast reconstruction patients. Plast Reconstr Surg 2020;145:11e–20e.

[19] Pusic AL, Matros E, Fine N, et al. Patient-reported outcomes 1 Year after immediate breast reconstruction: results of the mastectomy reconstruction outcomes Consortium study. J Clin Oncol 2017;35:2499–506.

[20] Santosa KB, Qi J, Kim HM, Hamill JB, Wilkins EG, Pusic AL. Long-term patient-reported outcomes in postmastectomy breast reconstruction. JAMA Surg 2018;153:891–9.

[21] Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A. The case for process fairness in learning: feature selection for fair decision making. In: 29th conference on neural information processing systems; 2016.

[22] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54(6):1–35. https://doi.org/10.1145/3457607. 115 published online Aug 22.

[23] Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA, J Am Med Assoc 2020;322:1806–16.

[24] Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6: e012799.

[25] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.

[26] Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16:385–95.

[27] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Software 2010;33:1–22.

[28] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232.

[29] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017. published online May 22, http://arxiv.org/abs/1705.07874.

[30] Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?': explaining the predictions of any classifier. published online Feb 16, https://arxiv.org/abs/1602.04938; 2016.

[31] Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. 2016 [published online June].

[32] Kuhn M, Wickham H. RStudio. Package 'recipes'. 2020.

[33] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.

[34] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.

[35] Nelson JA, Allen RJ, Polanco T, et al. Long-term patient-reported outcomes following postmastectomy breast reconstruction: an 8-year examination of 3268 patients. Ann Surg 2019;270:473–83.

[36] Romanoff A, Zabor EC, Stempel M, Sacchini V, Pusic A, Morrow M. A comparison of patient-reported outcomes after nipple-sparing mastectomy and conventional mastectomy with reconstruction. Ann Surg Oncol 2018;25: 2909–16.

[37] Pfob A, Sidey-Gibbons C, Heil J. Response prediction to neoadjuvant systemic treatment in breast cancer – yet another algorithm? JCO Clin Cancer Inform 2021;5:654–5. https://doi.org/10.1200/CCI.21.00033.

[38] Geerards D, Pusic A, Hoogbergen M, Van Der Hulst R, Sidey-Gibbons C. Computerized quality of life assessment: a randomized experiment to determine the impact of individualized feedback on assessment experience. J Med Internet Res 2019;21. https://doi.org/10.2196/12212.

[39] Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016;18. https://doi.org/10.2196/jmir.4790.