ELSEVIER

Journal of Clinical Epidemiology 152 (2022) 238−247

Journal of Clinical Epidemiology

REVIEW ARTICLE

The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias

Tom A. Hueting[a], Marissa C. van Maaren[a,b], Mathijs P. Hendriks[a,b,c], Hendrik Koffijberg[a], Sabine Siesling[a,b,*]

[a]*Department of Health Technology & Services Research, Technical Medical Centre, University of Twente, Enschede, The Netherlands*
[b]*Department of Research and Development, Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands*
[c]*Department of Medical Oncology, Northwest Clinics, Alkmaar, The Netherlands*

Accepted 20 October 2022; Published online 27 October 2022

Abstract

Objectives: To systematically review the currently available prediction models that may support treatment decision-making in breast cancer.

Study Design and Setting: Literature was systematically searched to identify studies reporting on development of prediction models aiming to support breast cancer treatment decision-making, published between January 2010 and December 2020. Quality and risk of bias were assessed using the Prediction model Risk Of Bias (ROB) Assessment Tool (PROBAST).

Results: After screening 20,460 studies, 534 studies were included, reporting on 922 models. The 922 models predicted: mortality (n = 417 45%), recurrence (n = 217, 24%), lymph node involvement (n = 141, 15%), adverse events (n = 58, 6%), treatment response (n = 56, 6%), or other outcomes (n = 33, 4%). In total, 285 models (31%) lacked a complete description of the final model and could not be applied to new patients. Most models (n = 878, 95%) were considered to contain high ROB.

Conclusion: A substantial overlap in predictor variables and outcomes between the models was observed. Most models were not reported according to established reporting guidelines or showed methodological flaws during the development and/or validation of the model. Further development of prediction models with thorough quality and validity assessment is an essential first step for future clinical application. © 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Clinical prediction models; Treatment decision support; Breast cancer; Systematic review; Risk of bias; Nomograms; Prognostic models

1. Introduction

Breast cancer is the most commonly diagnosed cancer in women worldwide. Disease severity and treatment options for breast cancer depend on various factors such as subtype, tumor stage, personal context, and genetic characteristics [1]. The heterogeneity of breast cancer challenges clinicians to optimize treatment for each individual patient. Pros and cons of different treatment options (i.e., improvement of prognosis vs. (late) adverse events) should be considered before treatment initiation on an individual patient level. Clinical prediction models can support clinical decision-making by estimating individual predictions on certain outcomes using combinations of different relevant patient and disease characteristics.

Multiple prediction models have been available to guide treatment decision-making for breast cancer patients in the past years. For example, Predict [2] is a prediction model

Compliance with ethical standards: Tom Hueting declares employment at Evidencio B.V, the other authors have no potential conflicts of interest to disclose.

Author contributions: Conceptualization: TH, MvM, HK, SS. Data curation: TH, MvM. Formal analysis: TH, MvM. Methodology: TH, MvM, MH, HK, SS. Resources: TH, MvM, SS. Supervision: HK, SS. Validation: MvM, SS. Visualization: TH, MvM. Roles/Writing - original draft: TH, MvM, MH, HK, SS. Writing - review and editing: TH, MvM, MH, HK, SS.

Declaration of interests: 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.

* Corresponding author: Netherlands Comprehensive Cancer Organisation (IKNL), POBox 19079, 3501 DB UTRECHT, The Netherlands; And Technical Medical Centre, University of Twente, POBox 217, 7500 AE Enschede, The Netherlands. Tel.: +31 53 489 7475.

E-mail address: s.siesling@iknl.nl (S. Siesling).

https://doi.org/10.1016/j.jclinepi.2022.10.016

What is new?

Key findings

- An abundance of clinical prediction models are available to support treatment decisions for breast cancer patients.
- The majority of clinical prediction models are poorly reported, show methodological flaws, or are at high risk of bias.

What this adds to what is known?

- This review systematically identified and critically appraised clinical prediction models that were developed to support treatment decisions in breast cancer patients.

What is the implication/what should change now?

- Development of new clinical prediction models should adhere to established methodological guidelines and need to be reported completely and transparently
- Existing models require thorough quality and validity assessment prior to their use in clinical practice.

that has been available as an online model to support decision-making on adjuvant treatment strategies. The use of Predict or other similar tools such as CancerMath [3] or the Nottingham Prognostic Index [4] has been recommended in international guidelines [5]. Yet there are more treatment decisions for breast cancer patients that could be well supported by prediction models. There may be potentially valuable models already available that are not currently used because their quality and reliability are unclear.

Before prediction models may be implemented in clinical practice, multiple steps should be performed. These methods include the steps for development, internal validation, external validation, updating, and impact assessment of prediction models [6–9]. Ideally, the development and validation of a model should be described according to the guideline for transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [10].

Previous systematic reviews have been conducted to assess clinical prediction models for breast cancer, but seem to have identified older and potentially outdated models and only a limited number of recently developed prediction models, possibly by only including models that predict a specific outcome, such as mortality or recurrence [11]. However, with regard to the application of prediction models aimed at supporting decision-making in breast cancer care, it is currently unknown how many different models have been developed, which outcomes can be (accurately) predicted and with which variables the outcomes can be predicted. We therefore aimed to systematically review prediction models that may be used to support treatment decision-making in breast cancer patients and to assess the quality of studies reporting on the development and (internal) validation of prediction models.

2. Methods

The systematic review study protocol has been registered in the International Prospective Register of Systematic Reviews (registration number: CRD42020134826). The Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist for transparent reporting of systematic reviews and meta-analysis was followed for reporting the results (Supplementary data S2) [12].

2.1. Search strategy

Medline (Pubmed) and Embase (Elsevier) were searched for studies published between January 1, 2010, and December 31, 2020. The search strategy was constructed using validated search filters to find prognostic prediction studies (Supplementary data S1) [13].

2.2. Study selection

Studies were included if they reported the development of a prediction model intended to be used for treatment decision-making in patients (both men and women) who have been diagnosed with breast cancer. Such outcomes include survival (overall or disease-specific), recurrence ([loco]-regional or second primary breast cancer), metastasis (including contralateral lymph nodes), adverse events, quality of life, and treatment response. Included studies must be aimed at providing predictions for breast cancer patients using a combination of two or more predictor variables, possibly demonstrated by providing a calculation method (i.e., logistic regression, neural network). Studies can report the development of multiple prediction models. We defined separate models when either the predictor-outcome association was different, or when the predictor-outcome association was the same, but a different baseline hazard or intercept was reported. Conference abstracts were excluded from the review as they are unable to provide sufficient details regarding the development and initial validation of the model in order to assess its quality. Two types of prediction models were distinguished, diagnostic and prognostic models. Diagnostic models aim to estimate the likelihood for currently having the outcome, whereas prognostic models aim to estimate the probability of the outcome at a specified future time. Subsequently, separate studies were searched for that described the external validation of one or more of the included models. The external validation studies were identified by searching via Google Scholar for studies citing the study describing the development of the model. The process of including and excluding studies for the review was performed by one reviewer (T.A.H.). To ensure that the study selection process was performed appropriately by one reviewer, a second reviewer (M.C.v.M.) performed the same process for a random sample of 1,600 studies. Discrepancies were resolved after discussions between the two reviewers.

2.3. Data extraction

Data extraction was performed using the checklist for critical appraisal and data extraction for systematic reviews of prediction modeling studies (CHARMS) [14].

The definition of the same predictors sometimes varied between different studies. For instance, HER2-status could be defined as negative or positive, or could be incorporated in a subtype variable including both HER2 and hormonal receptor status. We decided to report the definition of the predictor as reported in the study.

To assess whether the sample size was likely sufficient to develop the model, the events per variable (EPV) were estimated for each model. The EPV is a traditional criterion that is used to estimate how many predictors could be included in the multivariable analysis. Even though the EPV has its limitations, it provides a rough indication of whether the sample size was sufficient [15]. The sample size is likely to be insufficient with an EPV <10. An EPV between 10 and 20 could be sufficient, but is still fairly low, and an EPV >20 is likely to be sufficient.

2.4. Risk of bias (ROB)

To assess whether reviewed studies are at low or high ROB, the Prediction model Risk Of Bias Assessment Tool (PROBAST) was used [16]. The PROBAST tool includes 20 signaling questions in four domains: participants, predictors, outcome, and analysis. In addition, an overall conclusion regarding low or high ROB for the reviewed prediction models was determined. The participants domain covered the ROB related to study data and the methods used to enroll study participants. The predictors domain covered ROB caused by the measurement and definition of predictors. The outcome domain assessed the ROB caused by the estimation and definition of the outcome. The analysis domain covered the ROB related to the statistical methods used to develop and validate the model (Box 1).

Data extraction and the ROB assessment were performed for all prediction models by one reviewer (T.A.H.). For a

Box 1 The Prediction model Risk Of Bias Assessment Tool (PROBAST).

The PROBAST was developed to critically appraise the development and validation of prediction models. Even though the PROBAST can be used to assess both risk of bias (ROB) and concerns regarding applicability, it was mainly used in this review to assess the ROB. The PROBAST aims to judge the ROB in four domains. Each domain has a set of signaling questions that needs to be answered with either “(Probably) Yes,” “(Probably) No,” or “No information.” The ROB is subsequently judged as low, high, or unclear. The following domains are identified by the PROBAST:

1. Participants

The first domain has two signaling questions regarding the appropriateness of used data sources and the applied inclusion and exclusion criteria.

2. Predictors

This domain includes three signaling questions regarding uniformly described predictors, predictor assessment, and availability of predictors at the time the model is intended to be used.

3. Outcome

The outcome domain has six signaling questions regarding its determination, definition, and time interval between predictor measurement and outcome occurrence.

4. Analysis

The analysis domain has nine signaling questions regarding the statistical methods used to develop and validate the model. Topics include the sample size, handling of continuous predictors, inclusion of patients in the analysis, dealing with missing data, avoidance of univariable analysis, dealing with complexities in the data, appropriateness of performance measures, dealing with overfitting, underfitting, and optimism in the model, and whether the weights in the final model correspond with the results from the analysis.

All signaling questions should be answered with “(Probably) Yes” for a low ROB rating. At least one “(Probably) No” results in a high ROB, and at least one “No information” (and no “(Probably) No” results in an unclear ROB rating.

subset of 20 models, these processes were also performed by a second reviewer (M.C.v.M.) to identify potential discrepancies and to verify the quality of the review activities. Based on the similarities in ROB assessment between the two reviewers, the subset of 20 models assessed by the second reviewer was deemed sufficiently large to ensure high quality data extraction and ROB assessment.

3. Results

The search strategy identified 20,460 studies, of which the titles were screened. The abstract was screened for studies that could not be excluded based on the title alone. Subsequently, 1,345 studies were selected for full-text screening. Finally, a total of 534 studies were included, reporting on 922 models. The inclusion and exclusion criteria of the different studies and the reasons for excluding studies are shown in Figure 1.

3.1. Predictors

A total of 228 different model predictors were identified in the included 922 models. A total of 14 predictors were used in more than 100 different models: age (*n* = 426, 48%), tumor size (*n* = 373, 40%), lymph node involvement (*n* = 337, 37%), tumor grade (*n* = 297, 32%), ER-status (*n* = 187, 20%), HER2-status (*n* = 158, 17%), surgery (*n* = 149, 16%), radiotherapy (*n* = 141, 15%), chemotherapy (*n* = 141, 15%), subtype (*n* = 132, 14%), PR-status (*n* = 130, 14%), metastasis (*n* = 123, 13%), and genetic risk score (*n* = 115, 12%). In the supplementary materials (S4, sheet "predictors"), an overview of all predictors per outcome is displayed. The five most common predictors per outcome are shown in Table 1.

3.2. Outcome

The included studies described models that were developed to predict the following outcomes: mortality (*n* = 417, 45%), recurrence(-free survival) (*n* = 217, 24%), lymph node involvement (*n* = 141, 15%), adverse events (*n* = 58, 6%), treatment response (*n* = 56, 6%), and other outcomes (*n* = 33, 4%) such as menopausal status, quality of life, surgical margin, receiving treatment, cosmetic outcome, nipple-areola complex invasion. The number of models per outcome is displayed in Table 1.

The majority of the models predicted similar outcomes, although the models often differed in the specific definition of the outcome (i.e., lymph node involvement could include both sentinel and non-sentinel lymph node involvement), or the models used different inclusion and exclusion criteria to develop the model. Out of the 922 models, 693 (75%) were prognostic, and 229 (25%) were diagnostic models. The details of all included models were added as an additional spreadsheet in supplementary material S4.

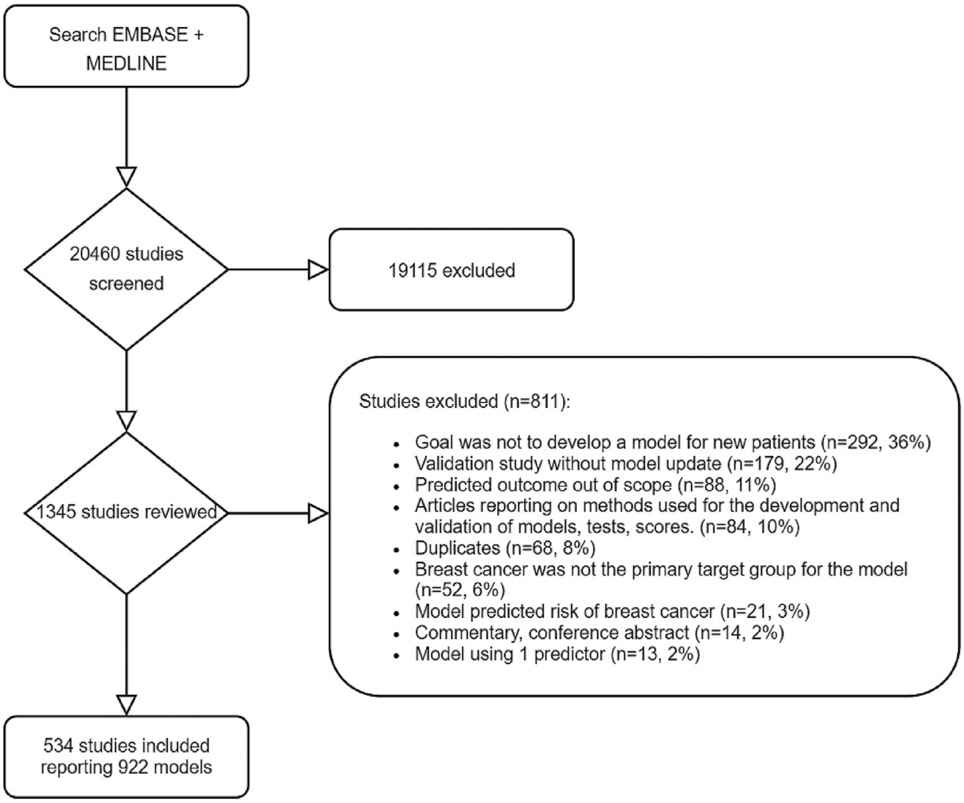
Search EMBASE + MEDLINE
20460 studies screened
19115 excluded
1345 studies reviewed
Studies excluded (n=811):
• Goal was not to develop a model for new patients (n=292, 36%)
• Validation study without model update (n=179, 22%)
• Predicted outcome out of scope (n=88, 11%)
• Articles reporting on methods used for the development and validation of models, tests, scores. (n=84, 10%)
• Duplicates (n=68, 8%)
• Breast cancer was not the primary target group for the model (n=52, 6%)
• Model predicted risk of breast cancer (n=21, 3%)
• Commentary, conference abstract (n=14, 2%)
• Model using 1 predictor (n=13, 2%)
534 studies included reporting 922 models

Fig. 1. Flowchart of study inclusion and exclusion criteria.

Table 1. Overview of models by outcome

Outcome	Specified outcome	Models (n = 922)	Average C-index	Most common predictors (n (X%))
Mortality	Overall survival	316	0.740	Age (184 (58%)), Tumor size (146 (46%)), Lymph node involvement (126 (40%)), Tumor grade (126 (40%)), Metastasis (82 (26%))
	Disease specific survival	94	0.763	Tumor grade (64 (68%)), Tumor size (62 (66%)), Age (54 (57%), Lymph node involvement (50 (53%)), ER status (39 (41%))
	Other cause specific survival	7	0.746	Age (7 (100%)), Tumor size (7 (100%), Surgery (4 (57%)), Chemotherapy (3 (43%)), Marital status (3 (43%))
Recurrence	Recurrence (free survival)	132	0.769	Lymph node involvement (54 (41%)), Age (42 (32%)), Tumor size (38 (29%)), Grade (35 (27%)), Genetic risk score (27 (21%))
	Locoregional recurrence	33	0.728	Age (27 (82%)), Tumor grade (24 (73%)), Tumor size (19 (58%)), Lymph node involvement (18 (55%)), Hormonal therapy (17 (52%))
	Metastasis	45	0.760	Lymph node involvement (30 (67%)), Tumor size (21 (47%)), Age (16 (36%)), Genetic risk score (13 (29%)), Subtype (7 (16%)), Lymphovascular invasion (7 (16%))
	Contralateral recurrence	7	0.589	Age (6 (86%)), Histology (5 (71%)), Radiotherapy (5 (71%)), Tumor size (4 (57%)), ER status (3 (43%)), Grade (3 (43%)), Surgery (3 (43%)), Hormonal therapy (3 (43%)), Family History (3 (43%))
Lymph nodes	Lymph node involvement	83	0.791	Tumor Size (31 (37%)), Age (27 (33%)), Lymph node status (21 (25%)), Grade (21 (25%)), Lymphovascular invasion (19 (23%))
	Sentinel lymph node involvement	13	0.763	Age (5 (38%)), Lymphovascular invasion (5 (38%)), Tumor size (3 (23%)), ER status (3 (23%)), PR status (3 (23%)), HER2 status (3 (23%)), Tumor location (3 (23%)), Multifocality (3 (23%))
	Non-sentinel lymph node involvement	45	0.758	Lymph node involvement (21 (47%)), Lymphovascular invasion (20 (44%)), Diameter largest lymph node (17 (38%)), Tumor size (13 (29%)), Lymph node ratio (11 (24%))
Treatment response	Pathologic complete response	56	0.812	ER status (21 (38%)), HER2 status (20 (36%)), Tumor size (16 (29%)), KI67 (15 (27%)), Age (9 (16%)), PR status (9 (16%)), Grade (9 (16%))
Adverse events	Lymphedema	15	0.775	BMI (10 (67%)), Radiotherapy (10 (67%)), chemotherapy (10 (67%)), Surgery (6 (40%)), Age (5 (33%)), Lymph nodes dissected (5 (33%)), Lymph node surgery (5 (33%))
	Cardiovascular complications	9	0.780	Age (8 (89%)), Chemotherapy (3 (33%)), BMI (2 (22%)), Tumor size (1 (11%)), Surgery (1 (11%))
	Pain	7	0.703	Preoperative pain (5 (71%)), BMI (4 (57%)), Age (3 (43%)), Lymph nodes dissected (3 (43%)), Postoperative pain (2 (29%))
	Other adverse events[a]	27	0.711	Age (10 (36%)), BMI (9 (32%)), Smoking status (8 (29%)), Comorbidities (8 (29%)), Radiotherapy (7 (25%))
Other outcomes	Menopausal status	8	0.814	Age (8 (100%)), BMI (4 (50%)), Chemotherapy (3 (38%)), Hormonal therapy (3 (38%)), FSH (3 (38%))
	Quality of life	8	Not applicable	Age (6 (75%)), Chemotherapy (5 (63%)), Hormonal therapy (5 (63%)), Radiation therapy (5 (63%)), Stage (5 (63%)), Surgery (5 (63%)), Complications (5 (63%)), menopausal status (5 (63%)), Ambulatory (5 (63%)), Charlson Deyo score (5 (63%)), Education (5 (63%)), Postoperative length of stay (5 (63%))
	Surgical margin	6	0.726	Tumor grade (2 (33%)), lymph node involvement (2 (33%)), ER-status (2 (33%)), Tumor size (2 (33%)), Her2-status (2 (33%)), PR-status (2 (33%)), Metastasis (2 (33%)), Histology (2 (33%)), Multifocality (2 (33%))
	Treatment as the outcome	7	0.705	Age (4 (57%)), Tumor size (3 (43%)), ER status (2 (29%)), Race (2 (29%)), Radiotherapy (2 (29%))
	Good cosmetic	2	0.790	Lymphovascular invasion (2), Multifocality (1 (50%)), total tumor

(*Continued*)

Table 1. Continued

Outcome	Specified outcome	Models (*n* = 922)	Average C-index	Most common predictors (*n* (X%))
	outcome			load (1 (50%)), Tumor size (1 (50%)), ER status (1 (50%)), Lymph node involvement (1 (50%)), Grade (1 (50%))
	Nipple-areola complex invasion	2	0.879	Distance from nipple (2 (100%)), Lymph node involvement (1 (50%)), Tumor size (1 (50%)), Location (1 (50%)), Imaging outcome (1 (50%)), Multicentricity (1 (50%))

The details of all included models were added as an additional spreadsheet in supplementary material S4.

[a] Other adverse events include pneumonitis, necrosis, seroma, infection, exposure, explantation, overall complication, falls, symptomatic skeletal events, fibrosis, rash, fatigue, neutropenia, and cognitive impairment.

3.3. Modeling methods

Relevant findings related to methods used to develop and validate the prediction models were rated (Table 2). To develop diagnostic models, logistic regression was mostly used (n = 197 [86%]). For prognostic models, Cox regression was used in 510 (74%) of the models. The majority of models were developed using data from patients in Asian (n = 319, 35%), North-American (n = 262, 28%), or European (n = 183, 20%) countries. A total of 429 (47%) models were developed with patient data from multiple centers, and 386 (42%) models were developed with data from a single institution.

The median number of participants used to develop a model was 699 (IQR 272– 2970), with a median number of events of 130 (IQR 58–416). Regarding the sample sizes used to develop the models, the EPV could not be determined for 269 (29%) models, 162 (18%) models were developed with an EPV <10, and 159 (17%) models with an EPV between 10 and 20. The remaining 332 (36%) models were developed with an EPV ≥20.

For 525 (57%) of the developed models, it was unclear how the developers dealt with missing data in the derivation dataset, 297 (32%) of the models were developed using complete-case analysis, and only 80 (9%) of the models were developed using an imputation (i.e., multiple or single) method to deal with missing data as recommended by the TRIPOD statement [10]. A total of 285 (31%) models were not reported with sufficient information to apply the model in practice. This was mostly caused by the absence of either the predictor coefficients (n = 119, 13%), the baseline hazard (n = 96, 10%), or the intercept (n = 51, 6%).

3.4. Risk of bias

The models were rated as either low (n = 27, 3%), high (n = 878, 95%), or unclear (n = 17, 2%) ROB. The majority of the models were considered at high ROB, mainly due to the assessment of the domain ‘analysis’ in the PROBAST tool. Figure 2 shows the general assessment of the ROB and supplementary table 1 displays the ROB assessment per model. Discrepancies in ROB assessments performed by the two reviewers were sometimes found between answers of signaling questions, but the assessment for each PROBAST domain was similar for all studies that were assessed by both reviewers. The studies with a low ROB were added in supplementary table S3.

Reasons for defining PROBAST domains to be unclear or high risk were often similar for different models. Figure 3 represents the ROB stratified per outcome. An unclear or high ROB in the outcome domain occurred more often in models predicting recurrence or adverse events compared to the other models. Only 50% and 47% of the models predicting recurrence and adverse events were deemed at low ROB in the outcome domain where this percentage was 93%, 87%, 96%, and 73% for mortality, lymph node involvement, response, and other outcomes, respectively. The reason for this difference is mostly due to differences in methods to define the outcome (i.e., assessment via telephone follow-up) or to lack of description on the intensity and method of the follow-up. Notably, the ROB for the ‘analysis’ domain was defined as high for the majority of the models (95%). Common reasons for the high ROB concerned inadequate dealing with missing data, using univariable analysis to select candidate predictors, or not dealing with overfitting or optimism in the model. Out of all the models that were high ROB in the analysis domain, 82% showed concerns on two or more of the signaling questions, whereas the PROBAST tool advises to assign high ROB already if one of the signaling questions is not appropriately addressed.

3.5. Model performance

The most commonly applied measure to assess the performance of the model concerned model discrimination. Discrimination was quantified using the C-index in the development of 814 (88%) of the models. The C-index could vary widely based on the outcome predicted by the model, the predictors incorporated in the final model, and the methods used to develop and validate the model. The average C-index per predicted outcome is shown in Table 1. The C-index was used in 96 (72%) of the external validation studies. Finally, there were 72 models for which a C-index was assessed at both model development and external validation. On average, the C-index at external validation (0.71) was lower than during model development (0.77). Only

Table 2. Summary of extracted items for all included models.

Item	Diagnostic models (*N* = 229)	Prognostic models (*N* = 693)	Total included models (*N* = 922)
Modeling method			
Cox regression	0 (0%)	510 (74%)	510 (55%)
Fine and Gray model	0 (0%)	25 (4%)	25 (3%)
Logistic regression	197 (86%)	93 (13%)	290 (31%)
Linear regression	2 (1%)	9 (1%)	11 (1%)
Machine learning	25 (11%)	41 (6%)	66 (7%)
Other[a]	4 (2%)	13 (2%)	17 (2%)
Unclear	1 (0.4%)	2 (0.3%)	3 (0.3%)
Location of participants used to develop the model			
Asian	121 (53%)	199 (29%)	319 (35%)
North-American	35 (15%)	227 (33%)	262 (28%)
European	61 (27%)	121 (17%)	183 (20%)
South-American	1 (0.4%)	4 (1%)	5 (1%)
African	1 (0.4%)	1 (0.1%)	2 (0.2%)
Oceania	0 (0%)	3 (0.4%)	3 (0.3%)
Multiple continents	4 (2%)	16 (2%)	20 (2%)
Unclear	6 (3%)	122 (18%)	128 (14%)
Database used to develop the model			
Single center	149 (65%)	237 (34%)	386 (42%)
Multicenter	52 (23%)	105 (15%)	157 (17%)
Registry	23 (10%)	249 (36%)	272 (30%)
Unclear	5 (2%)	102 (15%)	107 (12%)
Participants in derivation cohort (*n*)			
<100	24 (10%)	16 (2%)	40 (4%)
100 − 200	50 (22%)	63 (9%)	113 (12%)
200 − 500	67 (29%)	143 (21%)	210 (23%)
500 − 1,000	39 (17%)	130 (19%)	169 (18%)
1,000 − 10,000	40 (17%)	196 (28%)	236 (26%)
≥10,000	9 (4%)	119 (17%)	128 (14%)
Unclear	0 (0%)	26 (4%)	26 (3%)
Events per variable			
<10	55 (24%)	107 (15%)	162 (18%)
10 − 20	45 (20%)	114 (16%)	159 (17%)
20 − 50	62 (27%)	84 (12%)	146 (16%)
≥50	40 (17%)	146 (21%)	186 (20%)
Unclear	27 (12%)	242 (35%)	269 (29%)
Dealing with missing data			
Excluded patients with missing data	61 (27%)	236 (34%)	297 (32%)
Imputation (multiple, random, mean, single)	9 (4%)	71 (10%)	80 (9%)
Unknown modeled as covariate	4 (2%)	8 (1%)	12 (1%)
No missing data	1 (0%)	7 (1%)	8 (1%)
Unclear	154 (67%)	371 (54%)	525 (57%)
Model performance (discrimination)			
Quantified	215 (94%)	599 (86%)	814 (88%)
Not quantified	14 (6%)	94 (14%)	108 (12%)
Model performance (calibration)			
Plot (observed vs. expected)	89 (39%)	419 (60%)	508 (55%)
Hosmer−Lemeshow goodness of fit test	11 (5%)	22 (3%)	33 (4%)
Other[b]	3 (1%)	44 (6%)	47 (5%)

(*Continued*)

Table 2. Continued

Item	Diagnostic models (N = 229)	Prognostic models (N = 693)	Total included models (N = 922)
Unclear	126 (55%)	208 (30%)	334 (36%)
Validation method			
Apparent	41 (18%)	73 (11%)	114 (12%)
x-fold cross validation	20 (9%)	27 (4%)	47 (5%)
Bootstrap	30 (13%)	108 (16%)	138 (15%)
External validation cohort	29 (13%)	147 (21%)	176 (19%)
Temporal validation cohort	12 (5%)	29 (4%)	41 (4%)
Split sample	61 (27%)	171 (25%)	232 (25%)
Combination of multiple methods	22 (10%)	85 (12%)	107 (12%)
Unclear	14 (6%)	53 (8%)	67 (7%)
Model is reproducible			
No	79 (34%)	206 (30%)	285 (31%)
Yes	150 (66%)	487 (70%)	637 (69%)

The details of all included models were added as an additional spreadsheet in supplementary material S4. Percentages added together may not be equal to 100% due to rounding

[a] Other modeling methods include classification and regression trees (CART), parametric survival regression, principal component analysis, and structural equation modeling.

[b] Other calibration methods include the use of a table, description of observed vs. expected, or a bar chart.

a minority of models were externally validated. At the time of development, 176 (19%) of the models were validated using an external validation cohort. Subsequently, 82 (9%) of the models were externally validated in a separate study. Where 41 (50%) models were externally validated in multiple studies. The identified external validation studies were added to the supplementary data table.

4. Discussion

This systematic review identified a total of 534 studies published between 2010 and 2020, reporting the development of 922 different models. The patient's age, tumor size, and lymph node involvement were the most common predictors and were used in more than a third of the models. Models were categorized as either predicting a prognostic (n = 693, 75%) or a diagnostic (n = 229, 25%) outcome, The quality of the identified models was poor as only 35 models (4%) were developed with appropriate statistical methods according to the PROBAST tool, and only 27 models (3%) were deemed at low ROB overall.

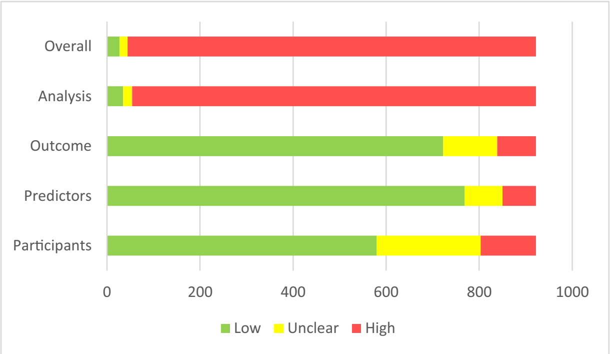

Fig. 2. Risk of bias by PROBAST domains. A rating of high was given for a subdomain when at least one signaling question was answered with a "No." A low risk of bias rating was given if all signaling questions were answered with "Yes." An unclear risk of bias is assigned if at least one signaling question could not be answered, and if the remaining signaling questions were answered with "yes."

Predictors used in the identified models were overlapping to a large extent. This makes sense as these predictors were proven to provide significant prognostic information regarding relevant health outcomes. ER status is an example of a predictor that was often used to predict different outcomes. ER status was mostly entered in the model as a dichotomous variable (i.e., negative, or positive). Even though the registration of such predictors as dichotomous variables is commonly applied and accepted, the dichotomization of continuous variables is regarded as bad practice [17]. As multiple predictors are commonly accepted as dichotomous variables in clinical practice, the use of these variables was not a reason for a high ROB rating as suggested in the PROBAST tool. Accepted dichotomous variables were ER status, PR status, HER2-status, KI67 status, and tumor stage. Even though the use of the EPV criterion is regarded as suboptimal [15], the EPV could not be determined for 269 models (29%) mainly due to the lack of reporting on the number of events.

This review identified a disproportionate number of models predicting the same outcome. The majority of identified prediction models in breast cancer were developed using suboptimal or inappropriate methodology. This result aligns with previous findings in other disease areas. The review by Damen et al. assessed 363 prediction models for cardiovascular disease and concluded that the most models

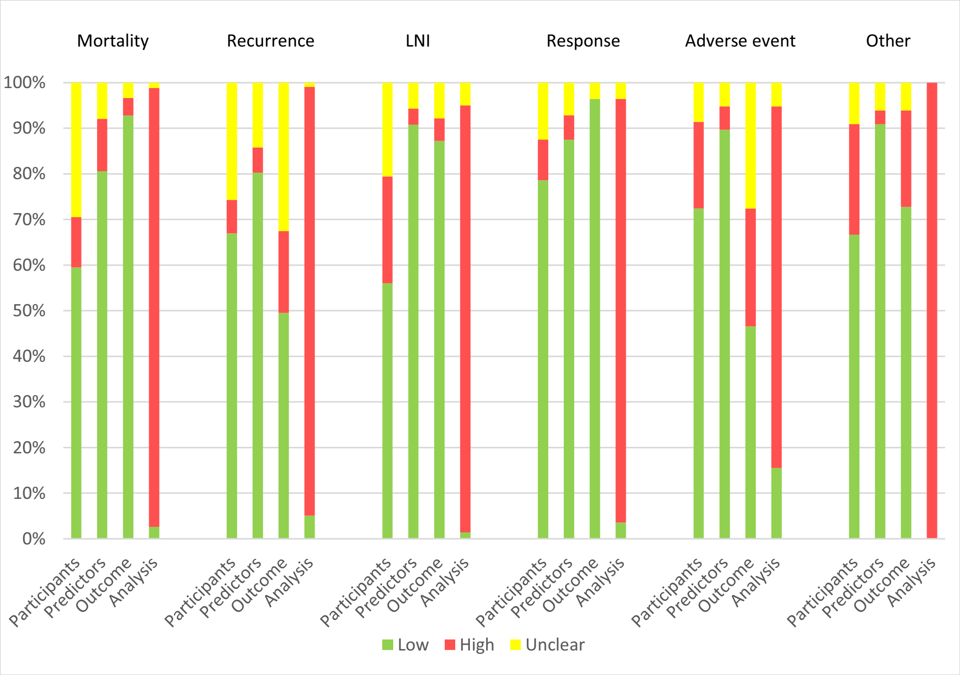

Fig. 3. Risk of bias assessment per outcome. LNI, lymph node involvement. A rating of high was given for a subdomain when at least one signaling question was answered with a ''No.'' A low risk of bias rating was given if all signaling questions were answered with ''Yes.'' An unclear risk of bias is assigned if at least one signaling question could not be answered, and if the remaining signaling questions were answered with ''yes.''

were reported inadequately according to the CHARMS checklist [18]. A more recent systematic review of predictions models for diagnosis and prognosis of COVID-19 found all models to be at high ROB [19]. A systematic review by Phung et al. on prognostic models for breast cancer focused more on their performance [11]. Even though only 58 models were identified, the authors concluded that the performance for most models was suboptimal in independent cohorts, which could have been expected as the methods for development of most models were also suboptimal. Clearly, the lack of adequate reporting and the methodological shortcomings for the development of prediction models is not specific for breast cancer, but seems to be a common problem within clinical research. Adherence to reporting guidelines such as the TRIPOD is necessary to improve the quality of developed prediction models. A limitation of this review concerns the fact that a subset of 1,600 of the identified studies in the search strategy was assessed by a second reviewer. From this subset of 1,600 studies, only three studies were additionally included for full-text analysis. For this reason, we do not expect that the review protocol led to exclusion of relevant studies. In addition, referenced models in studies reporting on the validation of prediction models, as well as references in previous systematic reviews on breast cancer prediction models were assessed to minimize the risk of missing relevant studies. The same limitation is applicable to the data extraction and ROB assessment. In our study, a second reviewer assessed 20 prediction models to assess the data extraction process and the ROB assessment (using the PROBAST). No model would have been rated differently even though some differences were found on the signaling questions. Despite the fact that performing data extraction in duplicate would ensure fewer errors, it is unlikely that this would have changed the conclusions of the review [20]. Potentially relevant studies describing the development or update of a model might have been missed during the review due to the exclusion of studies reported in languages other than English or Dutch. Besides, the most recently developed models that have been published since January 1, 2021 were not included in the review.

One of the most important findings in this review concerns the high proportion of models regarded at high ROB. The majority of the models were at high ROB due to the 'analysis' domain, in which the ROB due to statistical methods is assessed. Still, a high ROB rating does not necessarily mean that the model has no or limited clinical value and a low ROB rating does not automatically constitute a valuable model. For instance, studies reporting on the update of the Predict model were rated as low ROB (Supplementary data S3), but an external validation study demonstrated suboptimal performance of the model in different patient groups [21,22]. Besides, each model only predicts a single outcome, whereas clinical decision-making also requires individual estimates of other relevant outcomes such as adverse events. Before clinical use of a model can be justified, different steps have to be taken for the development, internal and external validation, update, and impact assessment. Even then, the model needs to be trusted and understood by clinicians or adopted in clinical guidelines, and both the preferences and context of the patient should be taken into account before widespread implementation of a model is accepted in daily clinical practice. Nevertheless, the development of a valuable model starts with a good performance on internal validation, carried out with the appropriate statistical methods. Further (external) validation of the models may ultimately conclude whether the models may be generalized to different patient cohorts and perhaps different health care settings [23]. Even when models were proven to perform sufficiently well in external populations, additional (clinical) evaluations should be performed to assess the clinical and health impact of a prediction model [24]. Besides, with changing regulations in the European Union, the majority of prediction models in the current review are very likely to require certification as a medical device according to the Medical Devices Regulation before clinical use is enabled [25]. The fact that such a low number of models ($n = 27$, 3%) were considered to be reported adequately based upon the model development stage underpin the need for improved reporting of prediction model development, perhaps now more than ever.

5. Conclusion

Many prediction models have been published during the past decade to predict outcomes related to breast cancer treatment. Nearly all published prediction models identified were deemed as high ROB. Mainly due to a lack of adequate reporting, many prediction models could not be implemented in clinical practice as the studies did not provide sufficient data for external validation studies or an impact assessment. Future studies should focus on improving currently available models, either by identifying specific subgroups for which no model is applicable, or by performing the required steps before clinical adoption can be justified (i.e., external validation and impact assessment) rather than developing more new models.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 394 CA: a cancer journal for clinicians global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.

[2] 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:R1.

[3] Michaelson JS, Chen LL, Bush D, Fong A, Smith B, Younger J. Improved web-based calculators for predicting breast carcinoma outcomes. Breast Cancer Res Treat 2011;128:827–35.

[4] Blamey RW, Pinder SE, Ball GR, Ellis IO, Elston CW, Mitchell MJ, et al. Reading the prognosis of the individual with breast cancer. Eur J Cancer 2007;43:1545–7.

[5] Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2019;30:1194–220.

[6] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. Plos Med 2013;10(2):e1001381.

[7] Steyerberg EW. A Practical Approach to Development, Validation, and Updating. Clinical prediction models. Springer; 2019. https://doi.org/10.1007/978-3-030-16399-0.

[8] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683–90.

[9] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691–8.

[10] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594.

[11] Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: A systematic review. BMC Cancer. London, UK: BioMed Central Ltd; 2019.

[12] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097.

[13] Geersing G-J, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG, et al. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One 2012;7(7):e32844.

[14] Moons KGMM, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11(10):e1001744.

[15] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans S, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2019;28:2455–74.

[16] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51–8.

[17] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41.

[18] Damen JAAG, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416.

[19] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.

[20] Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol 2006;59:697–703.

[21] Engelhardt EG, van den Broek AJ, Linn SC, Wishart GC, Rutgers EJT, van de Velde AO, et al. Accuracy of the online prognostication tools PREDICT and Adjuvant! for early-stage breast cancer patients younger than 50 years. Eur J Cancer 2017;78:37–44.

[22] van Maaren MCC, 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.

[23] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31.

[24] van Giessen A, Peters J, Wilcher B, Hyde C, Moons C, de Wit A, et al. Systematic review of health economic impact evaluations of risk prediction models: stop developing, start evaluating. Value Heal 2017;20:718–26.

[25] Medical devices regulation. 2017. Available at: https://eur-lex.europa.eu/eli/reg/2017/745/2017-05-05. Accessed February 8, 2021.