Journal of Clinical Epidemiology 159 (2023) 159−173

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Systematic metareview of prediction studies demonstrates stable trends in bias and low PROBAST inter-rater agreement

Liselotte F.S. Langenhuijsen[a], Roemer J. Janse[a], Esmee Venema[b,c], David M. Kent[d], Merel van Diepen[a], Friedo W. Dekker[a], Ewout W. Steyerberg[e], Ype de Jong[a,f,*]

[a]*Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands*
[b]*Department of Public Health, Erasmus MC University Medical Center, Rotterdam, The Netherlands*
[c]*Department of Emergency Medicine, Erasmus MC University Medical Center, Rotterdam, The Netherlands*
[d]*Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center, Boston, MA, USA*
[e]*Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands*
[f]*Department of Internal Medicine, Leiden University Medical Center, Leiden, The Netherlands*

Accepted 25 April 2023; Published online 2 May 2023

Abstract

Objectives: To (1) explore trends of risk of bias (ROB) in prediction research over time following key methodological publications, using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) and (2) assess the inter-rater agreement of the PROBAST.

Study Design and Setting: PubMed and Web of Science were searched for reviews with extractable PROBAST scores on domain and signaling question (SQ) level. ROB trends were visually correlated with yearly citations of key publications. Inter-rater agreement was assessed using Cohen's Kappa.

Results: One hundred and thirty nine systematic reviews were included, of which 85 reviews (containing 2,477 single studies) on domain level and 54 reviews (containing 2,458 single studies) on SQ level. High ROB was prevalent, especially in the Analysis domain, and overall trends of ROB remained relatively stable over time. The inter-rater agreement was low, both on domain (Kappa 0.04−0.26) and SQ level (Kappa −0.14 to 0.49).

Conclusion: Prediction model studies are at high ROB and time trends in ROB as assessed with the PROBAST remain relatively stable. These results might be explained by key publications having no influence on ROB or recency of key publications. Moreover, the trend may suffer from the low inter-rater agreement and ceiling effect of the PROBAST. The inter-rater agreement could potentially be improved by altering the PROBAST or providing training on how to apply the PROBAST.

Keywords: Prediction; Prognosis; Diagnosis; Bias; PROBAST; Inter-rater agreement

1. Introduction

Over the past few years, there has been a rapid increase in the development of prediction models [1,2]. Prediction models can be used to predict the presence of an outcome (diagnostic models) or the probability of reaching an outcome later in time (prognostic models) [3]. These models can be applied in personalized medicine to calculate an individual's risk. In theory, this allows for more tailored counseling and treatment options than can be achieved by relying on population-based risks [2,4−7]. Despite the abundance of prediction models, their uptake in clinical practice is limited, likely caused by a combination of methodological flaws, a limited number of external validations and impact analyses, and a lack of influence on clinical decision-making processes [7−13]. Methodological shortcomings in model development and

Funding: The work on this study by R.J.J. and M.v.D. was supported by a grant from the Dutch Kidney Foundation (20OK016).

Data availability statement: The data that support the findings of this study are available from the corresponding author, Y.d.J., upon reasonable request.

Author Contributions: Conceptualization and study design: L.L., E.S., and Y.d.J. Data acquisition: L.L. and Y.d.J. Formal analysis: R.J. and Y.d.J. Interpretation of data: L.L., R.J., E.V., D.K., M.v.D., F.D., E.S., and Y.d.J. Writing original draft: L.L. and Y.d.J. Writing review and editing: R.J., E.V., D.K., M.v.D., F.D., and E.S. Supervision: Y.d.J. Approval of final version: L.L., R.J., E.V., D.K., M.v.D., F.D., E.S., and Y.d.J.

* Corresponding author. Leiden University Medical Center, Leiden, The Netherlands. Tel.: +31-071-5264037.

E-mail address: y.de_jong@lumc.nl (Y. de Jong).

https://doi.org/10.1016/j.jclinepi.2023.04.012

What is new?

Key findings

- High risk of bias (ROB) is prevalent in all PROBAST domains, especially in the Analysis domain.
- ROB trends in prediction research remain stable over time.
- Most PROBAST domains and signaling questions suffer from poor inter-rater agreement.
- Lacking influence or recency of key publications, low inter-rater agreement and a ceiling effect may explain these stable ROB trends.

What this adds to what was known?

- This review includes PROBAST-scored prediction models from multiple medical specialties.
- Despite growing interest in methodology, as reflected by key-publication cites, trends in ROB remain stable.
- The inter-rater agreement of the PROBAST is poor between independent research groups.

What is the implication and what should change now?

- Poor inter-rater agreement limits using the PROBAST for ROB comparisons.
- The ceiling effect may be attenuated by reporting on a ROB scale or adding an 'intermediate' ROB category.
- Modifying signaling questions, decreasing the number of answering options or providing specialized training may improve the inter-rater agreement of the PROBAST.

validation can result in a high risk of bias (ROB) (e.g., overfitting, model instability, systematic differences between observed and predicted risks). This may lead to flawed or distorted conclusions regarding their predictive performance [9].

In 2019, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) was developed to critically assess the ROB and applicability of prediction model studies [1,14]. It appraises what "(…) shortcomings in study design, conduct, or analysis could lead to systematically distorted estimates of a model's predictive performance" [1]. The PROBAST contains 20 signaling questions (SQs) divided over four domains: Participants, Predictors, Outcome, and Analysis. One year after publication of the PROBAST, the tool was well received, as demonstrated by the high number of studies using the PROBAST [2]. On external validations, models with a low ROB as assessed using the PROBAST appear to perform better than models with a high ROB, as measured by the change in the area under the receiver operator characteristic curve [9]. Nevertheless, the PROBAST has received critique as well: the tool does not prioritize SQs, although some items are more likely to cause ROB than others [9]. Furthermore, some SQs are regarded as ambiguous and assessing ROB is regarded as time consuming and overly complex, ultimately resulting in a poor inter-rater agreement [9,15]. Therefore, a research group working independently from the original PROBAST team developed a shorter version of the PROBAST, which was found to closely reproduce the original classifications [9]. Two studies have assessed the inter-rater agreement of the PROBAST by comparing the assessments of researchers from their own research groups, which showed poor agreement [9,15]. The agreement may be even lower when comparing the assessments of independent research groups assessing the same study, but this has never been investigated.

Previous studies suggest a high ROB on all four domains of the PROBAST, but especially on the Analysis domain [9,15]. A number of publications aimed to inform researchers about prediction research, including textbooks, methodological papers, frameworks, guides, and other initiatives, including the PROGnosis RESearch Strategy (PROGRESS, 2013) [16−19], the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD, 2015) [20,21], and the Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK, 2018) [22,23]. To this date, no studies have investigated the trend in ROB as per PROBAST scores or evaluated the effect of these "key publications". In contrast, the effectiveness of studies aiming to improve the quality in other research fields has been investigated (e.g., the COnsolidated criteria for REporting Qualitative studies (COREQ) and the REMARK) [24,25]. The first aim of this study was to explore the trends in ROB in prediction research following key methodological publications. The second aim of this study was to investigate the inter-rater agreement of the PROBAST in independent research groups.

2. Methods

This review is reported as per the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 guidelines [26].

2.1. Search strategy and eligibility criteria

PubMed was searched for all reviews mentioning the PROBAST and Web of Science was searched for all reviews referring to the original PROBAST article [14] or

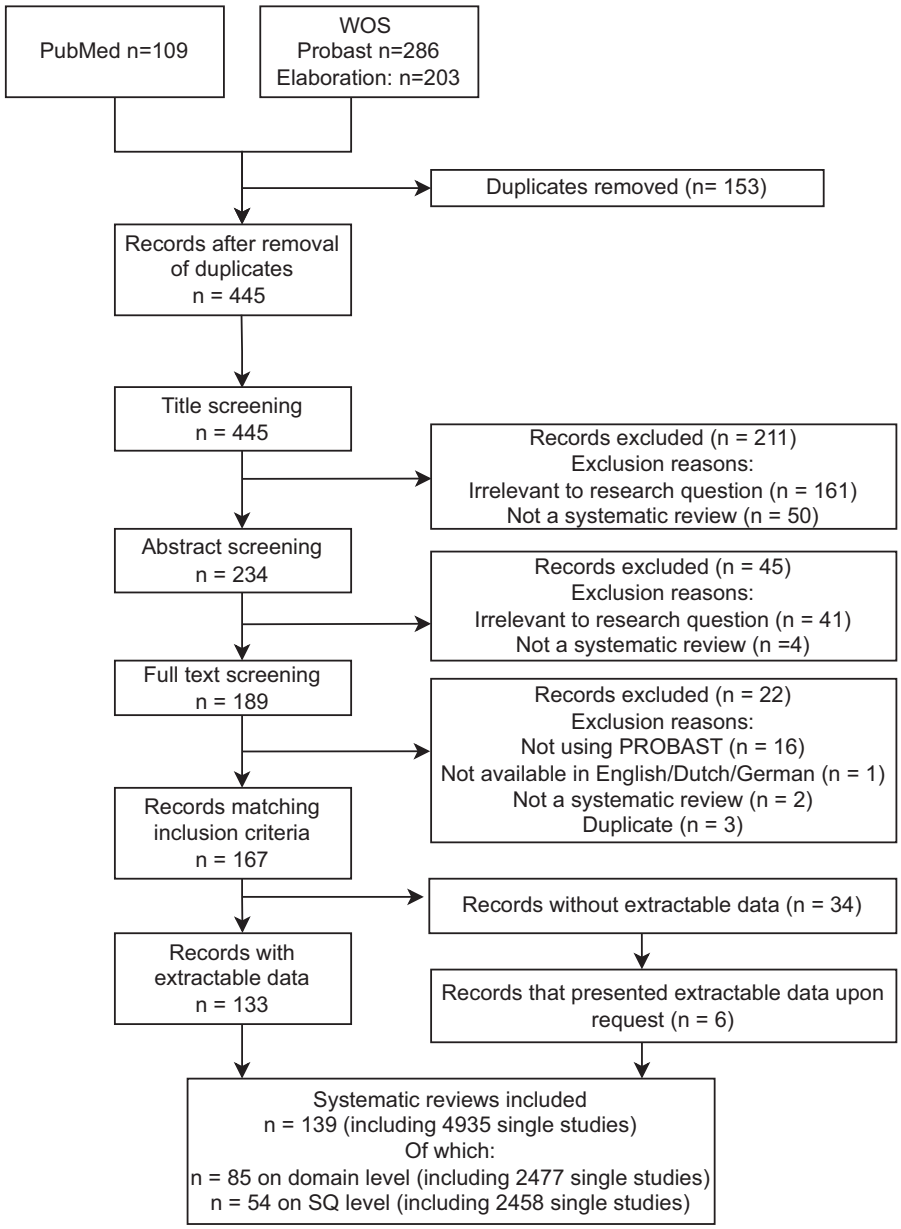

Fig. 1. Flowchart of study inclusion. PRISMA flowchart showing the total number of studies that were identified by a search in PubMed and Web of Science using our search strategy. The flowchart shows the study selection process with the number of excluded studies and the reasons for exclusion. This resulted in the inclusion of 85 reviews with extractable PROBAST scores on domain level and 54 reviews with extractable PROBAST scores on signaling question (SQ) level.

the PROBAST explanation and elaboration article [1]. This search was conducted on December 31, 2021. The search method is explained in more detail in Supplement section A (“Study selection methods”). After removal of duplicates, two authors (L.L./Y.d.J.) independently screened all titles, abstracts, and full texts for eligibility. Conflicts were discussed until consensus was reached. Articles were considered eligible for inclusion if they were (1) a systematic review, (2) contained at least one prediction model, and (3) used the PROBAST to assess the ROB of the included studies. Articles written in any other language than English, German, or Dutch were excluded.

2.2. *Data extraction*

Data extraction was conducted by two authors (L.L./Y.d.J.) using a predefined data extraction form. For the first aim, to evaluate the trends in ROB following key publications, we extracted the PROBAST scores for each original study included within the reviews (i.e., none of the studies were rated by us). The PROBAST scores of these ‘within-review’ studies were extracted per SQ or, if unavailable, per domain. If both scores were reported, only SQ scores were extracted and these were then recalculated to domain scores. We largely followed the PROBAST scoring rules:

a domain was rated as low ROB if all SQs of that domain were answered as 'yes', as high ROB if ≥ 1 SQ was answered as 'no', and as unclear ROB if ≥ 1 SQ was answered as 'probably no'/'probably yes'/'no information' while all other SQs were answered as 'yes'. This approach allowed us to use the most specific data (i.e., on SQ level) to obtain uniformly assessed domain scores. Low, unclear, or high ROB domains and SQs were subsequently recoded as 0, 1, and 2, respectively. We constructed three datasets: a dataset with SQ scores, a dataset with derived domain scores, which were the scores as extracted from the reviews, and a dataset which combined the calculated domain scores from SQs and the derived domain scores. Due to the large number of studies, we performed a pilot cross-check of the extracted data of the first 50 reviews. In case of unextractable data or unclarities, corresponding authors were contacted. Authors that only reported PROBAST scores at domain level were requested to share their data at SQ level. Key publications were identified during a consensus meeting with all authors. Inclusion was based on expert opinions on which articles have been influential, supplemented with searches in Google Scholar, PubMed, and Web of Science to identify articles with high numbers of citations/year. Total citation counts were extracted from Google Scholar and recalculated to citations per year. For the second aim, to assess the inter-rater agreement, we used the extracted Digital Object Identifiers (DOIs) and PubMed IDentifiers (PMIDs) of within-review studies to identify 'between-review' duplicates (i.e., studies included in multiple reviews by different authors). Duplicates were thus identified based on their identifier (i.e., DOI or PMID). These duplicates and their extracted PROBAST scores were manually cross-checked (L.L./Y.d.J.) to ensure that only between-review duplicates were included in our analysis. Identified 'within-review' duplicates (i.e., single studies included multiple times within the same review) were excluded.

2.3. Statistical analysis

For the first aim, to visualize the trend of ROB over time, we fit a LOcally Estimated Scatterplot Smoothing curve with a 95% confidence interval. The span was determined by generalized cross-validation using a local polynomial regression with an automatic smoothing parameter selection. Publication dates of the key publications and the number of citations per year were added to these plots. We also plotted a stacked bar chart indicating the number of studies with a low, unclear, and high ROB over time. The LOcally Estimated Scatterplot Smoothing curve and bar chart were also created for development and validation studies to separately investigate the ROB of these types of studies. All figures for our first aim were restricted on the data of the within-review studies published after 2000, because the number of studies published before 2000 was relatively low, resulting in a low power. For the second aim, to assess the inter-rater agreement in PROBAST scores, we analyzed the differences in PROBAST scores of between-review duplicates by Cohen's Kappa for the percentage of agreement per domain and SQ. All possible unique combinations of duplicates were included. For studies included in two reviews, there was one possible comparison; for studies included in three reviews, there were three possible comparisons; and for studies included in four reviews, there were six possible comparisons. R version 4.2.1 (R Foundation for Statistical Computing,

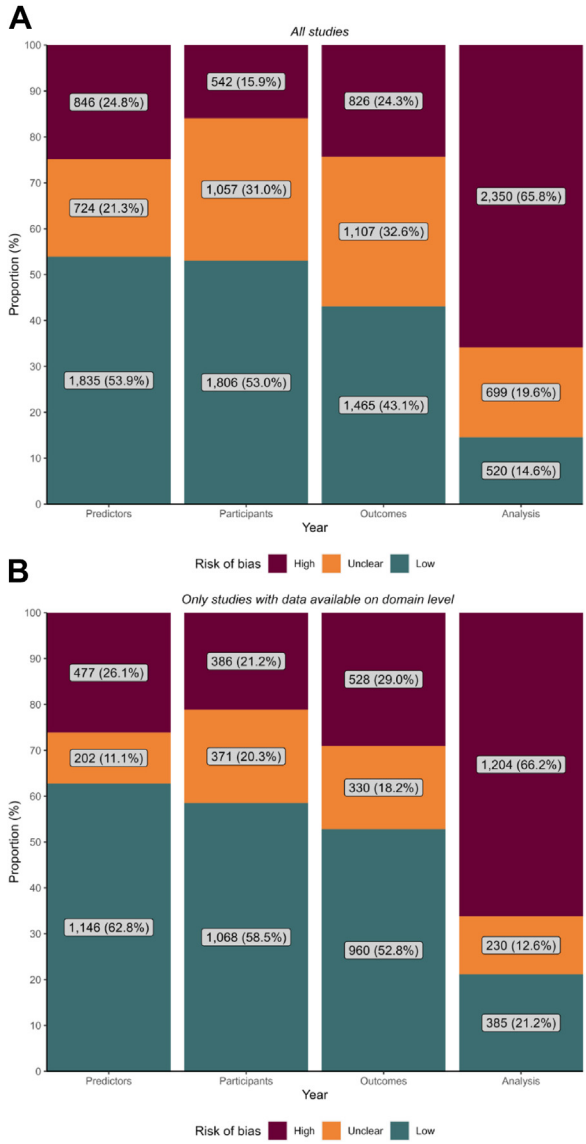

Fig. 2. (A and B) ROB per domain from 2000 to 2021, as assessed using the PROBAST. (A) includes data from 2000 to 2021 on derived domain level and SQ-calculated domain level (Predictors and Participants domain n = 3,405, Outcome domain n = 3,398, Analysis domain n = 3,569). (B) includes only the derived domain level data from 2000 to 2021 (Predictors and Participants domain n = 1,825, Outcome domain n = 1,818, Analysis domain n = 1,819). These figures show the absolute number and percentage of studies with a low, unclear, or high ROB on each domain.

Vienna, Austria) was used for all analyses. All annotated scripts are available online at https://github.com/rjjanse.

2.4. Sensitivity analyses

We conducted three sensitivity analyses. (1) We repeated our main analysis, in which duplicates with potentially different PROBAST scores were included multiple times, using only the average PROBAST score of between-review duplicates. (2) We repeated our main analysis, excluding the PROBAST scores of all between-review duplicates. (3) We recalculated Cohen's Kappa values of the between-review duplicates scored in three or more reviews by sequentially omitting reviews from the duplicate analysis. For instance, if review A, B, and C all assessed study X, our main analysis included comparisons AB, AC, and BC. When omitting review A, comparisons AB and AC were excluded, thus resulting in a Cohen's Kappa value based on only comparisons without A (i.e., BC). We then repeated this for B and C. This was done to assess whether the scoring of a single study had a large influence on the observed inter-rater agreement.

3. Results

3.1. Characteristics of included studies

The searches in PubMed and Web of Science resulted in a total of 598 articles, of which 445 were unique. Of these, 167 systematic reviews (hereafter referred to as 'reviews') met the eligibility criteria (see flowchart, Fig. 1). This included 133 reviews with extractable PROBAST data and 34 reviews without extractable PROBAST scores. However, upon request, three shared their data on domain level and three on SQ level. Of the reviews with PROBAST scores on domain level, five shared their data on SQ level upon request.

Table 1. Average PROBAST scores per domain and per signaling question

Domain/SQ	Total number of within-review studies	Within-review studies with extractable data	Completeness of data (%)	Mean score
Domain 1: Predictors	4,935	4,477	90.7	0.74
SQ 1.1.	2,458	2,000	81.4	0.48
SQ 1.2.	2,458	2,000	81.4	0.33
Domain 2: Participants	4,935	4,477	90.7	0.64
SQ 2.1.	2,458	2,000	81.4	0.35
SQ 2.2.	2,458	2,000	81.4	0.36
SQ 2.3.	2,458	2,000	81.4	0.19
Domain 3: Outcomes	4,935	4,470	90.6	0.82
SQ 3.1.	2,458	2,000	81.4	0.30
SQ 3.2.	2,458	2,000	81.4	0.29
SQ 3.3.	2,458	2,000	81.4	0.29
SQ 3.4.	2,458	2,000	81.4	0.36
SQ 3.5.	2,458	2,000	81.4	0.53
SQ 3.6.	2,458	1,978	80.5	0.22
Domain 4: Analysis	4,935	4,842	98.1	1.56
SQ 4.1.	2,458	2,371	96.5	0.83
SQ 4.2.	2,458	2,371	96.5	0.56
SQ 4.3.	2,458	1,913	77.8	0.66
SQ 4.4.	2,458	2,371	96.5	0.96
SQ 4.5.	2,458	2,000	81.4	0.77
SQ 4.6.	2,458	1,898	77.2	0.71
SQ 4.7.	2,458	1,910	77.7	0.71
SQ 4.8.	2,458	2,003	81.5	1.04
SQ 4.9.	2,458	1,544	62.8	0.49

The mean scores per domain are based on the derived and SQ-calculated domain scores. These mean scores can range from 0 to 2, with 0 indicating low ROB, that is, signaling questions (SQs) were scored as ''yes'', 1 indicating unclear ROB, that is, SQs scored as ''probably yes'', ''probably no'' or ''no information'', and 2 indicating high ROB, that is, SQs scored as ''no''. The completeness of data percentage for domains indicates the percentage of the 4,935 within-review studies that presented an extractable (derived or SQ-calculated) domain score for each specific domain. The completeness of data percentage for SQs indicates the percentage of the 2,458 within-review studies on SQ level that contained an extractable SQ score for each specific question.

In the end, 139 reviews (including 4,935 within-review studies) were included; there were 85 reviews on domain level (including 2,477 within-review studies for the derived domain dataset, Supplementary Table S1/S3) and 54 reviews on SQ level (including 2,458 within-review studies for the SQ dataset, Supplementary Table S2/S4). The remaining 28 reviews were excluded due to unextractable data. Because not all 139 reviews presented complete PROBAST

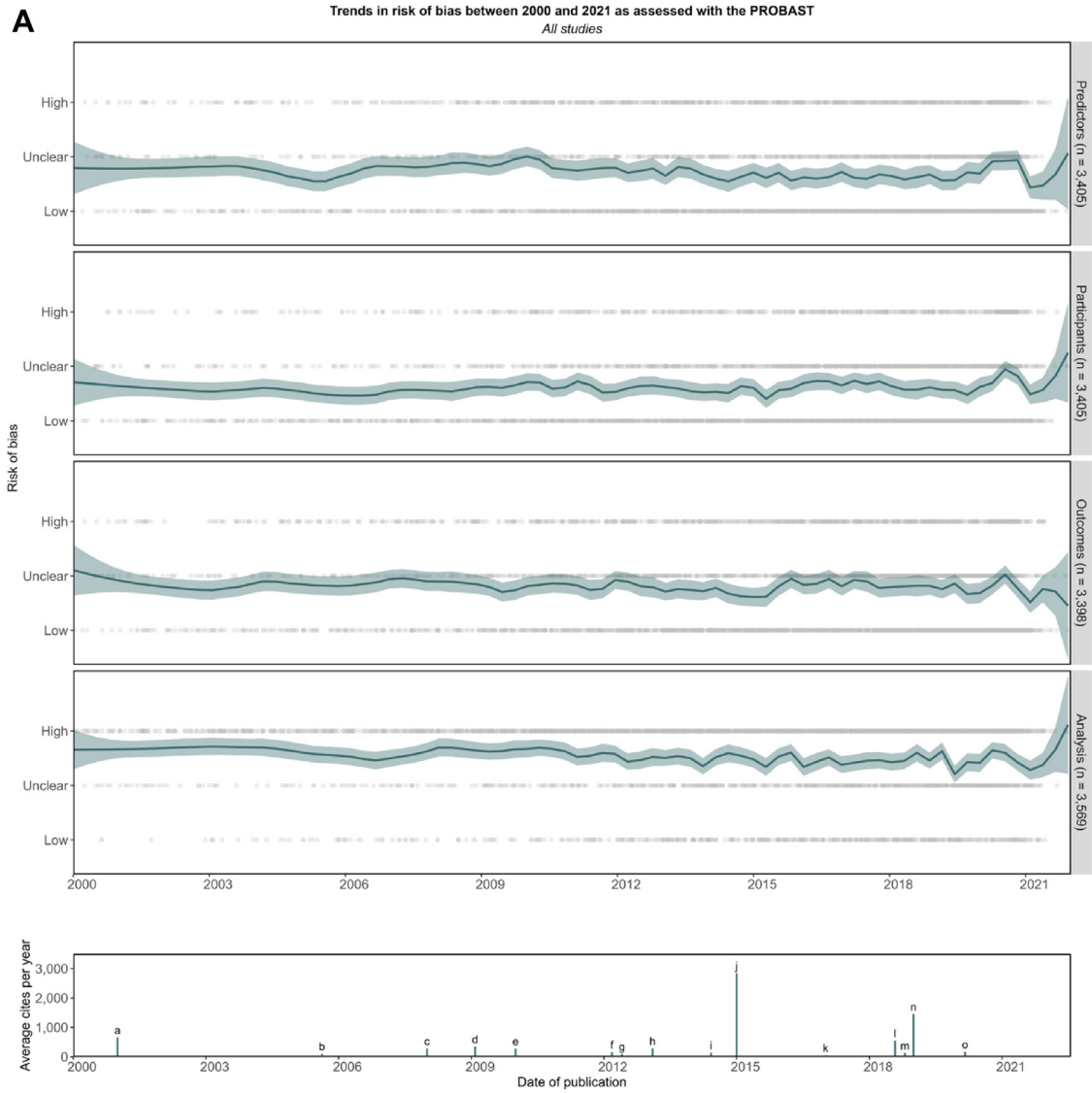

Fig. 3. (A and B) Trends in ROB from 2000 to 2021, as assessed by the PROBAST. (A) includes both data on derived domain level and SQ-calculated domain level (Predictors and Participants domain n = 3,405, Outcome domain n = 3,398, Analysis domain n = 3,569). (B) includes only data on derived domain level (Predictors and Participants domain n = 1,825, Outcome domain n = 1,818, Analysis domain n = 1,819). The vertical lines represent key publications in the field of prediction modelling, with higher lines indicating a higher average amount of citations per year. From left to right, the lines indicate the following key publications: A = 2001-01-01: Harrel's textbook first edition with on average 662 cites per year. B = 2005-08-17: REMARK with on average 106 cites per year. C = 2008-01-01: Steyerberg's textbook first edition with on average 285 cites per year. D = 2009-02-01: Moon's BMJ series with on average 346 cites per year E = 2010-01-01: Steyerberg et al. (Epidemiol 2010) on average 281 cites per year. F = 2012-03-07: Moon's Heart series with on average 159 cites per year. G = 2012-05-29: REMARK E&E paper with on average 97 cites per year. H = 2013-02-05: PROGRESS with on average 283 cites per year. I = 2014-06-04: Steyerberg et al. (Eur Heart J 2014) with on average 134 cites per year. J = 2015-01-01: TRIPOD and Harrel's textbook second edition with on average 2,848 cites per year. K = 2017-01-05: Debray et al. (BMJ 2017) with on average 51 cites per year L = 2018-08-01: REMARK with on average 548 cites per year. M = 2018-10-22 and 2018-10-24: Riley et al. (Stat Med 2019) and Riley et al. (Stat Med 2019) with on average 126 cites per year (no distinction could be made between cites of the first edition and of the second edition for the textbooks). N = 2019-01-01: Riley's textbook and Steyerberg's textbook second edition with on average 1,140 cites per year. O = 2020-03-01: Riley et al. (BMJ 2020) with on average 173 cites per year.

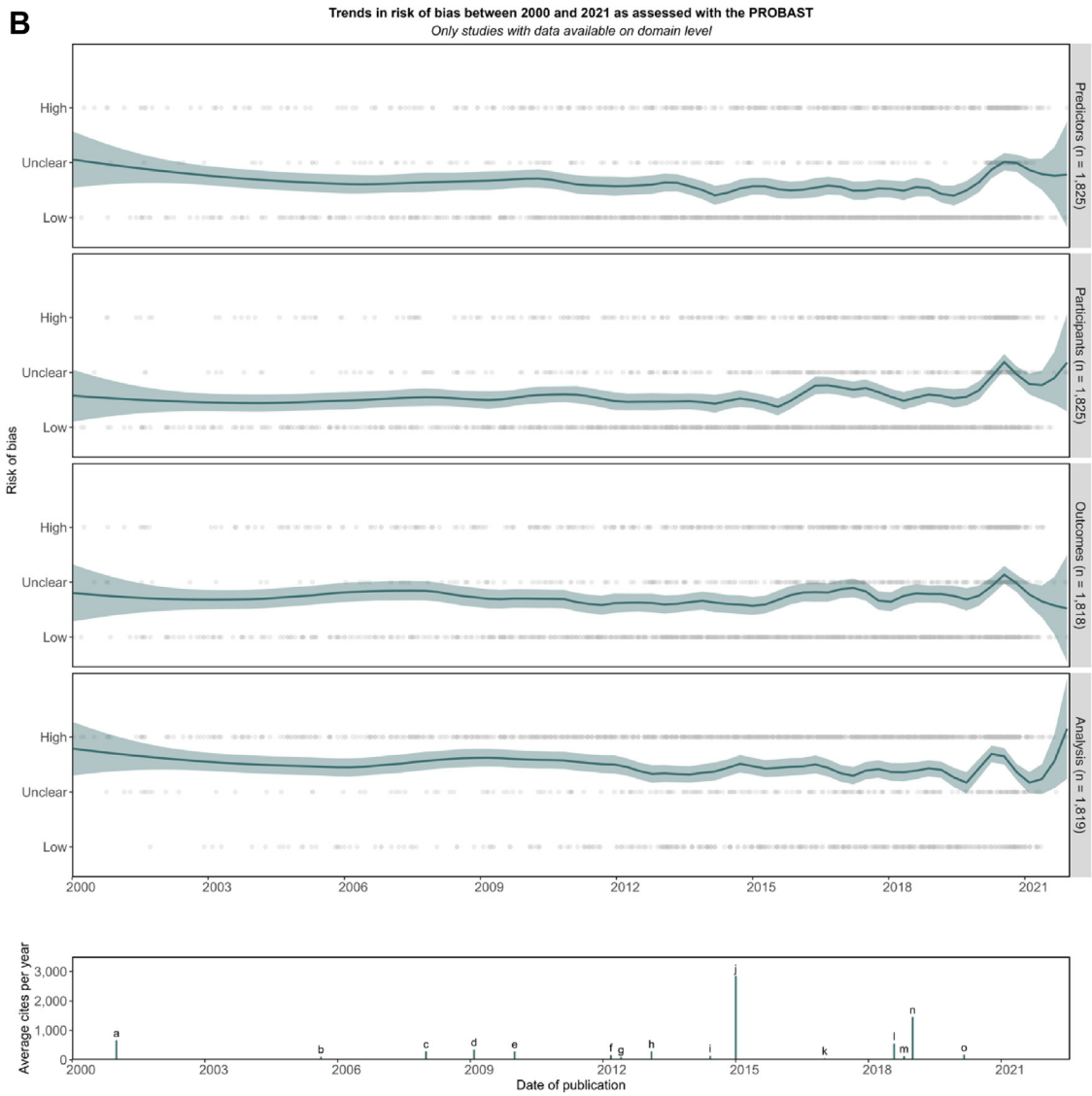

Fig. 3. Continued.

assessments (i.e., some reviews contained missing values on domain or SQ level), the *n* value in our analyses varies per domain and SQ. The included reviews were published between 2016 (using preliminary versions of the PROBAST) and 2022, including within-review studies from 1972 to 2021 at domain level and 1966–2021 at SQ level. The included reviews covered 18 medical specialties, ranging from surgery to psychiatry, as detailed in Supplementary section A (''Reviews per specialty''). Of the included reviews, 16 contained diagnostic prediction models, 100 contained prognostic prediction models, and 23 contained both diagnostic and prognostic prediction models. The reviews included 3,330 model development studies and 1,605 validation studies. Metadata could be automatically extracted for 4,590 DOI-tagged studies and for 217 studies using a PMID. Metadata of the remaining 128 studies were extracted by hand. In total, the 4,935 within-review studies of these 139 reviews contained 157 between-review duplicates, which were used for the inter-rater agreement analysis.

3.2. First aim: trends of risk of bias over time

On derived and SQ-calculated domain level, there were 3,405 within-review studies for the Participants and Predictors domain, 3,398 for the Outcome domain, and 2,569 for the Analysis domain. On derived domain level, there were 1,825 within-review studies for the Participants and Predictors domain, 1,818 for the Outcome domain, and 1,819 for the Analysis domain. Overall, high ROB was prevalent in all four domains, but especially in the Analysis domain (high ROB: 25% in the Predictors domain, 16% in the

Participants domain, 24% in the Outcome domain, and 66% in the Analysis domain; Fig. 2 and Table 1).

We identified 16 (groups of) publications as key publications. Their publication dates and number of citations were extracted from Google Scholar on November 5, 2022. For references of these publications and the number of citations, see Supplementary section A (“Overview of key publications”). ROB remained relatively stable following key publications (Fig. 3). See Figures 4 and 5 for the separate analyses of the ROB of development and validation studies, which show similar trends in ROB. The stacked bar chart of Figure 6 shows that over time, the proportion of studies at low ROB remained relatively stable, whereas the proportion of studies at high ROB slightly decreased and the proportion of studies at unclear ROB slightly increased. Figures using the data of all within-review studies without any time restriction are available in Supplementary Section C.

3.3. Second aim: inter-rater agreement of the PROBAST

On domain level, 68 within-review studies were included in two different reviews (equaling 68 possible comparisons), seven were included in three reviews (21 comparisons), and three were included in four reviews (18 comparisons) (Supplementary Table S5). On SQ level, 76 within-review studies were included in two reviews (76 comparisons) and three in three reviews (nine comparisons) (Supplementary Table S6). For clarity, reviews contained only the final PROBAST scores (i.e., individual scores from members of the same research team before merging them to these final scores were not reported in any included review). The inter-rater agreement of these studies showed a Cohen’s Kappa of 0.22 on the Predictors domain (53% inter-rater agreement), 0.04 on the Participant domain (56% inter-rater agreement), 0.26 on the Outcome domain (48% inter-rater agreement), and 0.06 on the Analysis domain (59% inter-rater agreement). The Cohen’s Kappa values of the individual SQs ranged from −0.14 to 0.49, as shown in Table 2 and Figure 7.

3.4. Sensitivity analyses

The first sensitivity analysis using the average PROBAST scores of between-review duplicates and the second sensitivity analysis excluding the PROBAST scores of between-review duplicates both showed a similar trend in ROB as our main analysis (Supplementary Figures S10-15). The third sensitivity analysis demonstrated that no single study considerably influenced Cohen’s Kappa (Supplementary Tables S7-9).

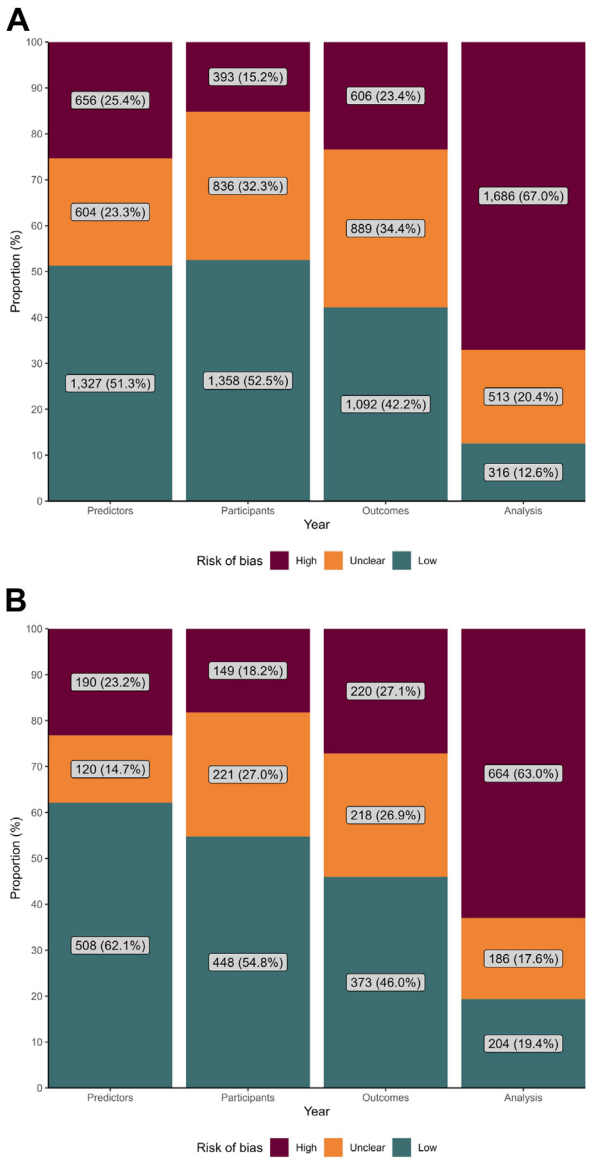

Fig. 4. (A and B) ROB per domain from 2000 to 2021, as assessed using the PROBAST, for development and validation studies separately. (A) includes data from 2000 to 2021 on derived and SQ-calculated domain level for development studies (Predictors and Participants domain n = 3,405, Outcome domain n = 3,398, Analysis domain n = 3,569). (B) includes data from 2000 to 2021 on derived and SQ-calculated domain level for validation studies (Predictors and Participants domain n = 1,825, Outcome domain n = 1,818, Analysis domain n = 1,819). These figures show the absolute number and percentage of studies with a low, unclear, or high ROB on each domain.

4. Discussion

In this metareview, we investigated the trends in ROB in prediction research following key methodological publications and explored the inter-rater agreement of the PROBAST. We demonstrated a stable trend in reported ROB, largely unaffected by key publications. High ROB was highly prevalent, especially in the Analysis domain, and remained consistently high over time. Furthermore, we demonstrated that the PROBAST has a low inter-rater agreement on both domain and SQ level; Kappa values ranging between 0.04 and 0.26 on domain level and

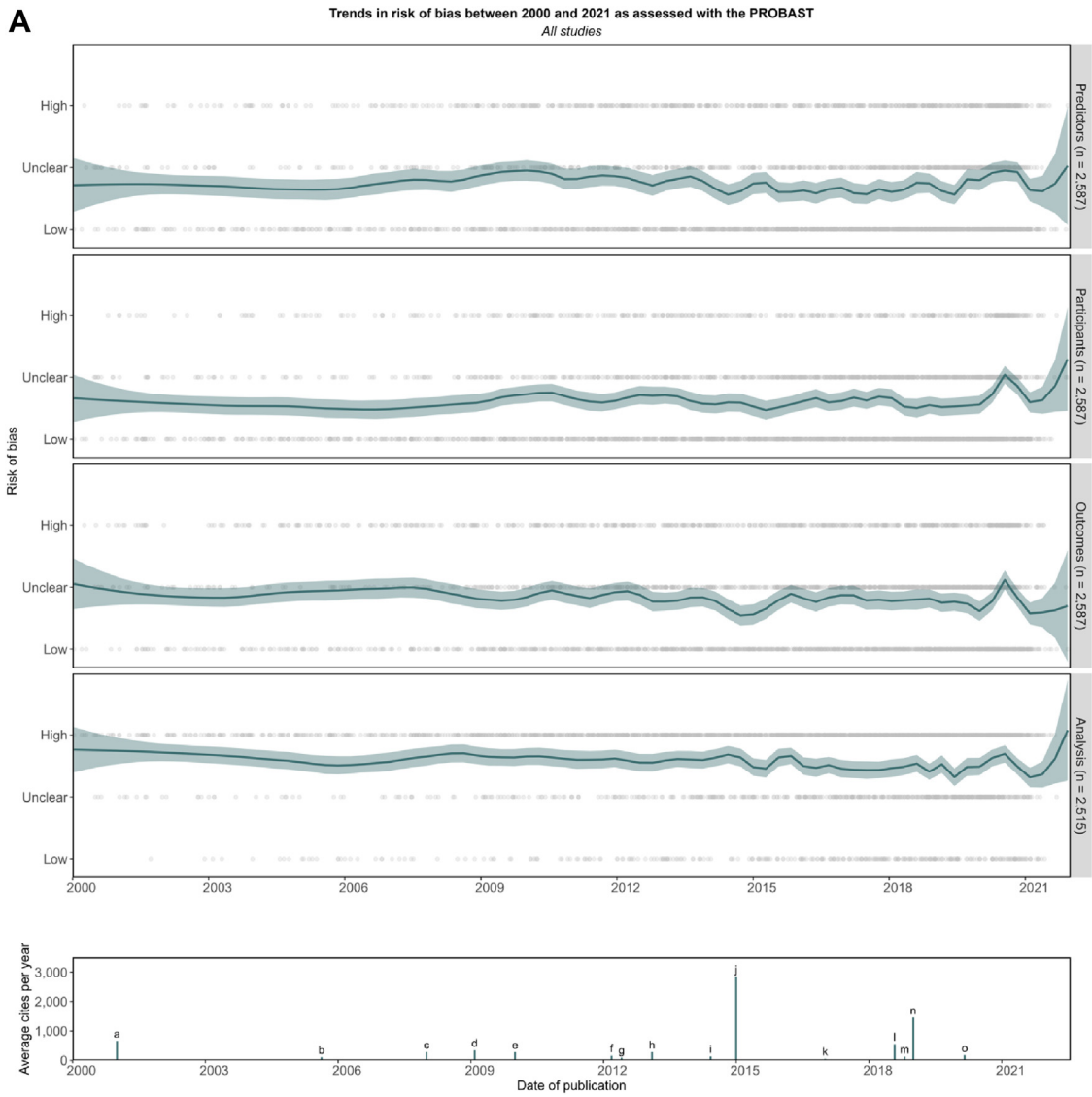

Fig. 5. (A and B) Trends in ROB from 2000 to 2021, as assessed by the PROBAST, for development and validation studies separately. (A) includes data on derived and SQ-calculated domain level for development studies (Predictors, Participants, and Outcome domain n = 2,587, Analysis domain n = 2,515). (B) includes data on derived and SQ-calculated domain level for validation studies (Predictors and Participants domain n = 818, Outcome domain n = 811, Analysis domain n = 1,054). The vertical lines represent key publications in the field of prediction modelling, with higher lines indicating a higher average amount of citations per year. From left to right, the lines indicate the following key publications: A = 2001-01-01: Harrel's textbook first edition with on average 662 cites per year. B = 2005-08-17: REMARK with on average 106 cites per year. C = 2008-01-01: Steyerberg's textbook first edition with on average 285 cites per year. D = 2009-02-01: Moon's BMJ series with on average 346 cites per year E = 2010-01-01: Steyerberg et al. (Epidemiol 2010) on average 281 cites per year. F = 2012-03-07: Moon's Heart series with on average 159 cites per year. G = 2012-05-29: REMARK E&E paper with on average 97 cites per year. H = 2013-02-05: PROGRESS with on average 283 cites per year. I = 2014-06-04: Steyerberg et al. (Eur Heart J 2014) with on average 134 cites per year. J = 2015-01-01: TRIPOD and Harrel's textbook second edition with on average 2,848 cites per year. K = 2017-01-05: Debray et al. (BMJ 2017) with on average 51 cites per year L = 2018-08-01: REMARK with on average 548 cites per year. M = 2018-10-22 and 2018-10-24: Riley et al. (Stat Med 2019) and Riley et al. (Stat Med 2019) with on average 126 cites per year (no distinction could be made between cites of the first edition and of the second edition for the textbooks). N = 2019-01-01: Riley's textbook and Steyerberg's textbook second edition with on average 1,140 cites per year. O = 2020-03-01: Riley et al. (BMJ 2020) with on average 173 cites per year.

between −0.14 and 0.49 on SQ level indicate poor agreement [27,28]. Some questions (e.g., all SQs with a Cohen's Kappa below 0.1, i.e., no agreement [27,28]: SQ 1.1, 2.1, 2.3, 3.2, 4.1, 4.3, 4.6, 4.7, 4.8, and 4.9) seem more at risk for divergent scoring than others (e.g., all SQs with a Cohen's Kappa above 0.4, i.e., moderate agreement [27,28]: SQ 3.3 and 3.4).

There can be several explanations for the lack of improvement following key publications. Although the key publications are well known as indicated by the number

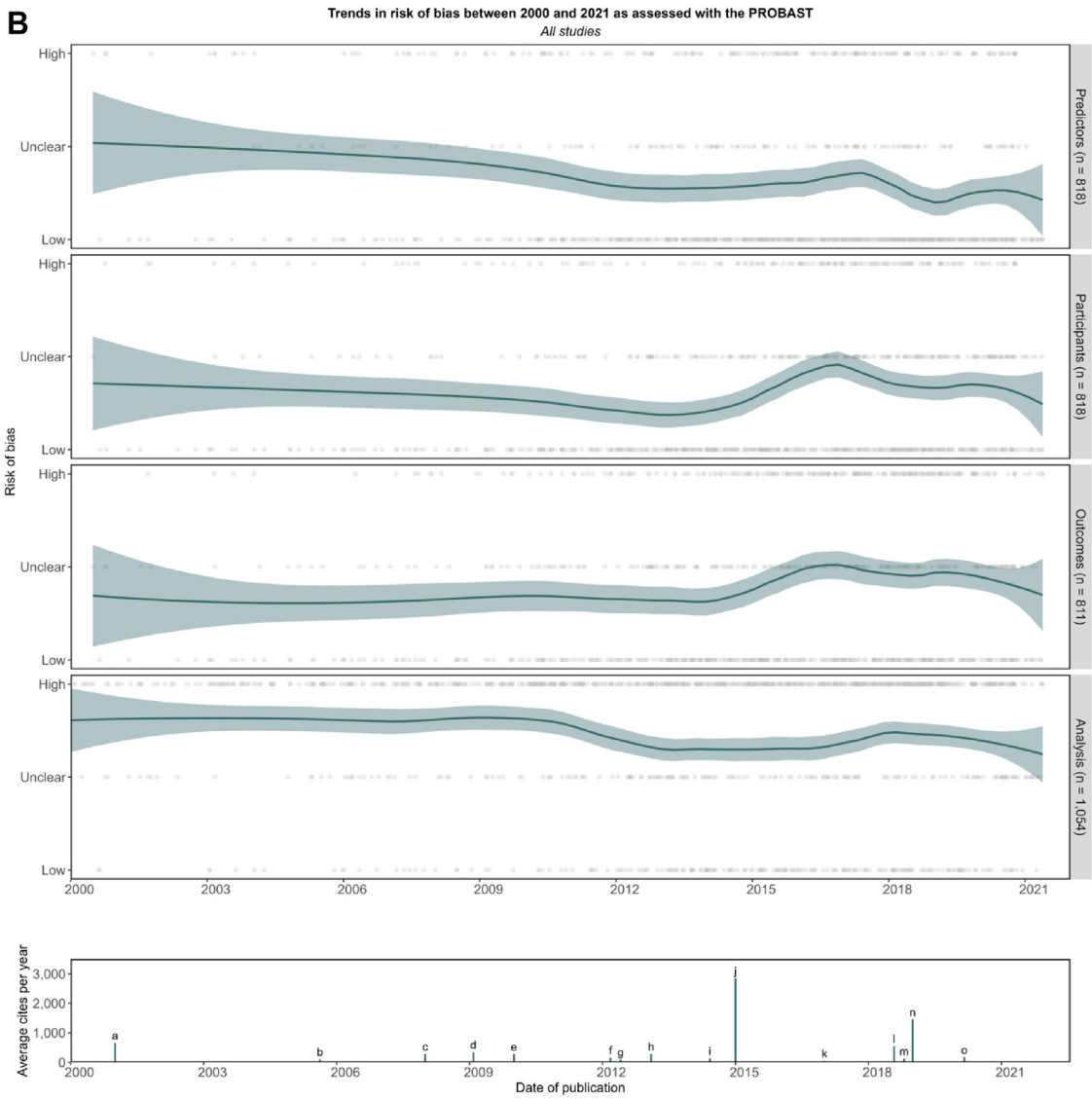

Fig. 5. Continued.

of citations, key publications simply might have had no effect on ROB in prediction studies. However, in other research fields, the publication of research guidelines (such as the COREQ for qualitative reviews) resulted in a positive trend in reporting quality [4]. Second, the time period between our analysis and publication of key publications might be too short to observe a decreasing trend in ROB. It may take years for publications to have an impact on the quality of research. Several years after the introduction of the REMARK, a similar study on its effect also showed that there were no significant improvements yet [25]. Third, the PROBAST may be an inappropriate tool to detect a trend in ROB due to its poor inter-rater agreement. Finally, our study highlights the ceiling effect of the PROBAST. It is impossible to distinguish domains with a high ROB based on one SQ answered as 'no' from domains with an extremely high ROB based on all SQs answered as 'no'. This leads to a loss of information and an inability to detect degrees of change in ROB. Introducing an ''intermediate'' ROB category or reporting ROB on a scale from low to high could be valuable [9,29], because there appears to be substantial variation in the methodological quality of studies at high ROB [9]. Besides revising the PROBAST, the quality of original prediction studies could be improved, potentially decreasing the ROB. For that aim, guidelines such as TRIPOD and REMARK have been developed. It has been suggested to preregister statistical analysis plans to improve their transparency and decrease the ROB in the Analysis domain [22,25,30,31]. A thorough methodology and reporting of analyses is essential to increase the

usability and uptake of prediction models in clinical practice [7,9,32–40].

The low inter-rater agreement implies that scoring of the PROBAST items is less straightforward than expected, especially so for the Analysis domain [9]. The subjectivity that the PROBAST allows for grading overall domain ROB may contribute to the low inter-rater agreement. Although the PROBAST provides grading rules, the authors state that any signaling question answered as 'no' or 'probably no' flags the potential for bias; assessors will need to use their own judgment to determine whether the domain should be rated as high, low, or unclear ROB. A 'no' answer does not automatically result in a high ROB rating [1]. A check of five randomly selected reviews with data on SQ and domain level showed that two studies scored all domains as per the grading rules [41,42]. The remaining three studies used the freedom to score domains as low ROB despite $\geq$ 1 SQ answered as 'no'. On within-review level (i.e., 86 within review studies, including 344 overall domain scores), these three reviews classified seven of the 344 domains as low ROB despite $\geq$ 1 SQ answered as 'no' [43–45]. Another explanation for the poor agreement may be the number of answering options (i.e., 'yes', 'probably yes', 'probably no', 'no', and 'no information'). Decreasing this to three options (i.e., 'yes', 'no', and 'unclear'), as suggested by the Cochrane Handbook and as applied by other ROB tools (e.g., the COREQ and the different Critical Appraisal Skills Program ROB tools), might increase the inter-rater agreement [46–48]. Limited data are available on the inter-rater agreement of other scoring tools. Near perfect agreement was reported for the PRISMA extension for Abstracts (PRISMA-A, Cohen's Kappa 0.81-0.92) [49,50]. A further explanation for the low inter-rater agreement of the PROBAST is that differences in scoring may arise from researchers with different statistical experience, which is required for answering some of the PROBAST questions [9,14,49]. This study did not focus on who performed the PROBAST assessments and how experienced these researchers were. However, a previous study showed that even experienced researchers have a low inter-rater agreement (Cohen's Kappa of 0.33) [9]. This implies that the low inter-rater agreement may be more dependent on the PROBAST tool itself than on the qualifications of the researchers using the instrument. Changes may be considered, especially to the Analysis domain. Another potential solution has been studied recently, showing that training induces significant improvements in the inter-rater agreement of two of the four domains and the overall ROB [15]. Regardless, the inter-rater agreement remained modest with Cohen's Kappa scores between 0.17 and 0.40. The low inter-rater agreement, in line with our study, may limit the usefulness of the tool for ROB assessment in prediction research. Finally, it has been argued that some SQs are more correlated to high ROB than others. This resulted in the development of a short form of the PROBAST, consisting of six SQs, with 98% sensitivity and 100% specificity to predict overall domain ROB [9]. Although the short form has been proven reliable in cardiovascular prediction models, external validity in other medical disciplines is yet to be investigated to further increase its reliability and usefulness.

Table 2. Inter-rater agreement of PROBAST scores at domain and SQ level

Domain/SQ	*N*	Agreement	% agreement	Cohen's Kappa
Predictors	76	41	53.2	0.22
SQ 1.1.	76	45	58.4	0.06
SQ 1.2.	76	50	64.9	0.20
Participants	76	43	55.8	0.04
SQ 2.1.	76	46	59.7	−0.14
SQ 2.2.	76	58	75.3	0.19
SQ 2.3.	76	58	75.3	−0.04
Outcomes	76	37	48.1	0.26
SQ 3.1.	76	55	71.4	0.23
SQ 3.2.	76	45	58.4	0.10
SQ 3.3.	76	59	76.6	0.44
SQ 3.4.	76	58	75.3	0.49
SQ 3.5.	76	47	61.0	0.26
SQ 3.6.	74	59	79.7	0.39
Analysis	75	44	58.7	0.06
SQ 4.1.	75	27	36.0	0.09
SQ 4.2.	75	50	66.7	0.29
SQ 4.3.	67	31	46.3	−0.02
SQ 4.4.	75	45	60.0	0.35
SQ 4.5.	75	47	62.7	0.26
SQ 4.6.	67	14	20.9	−0.05
SQ 4.7.	67	17	25.4	−0.05
SQ 4.8.	75	29	38.7	0.04
SQ 4.9.	67	24	35.8	−0.06

On derived domain level, 68 within-review studies were included in two different reviews, seven were included in three reviews, and three were included in four reviews. On SQ level, there were 76 within-review studies included in two reviews and three included in three reviews. The *n*-value column in this table stands for the number of comparisons that could be formed for each domain or SQ. The agreement column indicates the absolute number of comparisons with an identical PROBAST score on each domain or SQ. This table further includes the percentage of agreement and Cohen's Kappa per SQ and per domain.

4.1. Strengths and limitations of this study

Strengths of this study are the analysis of the trends in ROB over time, which has not been investigated before, the large number of included reviews with PROBAST data on domain and SQ level, and the independent assessment of PROBAST scores by researchers from different research groups. Our study reflects the current use of the PROBAST in prediction research. This study also has several

Fig. 6. Stacked bar chart of ROB from 2000–2021. (A) shows the ROB over time for studies on derived domain and SQ-calculated domain level (Predictors and Participants domain $n = 3{,}405$, Outcome domain $n = 3{,}398$, Analysis domain $n = 3{,}569$). (B) shows the ROB over time for the included studies on derived domain level only (Predictors and Participants domain $n = 1{,}825$, Outcome domain $n = 1{,}818$, Analysis domain $n = 1{,}819$). The bars indicate the percentage of studies with a low, unclear, and high risk of bias per year.

limitations. First, we could not include the PROBAST data of all 167 reviews because these data were unavailable or nonextractable for some reviews ($n = 28$). Additionally, only two bibliographic databases were searched for eligible reviews and the number of between-review duplicates was relatively low ($n = 157$), perhaps as a consequence of that. Identifying more reviews with extractable data and more between-review duplicates would have allowed for a better-powered conclusion on the inter-rater agreement, although we believe our conclusions are based on enough studies to warrant publication. Furthermore, we did not use the original division of PROBAST scores into five groups per signaling question (i.e., ''yes'', ''probably yes'', ''probably no'', ''no'', and ''no information''), but categorized them into three groups by composing a group with an unclear ROB. Although this reduced available information, we think it helped uniformize the calculated domain scores. Additionally, we have included figures of our analyses on derived domain scores and total (calculated and derived) domain scores, which were graphically similar. Next, as PROBAST scores were manually extracted, there was a risk of extraction errors. However, a cross-check of the extracted data of 50 of the 139 articles showed almost perfect agreement. Moreover, potential extraction errors will likely result in nondifferential misclassification. Another limitation of this study is the possible effect of unequal weighting of studies, with more weight on the PROBAST scores of reviews that contain the most within-review studies [9,51–54]. This effect may be minimal, because the differences in the number of within-review studies were relatively small. Our sensitivity analysis of Cohen's Kappa values showed that no single study disproportionately influenced our findings.

5. Conclusion

Our review demonstrates little change in the assessed ROB of published prediction model studies over time following key publications. Potential reasons for the lack

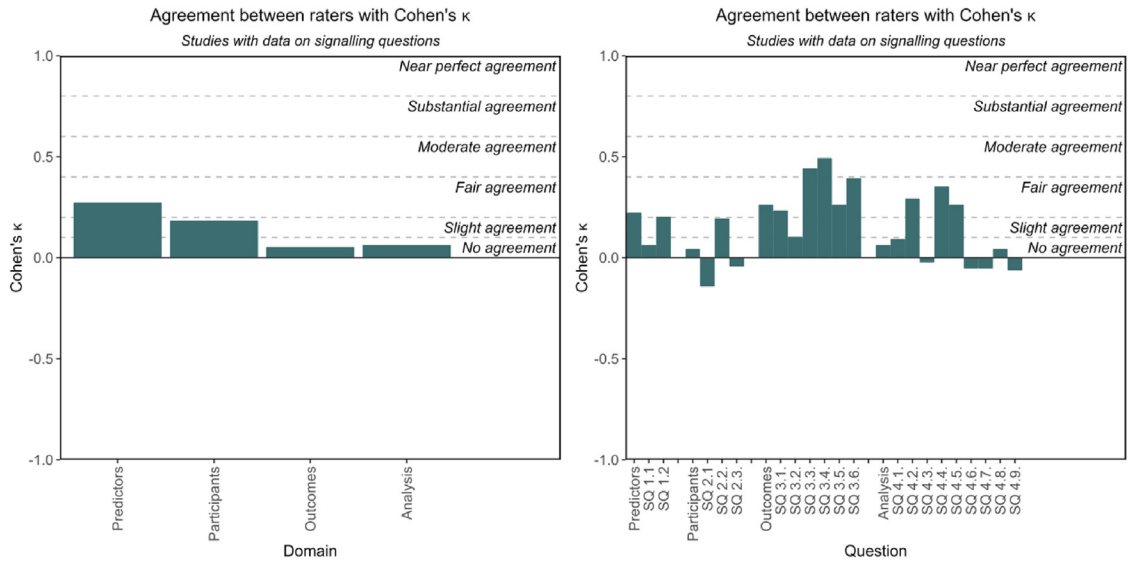

Fig. 7. Overview of inter-rater agreement of the PROBAST, using Cohen's Kappa. The figure on the left shows the Cohen's Kappa values per derived domain and the figure on the right shows the Cohen's Kappa values per signaling question (SQ) without any time restriction. On derived domain level, 68 within-review studies were included in two different reviews (equaling 68 unique comparisons), seven were included in three reviews (equaling 21 comparisons), and three were included in four reviews (equaling 18 comparisons). On SQ level, there were 76 within-review studies included in two reviews (equaling 76 unique comparisons) and three in three reviews (equaling nine comparisons). Kappa values can range from −1 (indicating no agreement) to 1 (indicating perfect agreement). Kappa values can be interpreted as no agreement (Kappa < 0.1), slight agreement (Kappa 0.1–0.2), fair agreement (Kappa 0.2–0.4), moderate agreement (Kappa 0.4–0.6), substantial agreement (Kappa 0.6–0.8), and near perfect agreement (Kappa 0.8–1.0) [27,28].

of improvement include that methodological quality may remain relatively unaffected by the key publications examined, that insufficient time may have passed to observe the influence of key publications or that the PROBAST may be incapable of assessing ROB trends because of the poor inter-rater agreement and the ceiling effect. Modification of the PROBAST itself focused on the SQ with the lowest inter-rater agreements, perhaps combined with specialized training for researchers using the PROBAST, may address these latter concerns.

Declaration of competing interest

The work on this study by R.J.J. and M.v.D. was supported by a grant from the Dutch Kidney Foundation (20OK016). All other authors declare no conflict of interest.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2023.04.012.

References

[1] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1–33.

[2] de Jong Y, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, van Diepen M. Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk of Bias ASsessment Tool (PROBAST). Nephrology (Carlton) 2021;26(12):939–47.

[3] Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

[4] Moons KG, 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.

[5] Moons KG, 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.

[6] Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009;338:b604.

[7] Helmrich I, Mikolić A, Kent DM, Lingsma HF, Wynants L, Steyerberg EW, et al. Does poor methodological quality of prediction modeling studies translate to poor model performance? An illustration in traumatic brain injury. Diagn Progn Res 2022;6(1):8.

[8] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J 2021;14(1):49–58.

[9] Venema E, Wessler BS, Paulus JK, Salah R, Raman G, Leung LY, et al. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. J Clin Epidemiol 2021;138:32–9.

[10] Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol 2013;66:268–77.

[11] Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68:25−34.
[12] Kent DM, Nelson J, Upshaw JN, Gulati G, Brazil R, Venema E, et al. PCORI final research reports. Using different data sets to test how well clinical prediction models work to predict patients' risk of Heart disease. Washington (DC): Patient-Centered Outcomes Research Institute (PCORI); 2021.
[13] Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, et al. External validations of cardiovascular clinical prediction models: a large-scale review of the literature. Circ Cardiovasc Qual Outcomes 2021;14(8):e007858.
[14] 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.
[15] Kaiser AB, Pfahlberg S, Mathes W, Uter K, Diehl T, Steeb MV, et al. Inter-rater agreement in assessing risk of bias in melanoma prediction studies using the prediction model risk of bias assessment tool (PROBAST): results from a controlled experiment on the effect of specific rater training. J Clin Med 2023;12(5):1976.
[16] Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ 2013;346:e5595.
[17] Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med 2013;10(2):e1001380.
[18] 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.
[19] Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013;346:e5793.
[20] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg 2015;102(3):148−58.
[21] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1−73.
[22] Sauerbrei W, Taube SE, McShane LM, Cavenagh MM, Altman DG. Reporting recommendations for tumor marker prognostic studies (REMARK): an abridged explanation and elaboration. J Natl Cancer Inst 2018;110:803−11.
[23] McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer 2005;93:387−91.
[24] de Jong Y, van der Willik EM, Milders J, Voorend CGN, Morton RL, Dekker FW, et al. A meta-review demonstrates improved reporting quality of qualitative reviews following the publication of COREQ- and ENTREQ-checklists, regardless of modest uptake. BMC Med Res Methodol 2021;21:184.
[25] Sekula P, Mallett S, Altman DG, Sauerbrei W. Did the reporting of prognostic studies of tumour markers improve since the introduction of REMARK guideline? A comparison of reporting in published articles. PLoS One 2017;12:e0178531.
[26] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J Clin Epidemiol 2021;134:178−89.
[27] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.
[28] Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Soc Adm Pharm 2013;9(3):330−8.
[29] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.
[30] Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Med 2012;10(1):51.
[31] Sauerbrei W, Haeussler T, Balmford J, Huebner M. Structured reporting to improve transparency of analyses in prognostic marker studies. BMC Med 2022;20(1):184.
[32] Wynants L, Van Calster B, Bonten MMJ, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020;369:11.
[33] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12−22.
[34] Wynants L, Van Calster B, Bonten MM, Collins GS, Debray TP, De Vos M, et al. Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection. BMJ 2020;369. https://doi.org/10.1101/2020.03.24.20041020.
[35] Steyerberg EW, Uno H, Ioannidis JPA, van Calster B. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol 2018;98:133−43.
[36] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17(1):230.
[37] Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 2010;8:20.
[38] Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC Med 2010;8:21.
[39] Ramspek CL, de Jong Y, Dekker FW, van Diepen M. Towards the best kidney failure prediction tool: a systematic review and selection aid. Nephrol Dial Transplant 2019;35:1527−38.
[40] de Jong Y, Ramspek CL, van der Endt VHW, Rookmaaker MB, Blankestijn PJ, Vernooij RWM, et al. A systematic review and external validation of stroke prediction models demonstrates poor performance in dialysis patients. J Clin Epidemiol 2020;123:69−79.
[41] Carrillo-Larco RM, Aparcana-Granda DJ, Mejia JR, Barengo NC, Bernabe-Ortiz A. Risk scores for type 2 diabetes mellitus in Latin America: a systematic review of population-based studies. Diabet Med 2019;36(12):1573−84.
[42] Mawdsley E, Reynolds B, Cullen B. A systematic review of the effectiveness of machine learning for predicting psychosocial outcomes in acquired brain injury: which algorithms are used and why? J Neuropsychol 2021;15(3):319−39.
[43] Groot OQ, Ogink PT, Lans A, Twining PK, Kapoor ND, DiGiovanni W, et al. Machine learning prediction models in orthopedic surgery: a systematic review in transparent reporting. J Orthop Res 2022;40:475−83.
[44] Fernandez-Felix BM, Barca LV, Garcia-Esquinas E, Correa-Pérez A, Fernández-Hidalgo N, Muriel A, et al. Prognostic models for mortality after cardiac surgery in patients with infective endocarditis: a systematic review and aggregation of prediction models. Clin Microbiol Infect 2021;27(10):1422−30.
[45] Van Remoortel H, Scheers H, De Buck E, Haenen W, Vandekerckhove P. Prediction modelling studies for medical usage rates in mass gatherings: a systematic review. PLoS One 2020;15:e0234977.
[46] Singh J. Critical appraisal skills programme. J Pharmacol Pharmacother 2013;4:76−7.
[47] Higgins JPTTJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane Handbook for Systematic Reviews of Interventions Cochrane. Cochrane; 2022:6.3.
[48] Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19:349−57.

[49] Maticic K, Krnic Martinic M, Puljak L. Assessment of reporting quality of abstracts of systematic reviews with meta-analysis using PRISMA-A and discordance in assessments between raters without prior experience. BMC Med Res Methodol 2019;19:32.

[50] Jia PL, Xu B, Cheng JM, Huang XH, Kwong JSW, Liu Y, et al. Assessment of the abstract reporting of systematic reviews of dose-response meta-analysis: a literature survey. BMC Med Res Methodol 2019;19:148.

[51] Sufriyana H, Husnayain A, Chen YL, Kuo CY, Singh O, Yeh TY, et al. Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: systematic review and meta-analysis. JMIR Med Inform 2020;8(11):e16503.

[52] Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ 2019;367:l5358.

[53] Oswald NK, Halle-Smith J, Mehdi R, Nightingale P, Naidu B, Turner AM. Predicting postoperative lung function following lung cancer resection: a systematic review and meta-analysis. EClinicalMedicine 2019;15:7−13.

[54] Austin PC, Escobar M, Kopec JA. The use of the Tobit model for analyzing measures of health status. Qual Life Res 2000;9:901−10.