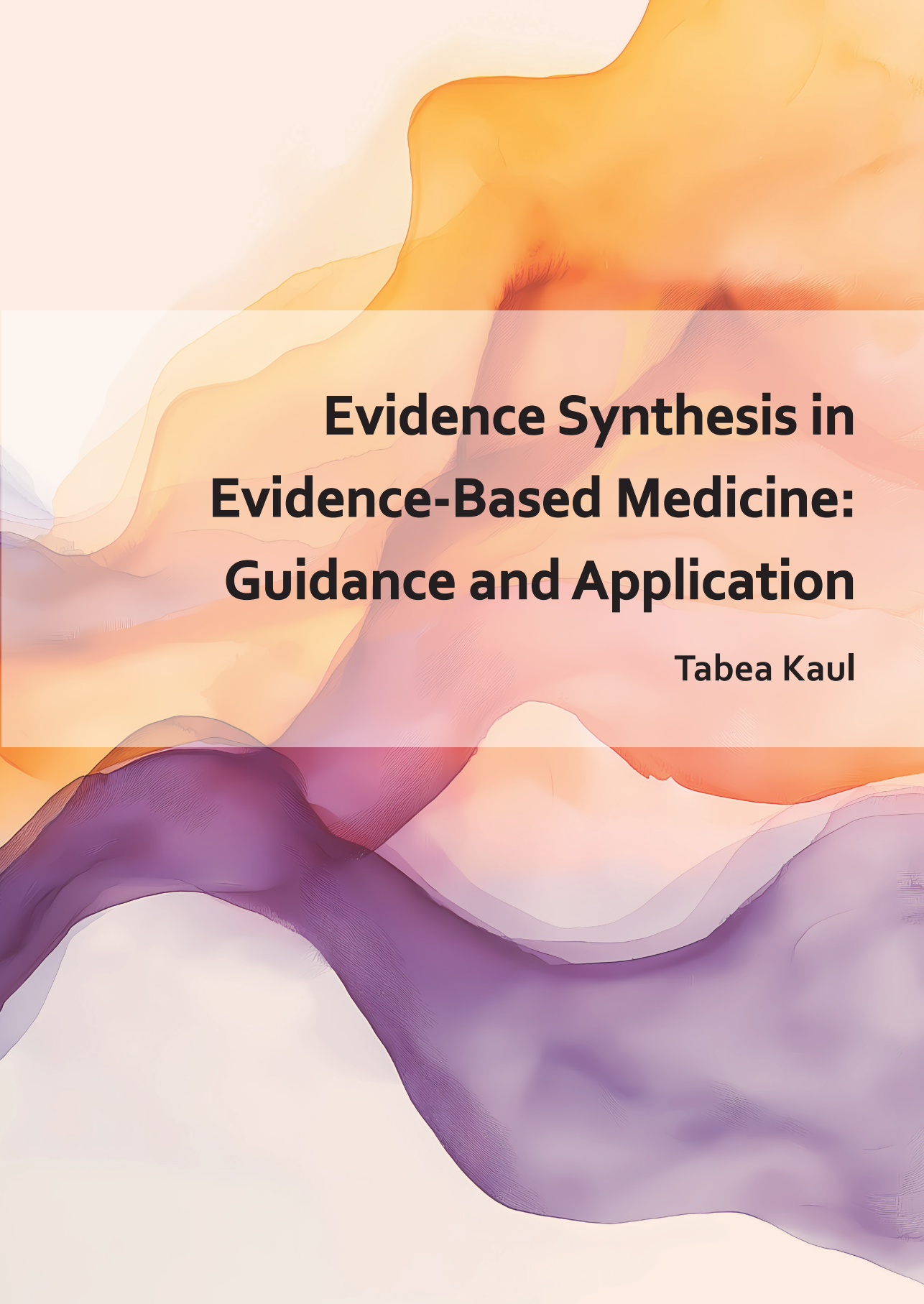


Utrecht University Repository

Title	Evidence Synthesis in Evidence-Based Medicine: Guidance and Application
Authors	Kaul, Tabea
Publication Date	2025-08-28
Link	https://dspace.library.uu.nl/handle/1874/461900
Citation	Kaul, T 2025, 'Evidence Synthesis in Evidence-Based Medicine: Guidance and Application', UMC Utrecht, Utrecht. https://doi.org/10.33540/3059
Versions / License	Publisher version
Rights	https://www.uu.nl/en/university-library/license-and-reuse-conditions
Universitair Medisch Centrum Utrecht (UMCU)	Yes



Evidence Synthesis in Evidence-Based Medicine: Guidance and Application

Tabea Kaul

Evidence Synthesis in Evidence-Based Medicine: Guidance and Application

Tabea Kaul

Copyright 2025 © Tabea Kaul

ISBN: 978-94-6522-259-2

DOI: <https://doi.org/10.33540/3059>

All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

Printing: Ridderprint, ridderprint.nl

Layout and design: Britt de Kroon, persoonlijkproefschrift.nl

Evidence Synthesis in Evidence-Based Medicine: Guidance and Application

Synthese van bewijsmateriaal in Evidence-Based Medicine: Richtlijnen en toepassing

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. ir. W. Hazeleger,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

donderdag 28 augustus 2025 des middags te 12.15 uur

door

Tabea Kaul

geboren op 23 augustus 1993
te Arnstadt, Duitsland

Promotoren:

Prof. dr. K.G.M. Moons

Prof. dr. L. Hooft

Copromotor:

Dr. J.A.A.G. Damen

Beoordelingscommissie:

Prof. dr. mr. W.J.J. Assendelft

Prof. dr. M.L. Bots

Prof. L.C.M. Kremer (voorzitter)

Prof. dr. A.M. May

Prof. dr. E.W. Steyerberg

Für Mama

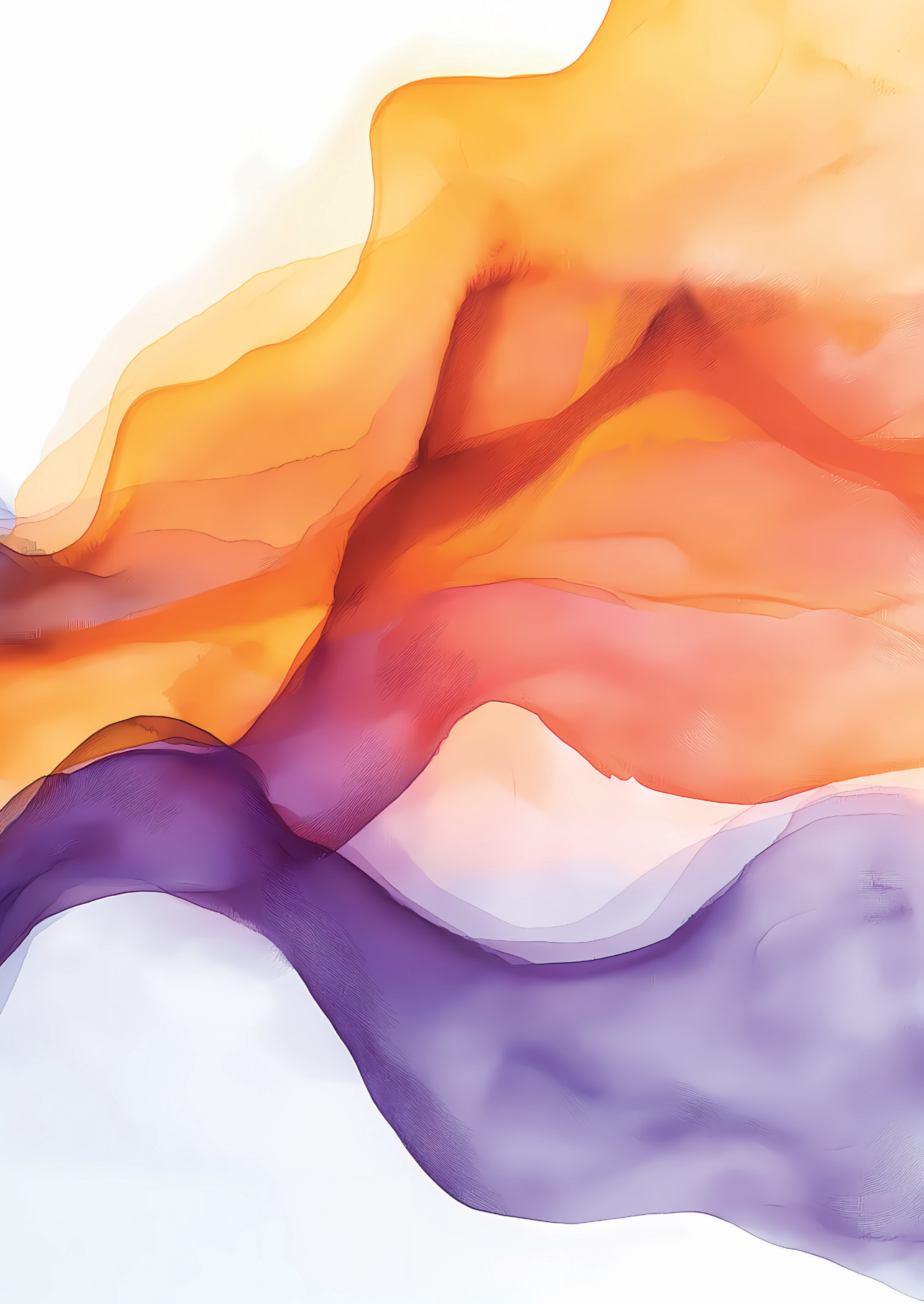
You can't stop the waves, but you can learn to surf.

- Jon Kabat-Zinn

-

TABLE OF CONTENTS

Chapter 1	General Introduction	9
Chapter 2	Both Clinical Trial Register and Electronic Bibliographic Database Searches Were Needed to Identify RCTs for Systematic Reviews: An Evaluation Study	23
Chapter 3	Improving the Reporting and Use of Trial Results in Clinical Trials Registries: Global Practices, Barriers, and Recommendations	45
Chapter 4	Development of Prognostic Models for In-Hospital Mortality and ICU Admission in Confirmed COVID-19 Patients: An IPD Meta-Analysis from LMICs	71
Chapter 5	Methodological Quality Assessment Tools for Diagnosis and Prognosis Research: Overview and Guidance	95
Chapter 6	Assessing the Quality of Prediction Models in Healthcare Using the Prediction model Risk Of Bias ASsessment Tool (PROBAST): An Evaluation of its Use and Practical Application	117
Chapter 7	PROBAST+AI: An Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods	145
Chapter 8	Adherence to the TRIPOD+AI Guideline: An Updated Reporting Assessment Tool	183
Chapter 9	Conclusions and General Discussion	199
Appendices		211
	English Summary	212
	Nederlandse Samenvatting	215
	Deutsche Zusammenfassung	219
	List of Manuscripts	224
	Author Affiliations	226
	PhD Portfolio	234
	Curriculum Vitae	236
	Acknowledgements	237



1

General Introduction

To start with a personal note:

Due to family history of breast cancer, my mother received her first mammogram at the age of 40. In 2021, after a two-year delay caused by the COVID-19 pandemic, her mammogram showed abnormalities. A vacuum biopsy was recommended, and it revealed pre-cancerous cells.

My mother's doctor explained that the identified cells they discovered are known to quickly become invasive and recommended a total mastectomy, though chemotherapy would not be necessary. While relieved to avoid chemotherapy, my mother struggled with the idea of losing her breasts. Determined to explore alternatives, she researched clinical guidelines, talked to experts in the field, and discovered a novel breast-conserving procedure, which, though less commonly practiced, aligned better with her personal preferences.

At her next appointment, my mother proposed this alternative. Initially hesitant, her doctor eventually agreed after reviewing the evidence she presented from guidelines and recent research results. While the doctor only agreed after reviewing the evidence in detail, considering her individual risk of disease progression, the doctor's clinical expertise, and my mother's preference, the doctor finally agreed to proceed with the proposed procedure. The shared decision was made under the condition that my mother shows up frequently for check-ups. This process exemplified the principles of evidence-based medicine: combining the best available scientific evidence, clinical expertise, and patient preferences to make evidence-informed healthcare decisions.

Thanks to this collaborative and evidence-informed approach, my mother received the care she felt was right for her. To this day, she remains in good health.

What is evidence-based medicine?

Evidence-based medicine (EBM), also referred to as evidence-based practice, is a systematic approach to healthcare that integrates the best available evidence from research, clinical expertise, and patient's preference to make more informed health decisions (1-3). The *best available evidence from research* includes high-quality evidence preferably synthesized from multiple studies on a specific topic. *Clinical expertise* includes a healthcare provider's knowledge, skills, and experience gained through their medical training and practice. *Patient preference* relates to the values, needs, and preferences of patients regarding their care options, including personal priorities. The combination of these three elements ensures the most appropriate management plan for the patient at a given time (1, 4).

Types of research in EBM encompasses several key clinical domains (also referred to as the DEPT_H [Diagnosis, Etiology, Prognosis, Therapy] model (5)), including:

- **Diagnosis research:** This domain includes studies that evaluate the reliability, accuracy, and utility of diagnostic tests, tools or models in identifying, classifying or predicting the presence or absence of diseases or conditions.
- **Etiology research:** This domain focuses on understanding the causes or risk factors for diseases and explores the relationships between exposures (e.g., smoking, environmental factors) and health outcomes.
- **Prognosis research:** This domain includes studies that investigate performance of prognostic factors or models in predicting the future occurrence of a certain health outcome, e.g., forecasting the future progression of a condition).
- **Therapy/Intervention research:** This domain evaluates the effects of therapeutic or preventive interventions, such as treatments, medications, or lifestyle changes on health outcomes.

Information of these domains collectively ensures that EBM integrates (scientific) evidence into clinical decision-making to optimise patient care and health outcomes.

Clinical guidelines, which support decisions related to the key clinical domains mentioned above, are informed by synthesis of scientific evidence (6). However, the rapid growth of scientific literature (with thousands of papers published every day) necessitates systematic methods to identify and synthesize only the most relevant and reliable evidence (6-8). This is where systematic reviews play a pivotal role.

What are systematic reviews?

Systematic reviews provide comprehensive and up-to-date information on a specific topic (8, 9). By identifying, summarizing, and critically appraising studies, they ensure that decisions affecting people's lives are informed by a solid body of evidence (8, 9). To ensure high-quality evidence syntheses, it is pivotal that systematic reviews are performed by a research team with relevant expertise, are carried out with careful and transparent pre-planning (e.g., by registering protocols on open access repositories such as PROSPERO), adhere to robust methodological guidelines (i.e., by using the Cochrane Handbooks (10, 11)), and follow reporting standards (e.g., Preferred Reporting Items for Systematic reviews and Meta-Analyses [PRISMA] or Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses [TRIPOD-SRMA]) (12-14). Systematic reviews typically follow several steps (15):

1. Formulating a research question;
2. Identifying and selecting relevant studies;
3. Performing data extraction;
4. Assessing the methodological quality of studies;
5. Summarizing findings;
6. Interpreting and reporting the findings.

Each of these steps ensures that systematic reviews produce the highest quality evidence to inform clinical practice.

Systematic reviews share common steps across different domains, such as intervention, diagnosis, prognosis, and etiology research. However, specific methodologies differ (e.g., which risk of bias or reporting tool to apply). To accommodate the specific methodological aspects of the various types of systematic reviews, several methodological best practice guidelines have been developed (e.g., Cochrane Handbook for Systematic Reviews of Intervention (11), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (10), Cochrane Handbook for Systematic Reviews of Prognosis Studies [in preparation, expected 2026](16)).

Systematic reviews and EBM: challenges

Systematic reviews are inherently dependent on the quality and completeness of the underlying body of evidence. This results in certain challenges:

Challenge 1: Unavailable and incomplete evidence

Previous research has shown that only about half of all conducted research is ever published and that most outcomes are not fully reported (17, 18). Incomplete or missing evidence in systematic reviews can increase bias and distort conclusions.

Evidence may be missing due to the non-publication of research, delays in reporting, or the selective dissemination of results. Identifying and addressing missing evidence is essential to minimize its impact. Studies with statistically significant or favourable results are more likely to be published, while studies with unfavourable results are more likely to remain unpublished (i.e., publication bias)(19, 20). In addition, researchers may selectively report only certain outcomes (i.e., selective outcome reporting bias), particularly those that are 'positive' or align with their hypotheses, leaving other data unpublished (21). Such biases can distort the pooled results in systematic reviews, leading to skewed interpretations of the available evidence.

Certain measures can be taken to get a vast idea about whether studies are missing, such as creating a funnel plot to look at the distributions of effect estimates or to search for unpublished trials registered on clinical trials registries to explore whether any trials are missing (22-24). Comprehensive searches for literature in multiple databases reduce the likelihood of missing relevant studies, including those published in less accessible journals, gray literature, and clinical trials registries (23).

Clinical trials registries are databases that allow researchers to register their clinical trials (ideally, prospectively) and provide information on planned, ongoing, completed, or discontinued trials; some allow the reporting of trial results (25-27). By making trial information publicly available (without needing to go through formal journal publication processes) clinical trials registries contribute to reducing research duplication, improve research transparency and trust in scientific findings (25, 28, 29). Further, trial registration helps to expose various biases affecting evidence synthesis, such as selective outcome reporting or publication bias (25, 30, 31). Currently, numerous agreements and policies exist to promote trial registrations, with a primary focus on clinical trials (25, 28, 29, 32-34). This makes clinical trials registries valuable resources to identify trials for systematic reviews.

Challenge 2: Biased evidence

Bias is typically defined as systematic deviation from the truth in research findings and a critical concern in evidence synthesis (35). Bias can compromise the validity of systematic reviews, either through “limitations in the design, conduct or data collection of primary studies or through analysis that may lead to systematic over- or underestimation of study results” (35, 36).

To assess the risk of bias in primary studies, researchers use tailored risk of bias tools with specific assessment criteria for different study designs and research domains, ensuring that their limitations are acknowledged during evidence synthesis (36-39). An example of a widely used risk of bias tool is the Cochrane’s Risk of Bias 2 tool (RoB2) (40), which includes criteria for the assessment of randomised controlled trials. Another commonly used example comprises the Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) (41, 42) which provides criteria to assess primary prediction model development and evaluation studies.

Challenge 3: Evidence synthesis for prediction research still under development

While guidance for systematic reviews in therapeutic or preventive intervention research is widely established (11), it is still developing for prediction (including diagnosis

and prognosis) research. Although extensive guidance on how to perform these evidence syntheses has been described in scientific articles, the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Studies (10) was only published in 2023, while we await the Cochrane Handbook for Systematic Reviews of Prognosis Studies (in preparation, expected in 2026) (16).

Since the early 2010s, key methodological papers for prediction models in healthcare have been published (41-53). The TRIPOD+AI statement (43) highlights essential reporting items for prediction models developed using regression or machine learning methods, while the Prediction model Risk Of Bias ASsessment Tool (PROBAST) (41, 42) aids researchers in assessing their risk of bias and applicability. PROBAST may also be used to guide the design of diagnostic and prognostic prediction model studies. Nevertheless, advances in machine learning methods are further expanding and challenging evidence synthesis in prediction research, requiring updates of the currently available guidance (54-56).

Well-known clinical prediction models such as the Wells' score for the diagnosis of deep vein thrombosis (DVT) (57), or PREDICT, which estimates how treatment after surgery in women with early-stage invasive breast cancer might improve 15-year survival probability (58-60), compete against a substantial number of developed prediction models (61-64). This complicates the selection of a model for external validation and integration into the clinical care pathway and exemplifies the importance of evidence synthesis in supporting clinical decision-making. As the field advances, systematic reviews in prediction research will need to adapt to ensure rigor and relevance.

OUTLINE OF THIS THESIS

In this thesis, we critically examine various aspects of evidence synthesis in evidence-based medicine with the aim to provide more reliable synthesis methods. We specifically focused on aspects related to methodological guidance and its application aiming to provide the best scientific evidence in medicine.

Above, we provided the typical steps taken for systematic reviews. **Chapter 2** and **Chapter 3** of this thesis are about evidence synthesis of intervention research, while the second part, following **Chapter 4**, focusses on methods used for, e.g., evidence synthesis of prediction research.

For systematic reviews of intervention, which include RCTs, the Cochrane Handbook of Systematic Reviews of Interventions (11) recommends searching in both medical journal databases (such as EMBASE) and clinical trials registries. A dual search

approach minimizes the risk of missing any RCTs. However, is this necessary? For RCTs, agreements and policies are in place that require registering an RCT before the trial is conducted. Therefore, searching both databases might be a duplicate of effort. In **Chapter 2**, we compared the different search approaches that can be taken to identify RCTs for systematic reviews. Here, we compared searching in medical journal databases with identifying trials in clinical trials registries. In **Chapter 3**, we performed a comprehensive overview of reporting practices of trial results across various clinical trials registries and surveyed registrants (i.e., people registering the trial results in registries) and evidence users (i.e., people using results from registries for secondary purpose, such as systematic reviews) to identify how reporting and therefore usability of trial results available in clinical trials registries could be improved.

In **Chapter 4**, we developed two clinical prediction models to calculate the individual risk of patients admitted to the hospital in low- and lower-middle-income countries (LMICs) with confirmed COVID-19 to experience in-hospital mortality or to require admission to the intensive care unit. Currently, no prognostic model performs well in the LMIC setting (65). We pooled data into an individual participant data meta-analysis using data from the World Health Organization Global Clinical Platform database for COVID-19 from over 15,000 patients and fitted the developed models per included country.

Systematic reviews require the methodological quality assessment of a study's validity. In diagnosis and prognosis research, many methodological quality (risk of bias) tools exist. In **Chapter 5**, we provide an overview of methodological quality assessment tools for diagnosis and prognosis research and developed five guiding questions that help users select a tool according to their specific research question and context. In **Chapter 6**, we assess the application of the Prediction model Risk Of Bias ASsessment Tool (PROBAST) and evaluate aspects related to the use of the tool in practices, such as inter-rater reliability, use of consensus meetings, and aspects leading to specific ratings.

Informed by the results of a three-stage Delphi survey procedure involving experts from across the globe and by findings of Chapter 5, we developed an updated version of PROBAST in **Chapter 7**, which now covers aspects of prediction model studies related to regression and artificial intelligence/ machine learning based methods, i.e., PROBAST+AI.

While searching, selecting, and methodological quality assessments are important steps in identifying the best available evidence for systematic reviews, the evidence cannot be used without proper reporting. Several reporting guidelines have been developed to support complete reporting of healthcare studies. Completeness of

reporting can be measured by assessing the adherence to the reporting guidelines. However, assessments lack uniformity. In 2019, a tool to measure the adherence to the TRIPOD statement (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) was developed to enable standardized adherence assessments. With recent advances in artificial intelligence/machine learning based methods, the TRIPOD statement was replaced by TRIPOD+AI. Therefore, an update of the adherence tool became necessary, which we provide in **Chapter 8**.

In **Chapter 9**, we provide the conclusions of this thesis and end with a general discussion.

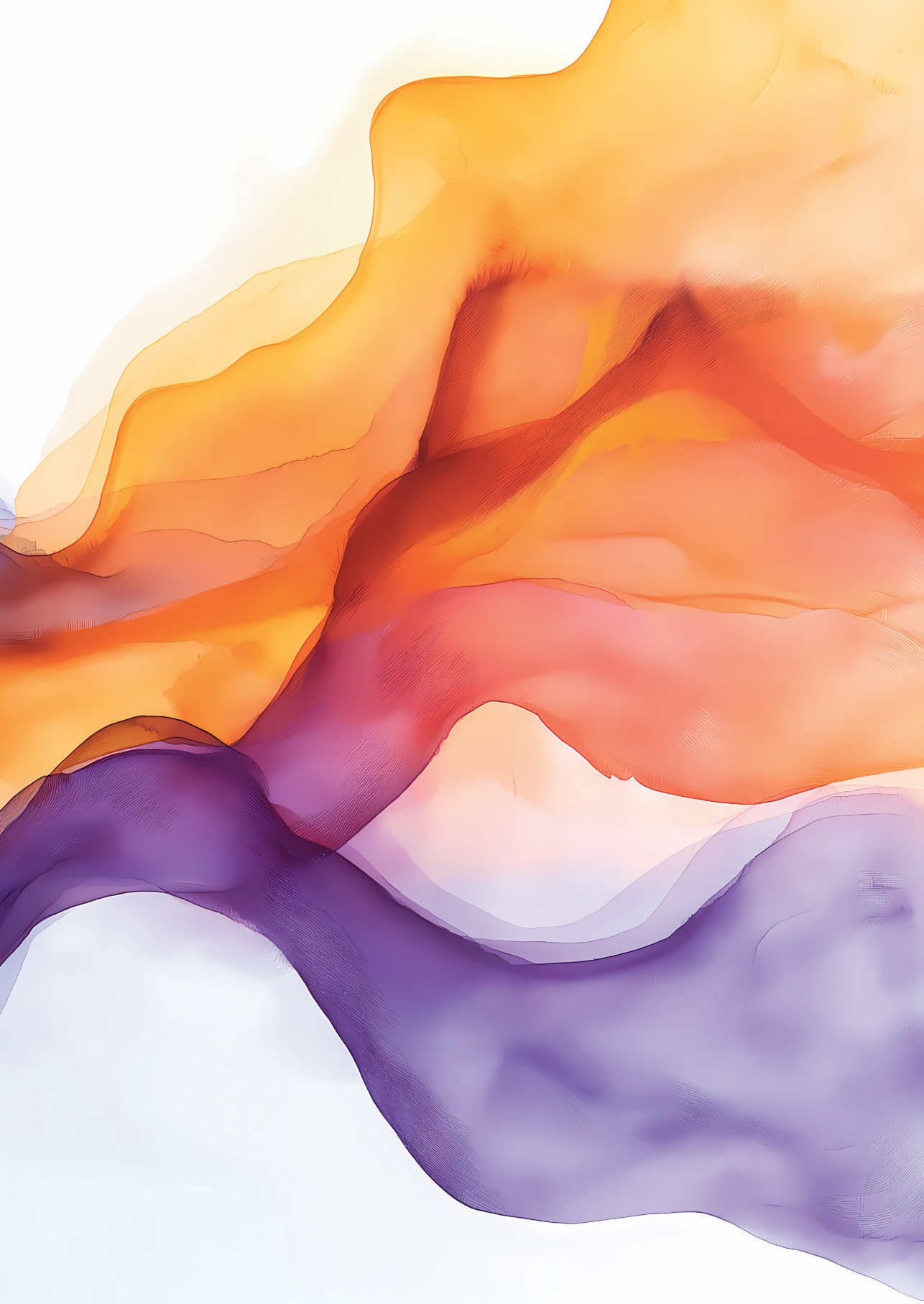
REFERENCES

- (1) Guyatt G, Cairns J, Churchill D, Cook D, Haynes B, Hirsh J, et al. Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine. *JAMA*. 1992;268(17):2420-5.
- (2) Guyatt G, Jaeschke R, Wilson MC, Montori VM, Richardson WS. What Is Evidence-Based Medicine? In: Guyatt G, Rennie D, Meade MO, Cook DJ, editors. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*, 3rd ed. New York, NY: McGraw-Hill Education; 2015.
- (3) National Cancer Institute. evidence-based medicine 2024 [cited 2024 12 December]. Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/evidence-based-medicine>.
- (4) Connor L, Dean J, McNett M, Tydings DM, Shrout A, Gorsuch PF, et al. Evidence-based practice improves patient outcomes and healthcare system return on investment: Findings from a scoping review. *Worldviews Evid Based Nurs*. 2023;20(1):6-15.
- (5) Grobbee DE, Hoes AW. *Clinical epidemiology: principles, methods, and applications for clinical research*: Jones & Bartlett Publishers; 2014.
- (6) Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. *Bmj*. 1999;318(7183):593-6.
- (7) Landhuis E. Scientific literature: Information overload. *Nature*. 2016;535(7612):457-8.
- (8) Cochrane. Evidence Synthesis - What is it and why do we need it? online2023 [cited 2024 16 December]. Available from: <https://www.cochrane.org/news/evidence-synthesis-what-it-and-why-do-we-need-it>.
- (9) Cumpston M, Flemyng E, Thomas J, Higgins JP, Deeks JJ, Clarke MJ. Chapter I: introduction. *Cochrane handbook for systematic reviews of interventions*. 2023;6.
- (10) Deeks JJ, Bossuyt PM, Leeftang MM, Takwoingi Y. *Cochrane handbook for systematic reviews of diagnostic test accuracy*: John Wiley & Sons; 2023.
- (11) Chandler J, Cumpston M, Li T, Page MJ, Welch V. *Cochrane handbook for systematic reviews of interventions*. Hoboken: Wiley. 2019.
- (12) Cumpston M, Chandler J. Chapter II: planning a Cochrane review. *Cochrane handbook for systematic reviews of interventions* Cochrane. 2022.
- (13) Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*. 2010;8(5):336-41.
- (14) Snell KIE, Levis B, Damen JAA, Dhiman P, Debray TPA, Hooft L, et al. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). *BMJ*. 2023;381:e073538.
- (15) Khan KS, Kunz R, Kleijnen J, Antes G. Five steps to conducting a systematic review. *J R Soc Med*. 2003;96(3):118-21.
- (16) Moons KGM, Hooft L, Williams K, Hayden JA, Damen J, Riley RD. Implementing systematic reviews of prognosis studies in Cochrane. *Cochrane Database of Systematic Reviews*. 2018(10).
- (17) Schmucker C, Schell LK, Portalupi S, Oeller P, Cabrera L, Bassler D, et al. Extent of Non-Publication in Cohorts of Studies Approved by Research Ethics Committees or Included in Trial Registries. *PLOS ONE*. 2014;9(12):e114023.
- (18) Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *Bmj*. 2010;340:c365.

- (19) Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*. 2009(1).
- (20) Showell MG, Cole S, Clarke MJ, DeVito NJ, Farquhar C, Jordan V. Time to publication for results of clinical trials. *Cochrane Database of Systematic Reviews*. 2024(11).
- (21) Thomas E, Heneghan C. Outcome reporting bias 2017 [cited 2025 20 February]. Available from: <https://catalogofbias.org/biases/outcome-reporting-bias/#cite>.
- (22) Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J Clin Epidemiol*. 2005;58(9):894-901.
- (23) Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Searching for and selecting studies. *Cochrane Handbook for systematic reviews of interventions*. 2019:67-107.
- (24) Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. *Journal of Clinical Epidemiology*. 2011;64(12):1277-82.
- (25) Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*. 2004;351(12):1250-1.
- (26) World Health Organization. WHO Registry criteria: Registry criteria (Version 2.1, April 2009), 2009 [cited 2024 November 14]. Available from: <https://www.who.int/clinical-trials-registry-platform/network/registry-criteria>.
- (27) World Health Organization. WHO Trial Registration Data Set (Version 1.3.1) 2024 [cited 2024 12 March]. Available from: <https://www.who.int/clinical-trials-registry-platform/network/who-data-set>.
- (28) International Clinical Trial Registry Plattform. Joint statement on public disclosure of results from clinical trials, : World Health Organization; 2017 [cited 2025 20 February]. Available from: <https://www.who.int/news/item/18-05-2017-joint-statement-on-registration>.
- (29) Strengthening clinical trials to provide high-quality evidence on health interventions and to improve research quality and coordination, WHA 75.8. Sect. 75.8 (2022).
- (30) Simes R. Publication bias: the case for an international registry of clinical trials. *Journal of Clinical Oncology*. 1986;4(10):1529-41.
- (31) Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA*. 2009;302(9):977-84.
- (32) U.S. Government. Food and Drug Administration Amendments Act of 2007: US Government Printing Office; 2007 [Available from: <https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf>].
- (33) Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov — The Final Rule. *New England Journal of Medicine*. 2016;375(20):1998-2004.
- (34) European Parliament and the Council of the European Union. Regulation (EU) No 536/2014 of the European Parliament and of the council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC, [Regulation]. *Official Journal of the European Union*; 2014 [updated 27 May 2014; cited 2024 5 April]. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0536>.
- (35) Boutron IP, MJ; Higgins, JPT; Altman, DG; Lundh, A; Hróbjartsson, A. Chapter 7: Considering bias and conflicts of interest among the included studies. 2024 last updated August 2022. In: *Cochrane Handbook for Systematic Reviews of Interventions* version 65 [Internet]. Cochrane. Available from: www.training.cochrane.org/handbook.
- (36) Kaul T, Kellerhuis BE, Damen JAA, Schuit E, Jenniskens K, van Smeden M, et al. Methodological quality assessment tools for diagnosis and prognosis research: overview and guidance. *Journal of Clinical Epidemiology*. 2025;177:111609.

- (37) Ma L-L, Wang Y-Y, Yang Z-H, Huang D, Weng H, Zeng X-T. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military Medical Research*. 2020;7(1):7.
- (38) LATITUDES network. Validity assessment tools for evidence synthesis: your one-stop-shop 2025 [cited 2025 5 January]. Available from: <https://www.latitudes-network.org/>.
- (39) Higgins JP, Savović J, Page MJ, Elbers RG, Sterne JA. Assessing risk of bias in a randomized trial. *Cochrane handbook for systematic reviews of interventions*. 2019:205-28.
- (40) 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.
- (41) 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(1):W1-w33.
- (42) 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(1):51-8.
- (43) Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
- (44) 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.
- (45) 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.
- (46) 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.
- (47) 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.
- (48) Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
- (49) Collins GS, Dhiman P, Ma J, Schlusser MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. 2024;384:e074819.
- (50) Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820.
- (51) Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *Bmj*. 2024;384:e074821.
- (52) Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*. 2015;13(1):1.
- (53) 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(1):W1-73.
- (54) Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Systematic review finds “spin” practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol*. 2023;158:99-110.
- (55) Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J Clin Epidemiol*. 2023;154:8-22.

- (56) Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
- (57) Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, et al. Accuracy of clinical assessment of deep-vein thrombosis. *Lancet*. 1995;345(8961):1326-30.
- (58) Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res*. 2010;12(1):R1.
- (59) Grootes I, Wishart GC, Pharoah PDP. An updated PREDICT breast cancer prognostic model including the benefits and harms of radiotherapy. *npj Breast Cancer*. 2024;10(1):6.
- (60) Winton Centre for Risk and Evidence Communication UoC. Overview 2024 [cited 2025 12 January]. Available from: <https://breast.predict.cam/about/overview/overview>.
- (61) Kafeza M, Shalhoub J, Salooja N, Bingham L, Spagou K, Davies AH. A systematic review of clinical prediction scores for deep vein thrombosis. *Phlebology*. 2016;32(8):516-31.
- (62) Engelhardt EG, Garvelink MM, de Haes JH, van der Hoeven JJ, Smets EM, Pieterse AH, et al. Predicting and communicating the risk of recurrence and death in women with early-stage breast cancer: a systematic review of risk prediction models. *J Clin Oncol*. 2014;32(3):238-50.
- (63) Altman DG. Prognostic Models: A Methodological Framework and Review of Models for Breast Cancer. *Cancer Investigation*. 2009;27(3):235-43.
- (64) Shen J-H, Chen H-L, Chen J-R, Xing J-L, Gu P, Zhu B-F. Comparison of the Wells score with the revised Geneva score for assessing suspected pulmonary embolism: a systematic review and meta-analysis. *Journal of Thrombosis and Thrombolysis*. 2016;41(3):482-92.
- (65) Damen JAA, Arshi B, van Smeden M, Bertagnolio S, Diaz JV, Silva R, et al. Validation of prognostic models predicting mortality or ICU admission in patients with COVID-19 in low- and middle-income countries: a global individual participant data meta-analysis. *Diagn Progn Res*. 2024;8(1):17.



2

Both Clinical Trial Register and Electronic Bibliographic Database Searches Were Needed to Identify RCTs for Systematic Reviews: An Evaluation Study

Tabea Kaul, Julia MT Colombijn, Robin WM Vernooij, Rene Spijker, Demy L Idema,
Linde F Huis in 't Veld, Johanna AA Damen, Lotty Hooft

J Clin Epidemiol. 2024 May;169:111300

ABSTRACT

Objective: To determine whether clinical trial register (CTR) searches can accurately identify a greater number of completed randomized clinical trials (RCTs) than electronic bibliographic database (EBD) searches for systematic reviews of interventions, and to quantify the number of eligible ongoing trials.

Study Design and Setting: We performed an evaluation study and based our search for RCTs on the eligibility criteria of a systematic review that focused on the underrepresentation of people with chronic kidney disease in cardiovascular RCTs. We conducted a combined search of ClinicalTrials.gov and the WHO International Clinical Trials Registry Platform (ICTRP) through the Cochrane Central Register of Controlled Trials (CENTRAL) to identify eligible RCTs registered up to June 1st, 2023. We searched Cochrane CENTRAL, EMBASE, and MEDLINE for publications of eligible RCTs published up to June 5th, 2023. Finally, we compared the search results to determine the extent to which the two sources identified the same RCTs.

Results: We included 92 completed RCTs. Of these, 81 had results available. 66 completed RCTs with available results were identified by both sources (81% agreement (95% CI 71 – 88)). We identified seven completed RCTs with results exclusively by CTR search (9% (95% CI 4 – 17)) and eight exclusively by EBD search (10% (95% CI 5 – 18)). 11 RCTs were completed but lacked results (four identified by both sources (36% (95% CI 15 – 65)), one exclusively by EBD search (9% (95% CI 1 – 38)), and six exclusively by CTR search (55% (95% CI 28 – 79)). Also, we identified 42 eligible ongoing RCTs: 16 by both sources (38% (95% CI 25 – 53)) and 26 exclusively by CTR search (62% (95% CI 47 – 75)). Lastly, we identified four RCTs of unknown status by both sources.

Conclusion: CTR searches identify a greater number of completed RCTs than EBD searches. Both searches missed some included RCTs. Based on our case study, researchers (e.g., information specialists, systematic reviewers) aiming to identify all available RCTs should continue to search both sources. Once the barriers to performing CTR searches alone are targeted, CTR searches may be a suitable alternative.

INTRODUCTION

Randomized clinical trials (RCTs) are pivotal in providing robust evidence on treatment effectiveness (1). RCTs inform interventional systematic reviews and meta-analysis and are key to clinical guideline development. To identify RCTs, methodological guidelines for systematic reviews (i.e. Chapter 4: Searching for and selecting studies, Cochrane Handbook for Systematic Reviews of Interventions) recommend combining electronic bibliographic database (EBD) searches with clinical trial register (CTR) searches (2). Searching both sources minimizes the risk of missing relevant RCTs: whereas searches for RCTs in EBDs are limited to the identification of published articles, CTRs allow the identification of RCTs regardless of their publication status (3-7).

RCTs may be missed for various reasons, such as incorrect indexing in search databases, publication in journals not included in the systematic search, or not being published at all (8-10). Consequently, the effect estimates identified by the systematic review may be biased (8-12). Another issue contributing to biased effect estimates in systematic reviews is publication bias: the selective publication of studies due to the strength or direction of the results (5, 8, 13-17). This might result in distorted effect sizes and incorrect conclusions in systematic reviews (5, 11, 12).

Exclusively searching CTRs for RCTs may be a suitable alternative to the current recommendation. Clinical trial (CT) registration is typically performed a priori, enabling CTR searches to identify both completed and ongoing trials that lack a published article, and which might have otherwise been missed with an EBD search alone (3, 7, 8). Consequently, CTR searches help to assess and potentially minimize the impact of publication bias on systematic review findings. Furthermore, while completed RCTs directly inform systematic reviews, the identification of ongoing trials serves as a useful indicator for determining the timing of systematic review updates (7).

Inconsistent CT registration and inadequate results reporting (i.e., no link to published article or posting in CT record) can hinder exclusively performing CTR searches for systematic reviews (18-20). A study from 2014 that compared systematic review search strategies reported a low yield of identified RCTs from CTR searches and inadequate adherence to CT registration, while a review from 2018 on RCTs of newly developed drugs suggested that the completeness of CT registers improved in recent years (18, 19). This development leads us to reevaluate current search approaches for identifying RCTs for systematic reviews of interventions.

In this study, we aimed to determine whether CTR searches can accurately identify a greater number of completed RCTs than EBD searches for systematic reviews of

interventions, and to quantify the number of eligible ongoing trials. Based on a case study of cardiovascular RCTs, we conducted similar searches, adjusting them accordingly to the specific source, in both CTRs and EBDs, and compared the search results.

METHODS

Study design

We performed an evaluation study based on a case study to assess the accuracy of CTR and EBD searches in identifying RCTs for systematic reviews of interventions. We registered our study protocol on the Open Science Framework (OSF) (21).

Case study

This is a sub-study of a systematic review that assessed the underrepresentation of people with chronic kidney disease (CKD) in cardiovascular RCTs (22). The systematic review searched for eligible RCTs in ClinicalTrials.gov through the Cochrane Central Register of Controlled Trials (CENTRAL) with a combination of keywords for cardiovascular diseases (CVD), cardiovascular risk factors, and the interventions of interest (23-25). The systematic review’s aim and protocol have been reported previously (22). The protocol is registered in the International Prospective Register for Systematic Reviews (PROSPERO CRD42022296746).

Eligibility criteria

The eligibility criteria of our study were identical to the case study (22). Due to the proof-of-concept character of our study, we limited our search to RCTs evaluating direct anticoagulants (DOACs). Table 1 shows the PICOS of our study.

Table 1 PICOS

Elements	Description	
Population	Adults with a history of CVD (22): • Coronary artery disease • Atrial fibrillation • Congestive heart failure • Peripheral arterial disease • Stroke	Adults with ≥ 1 cardiovascular risk factor (26-28): • Overweight and obesity • Hypertension • Hyperglycemia and diabetes mellitus • Chronic- or end-stage kidney disease

Table 1 PICOS (*continued*).

Elements	Description	
Intervention*	Direct factor Xa inhibitors (B01AF): • Apixaban (B01AF02) • Betrixaban (B01AF04) • Edoxaban (B01AF03) • Rivaroxaban (B01AF01)	Direct thrombin inhibitors (B01AE): - Argatroban (B01AE03) - Dabigatran (B01AE07)
Comparator	• No treatment • Placebo • Standard of Care	• Another treatment • Different dosage
Outcome	• All-cause mortality • Cardiovascular mortality • Major cardiovascular event (MACE)/another composite cardiovascular endpoint • Coronary artery disease	• Cerebrovascular disease • Peripheral arterial disease • Hospitalization for heart failure • Development of heart failure stage III or IV (New York Heart Association classification) • End-stage kidney disease
Study type	RCTs with at least 100 participants	

*The interventions are listed with their ATC code in brackets.

Information sources

We based our search on the databases recommended by ‘Chapter 4: Searching for and selecting studies’ of the Cochrane Handbook of Systematic Reviews of Interventions (7). To identify trial registration records of eligible RCTs, we performed a search for ClinicalTrials.gov and International Clinical Trials Registry Platform (ICTRP) records within the Cochrane Central Register of Controlled Trials (CENTRAL) as part of the Cochrane library (<https://www.cochranelibrary.com/>) on June 1st, 2023 (24, 25, 29). We obtained the published reports from the completed CT records by following the provided link or by manual search. For the EBD search, we searched CENTRAL (<https://www.cochranelibrary.com/>), EMBASE (embase.com 1947-), and MEDLINE (Ovid MEDLINE(R) ALL 1946) on June 5th, 2023, for published articles of RCTs either describing a study protocol or reporting the main results of the RCT (23, 30, 31). No limitations on publication year were applied to either search (Supplement, Section A and B).

Screening and data collection

Screening

Clinical trial register search

Two reviewers (TK, DI) screened the CT records in duplicate. The same two reviewers performed full-text screening to determine whether a trial met the inclusion criteria.

We based our decision on the CT record only in case no published manuscript was available. Any disagreements were resolved through discussion to reach consensus or by involving a third reviewer.

Electronic bibliographic database search

One reviewer (TK) performed title-abstract screening while a second reviewer (DI) screened a 20% random sample. Any disagreements were resolved through discussion to reach consensus or by involving a third reviewer. We regarded an interrater reliability of 0.7 of Cohen's κ as sufficient for agreement between the two reviewers (32, 33). Prior to full-text screening, we held a calibration session. In this session, the reviewers screened 20 full-texts and discussed whether to include a published article. Then, one reviewer (TK, DI, or LHV) proceeded independently with the remaining set for full-text screening.

Data collection

For both approaches, one reviewer extracted the data (TK, DI, or LHV), while a second (TK, DI, or LHV) verified them. Collected items included clinical trial registration number, name of clinical trial registry, study acronym, first author, last author, contact details, study year, study location/s, funding, and intervention (Supplement, Section C).

Comparison of the search strategies

Matching procedure

To compare the accuracy of the search strategies, we retrieved the CT registration numbers (if any) from the published articles retrieved by EBD search, and combined them with the records retrieved by CTR search in one spreadsheet in Microsoft Excel (Version 16.68) (34). Then, we compared the identified RCTs by CT registration number, study acronym, study location, and intervention.

Matched RCTs

We considered an RCT as 'identified by both sources' if the same CT registration number was present in both sets or if the study acronym and/or study location and/or intervention presented itself as identical (Supplement, Section D).

*Unmatched RCTs*RCTs identified by CTR search only

For unmatched RCTs identified exclusively by CTR search, we first checked whether the CT record referred to a published article. If not, we manually searched CENTRAL, EMBASE, and MEDLINE using the CT registration number and/or the study acronym and/or first and last author and/or study location and/or intervention. If we identified a published article, we classified the RCT as ‘only identified by CTR search and published in indexed journals’ (falsely missed by EBD search) (Supplement, Section D). Additionally, we checked (i.e., by Google search) whether an article had been published in a journal not indexed by CENTRAL, EMBASE, or MEDLINE (not indexed in searched EBDs). Otherwise, we considered the RCT as ‘only available in CT record’ (Supplement, Section D).

RCTs identified by EBD search only

For unmatched RCTs identified exclusively by EBD search, we manually searched ClinicalTrials.gov and the ICTRP using the CT registration number (if provided) and/or the study acronym and/or the study location and/or the intervention. If we identified a CT registration, we considered the RCT as ‘identified by EBD search only and registered’ (falsely missed by CTR search) (Supplement, Section D). In case we failed to identify a CT registration by manual search, we contacted study authors for more information. If we did not receive a reply to our enquiries, we considered the RCT as ‘non-registered’ (not included in CTRs) (Supplement, Section D).

Classification of identified RCTs

After the matching, we classified RCTs according to their completion status and their results availability as either ‘completed with(out) results’ or ‘ongoing’, based on the (planned) study completion date available in the CT record (Supplement, Section D). We classified RCTs with a study completion date before January 1st, 2022, as ‘completed with results’, if any results were available. In case no results were available, we contacted study authors for more information. In case we did not receive a reply to our enquiries, we considered the RCT as ‘completed without results’ (protocol only) (Supplement, Section D). Ongoing RCTs had a planned study completion recorded in their CT record after December 31st, 2021. We chose December 31st, 2021, as a cut-off as we anticipated that the main results would be published within 12 months after study completion (35-37). In case we identified RCTs of unknown completion status, we contacted study authors for more information on the respective RCTs. In case we did not receive a

response to our enquiries, we categorized these RCTs as ‘RCTs of unknown status’ (Supplement, Section D).

Statistical analysis

For our analysis, we focused on completed RCTs, since we expected these to be available in both sources. We treated ongoing RCTs as a separate outcome because we assumed that they would be available primarily in CTRs. RCTs of unknown status were reported as exploratory outcomes (Supplement, Section D).

We calculated counts and percentages or median and range to describe the characteristics of the included RCTs. We calculated counts, percentages, and the corresponding 95%-confidence interval for the outcomes using the Wilson score interval (38). We performed all statistical analyses in RStudio Version 2022.07.2 (39).

RESULTS

Study flow and sample characteristics

The EBD search identified a total of 9,395 records for screening. We excluded 8,648 of these records during title-abstract screening as they did not meet the eligibility criteria. We excluded 396 records of these because they were conference abstracts or secondary analyses. We screened 351 reports on full-text for eligibility and finally included 97 RCTs by EBD search (Figure 1). The CTR search identified a total of 523 records. We excluded 214 of these records during record screening, assessed 309 reports for eligibility and finally included 129 eligible RCTs (Figure 2).

After the comparison (Section 2.6 Comparison of the search strategies) we identified a total of 138 unique RCTs from both sources. Of these, 92 were completed (Table 2). 80 completed RCTs had results available, while 12 did not. We included 42 ongoing RCTs and four of unknown study status (Table 3&4). Most RCTs were from Asia (38%) and funded by industry (53%) (Table 5). The median year of publication for published articles (n=101) was 2018 (range: 2007– 2023). The median (planned) completion year of RCTs without published articles as listed in the CT record (n=21), was 2021 (IQR: 2015 – 2026) (Table 5).

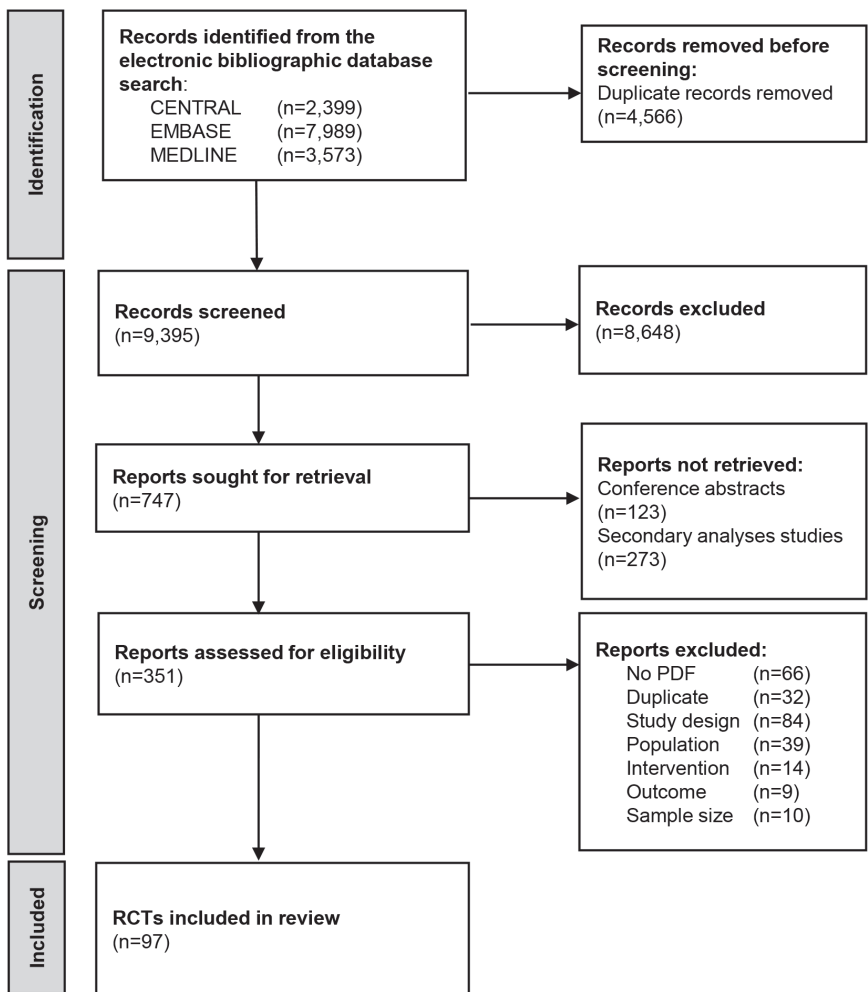


Figure 1 Flow diagram of the electronic bibliographic database search

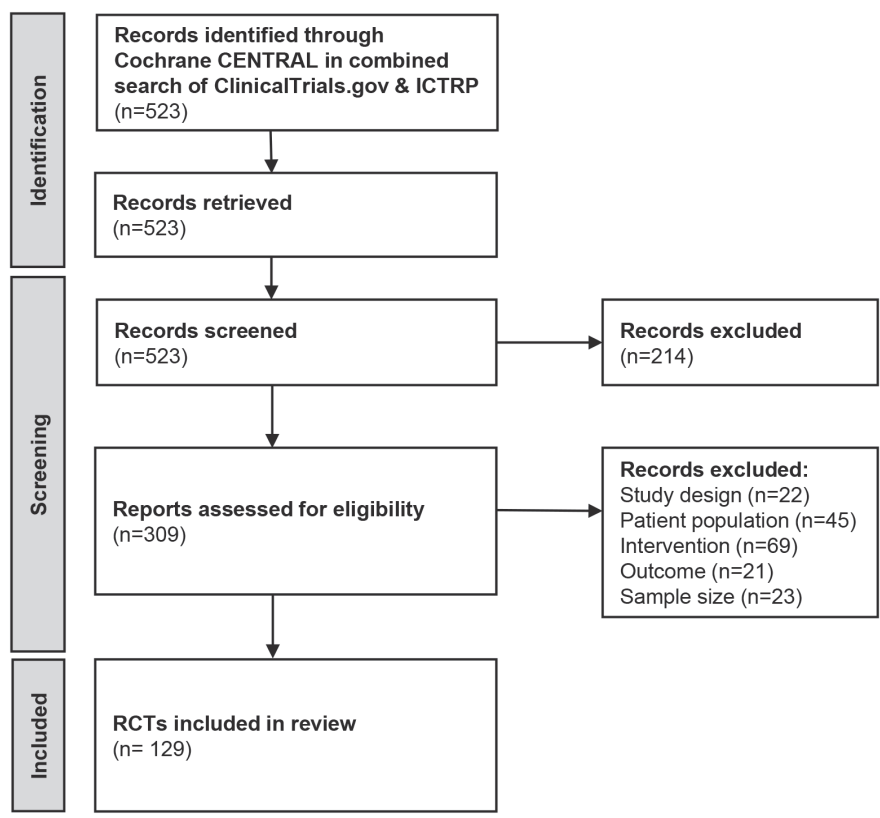


Figure 2 Flow diagram of the clinical trial register search

Comparison of the search strategies

Completed RCTs with results

We included a total of 81 completed RCTs with results. Out of these, we identified 66 RCTs in both sources (representing 81% agreement (95% CI 71 - 88)) (Table 3). We also identified eight eligible RCTs by EBD search only, accounting for 10% of the total ((95% CI 5 - 18)) (Table 2). Among these, three RCTs (4% (95% CI 1 - 10)) had a CT registration while five did not, (6% (95% CI 3 - 14) (Table 2). We were unable to obtain information regarding the missing CT registration from the study authors upon request. Two of the ClinicalTrials.gov-registered RCTs were not present in the Cochrane CENTRAL database at the time of searching, while one was available, but our CTR search strategy did not identify it (Section E, Supplement). Moreover, we identified seven eligible RCTs exclusively by CTR search, constituting 9% of the total (95% CI 4 - 17) (Table 3). Among these, one RCTs (1% (95% CI 0 - 7)) had an article published in indexed journals but

was not identified by the EBD search strategy we employed (thus falsely missed by EBD search, Section E, Supplement), zero in non-indexed journals (0% (95% CI 0 – 1)), and six had results only available in their CT record (not available in EBDs) (8% (95% CI 3 – 15)) (Table 2).

Completed RCTs without results

We included a total of 11 completed RCTs without results. Out of these, we identified four RCTs by both sources (representing 36% agreement (95% CI 15 - 65)) (Table 2). We also identified one RCT exclusively by EBD search, accounting for 9% of the total (95% CI 1 – 38) (Table 2). Moreover, we identified six eligible RCTs exclusively by CTR search, constituting 58% of the total (95% CI 28 – 79) (Table 2). Among these, all six only had a CT record available (protocol only, without results) (55% (95% CI 28 – 79)) (Table 2). None of our requests for information on the completed RCTs without results succeeded.

Table 2 Identification source of completed RCTs per results availability

Outcomes	N (%)	(95%-CI)
Completed RCTs		
With results (n=81)		
Identified by both sources*	66 (81)	71 – 88
Identified by EBD search only*	8 (10)	5 – 18
Registered RCT ^a	3 (4)	1 – 10
Non-registered RCT*	5 (6)	3 – 14
Identified by CTR search only*	7 (9)	4 – 17
Results published in indexed journals ^a	1 (1)	0 – 7
Results published in non-indexed journals*	0 (0)	0 – 1
Results available in clinical trial record only*	6 (8)	3 – 15
Without results (n=11)		
Identified by both sources**	4 (36)	15 – 65
Identified by EBD search only ^{a*}	1 (9)	1 – 38
Identified by CTR search only**	6 (55)	28 – 79
Protocol published in indexed journals**	0 (0)	0 – 79
Protocol published in non-indexed journals**	0 (0)	0 – 79
Protocol available in clinical trial record only**	6 (55)	28 – 79

*Denominator: Completed RCTs with results (n= 81)

**Denominator: Completed RCTs without results (n= 11)

^aAdditional information provided in Section E of the supplement

Ongoing RCTs

We included 42 eligible ongoing RCTs (Table 3). We identified 16 of these by both sources (representing 38% agreement (95% CI 25 - 53)). The remaining 26 were identified through CTR search and only available as CT records, accounting for 62% of the total (95% CI 47 - 75).

Table 3 Identification source of ongoing RCTs (n=42)

Ongoing RCTs	N (%)	(95%-CI)
<i>Protocol identified by both sources</i>	16 (38)	25 – 53
<i>Protocol identified by EBD search only</i>	0 (0)	0 – 1
Registered RCT	0 (0)	0 – 1
Non-registered RCT	0 (0)	0 – 1
<i>Protocol identified by CTR search only</i>	26 (62)	47 – 75
Protocol published in indexed journal	0 (0)	0 – 1
Protocol published in non-indexed journal	0 (0)	0 – 1
Protocol available in clinical trial record only	26 (62)	47 – 75

RCTs of unknown status

We identified four RCTs of unknown status by both sources (100% (95% CI 51 – 100)) (Table 4). None of our request for information on these trials succeeded.

Table 4 Identification source of RCTs of unknown status (n=4)

RCTs of unknown status	N (%)	(95%-CI)
<i>Protocol identified by both sources</i>	4 (100)	51 – 100
<i>Protocol identified by EBD search only</i>	0 (0)	0 – 49
<i>Protocol identified by CTR search only</i>	0 (0)	0 – 49

Table 5 *Characteristics of included RCTs (n=138)*

Characteristics	N (%)
Study Location	
Africa	0
Asia	54 (38)
Australia	1 (1)
Europe	25 (18)
North America	11 (9)
South America	3 (2)
Multi-Continental	44 (32)
Funding	
Government	9 (7)
Industry	73 (53)
Institutional	28 (20)
Unspecified/miscellaneous	8 (6)
Government, Institutional	3 (2)
Industry, Institutional	16 (11)
Industry, Government	1 (1)
Year of publication (median, range)*	2018 (2007 - 2023)
Year of planned completion (median, range)**	2021 (2015 - 2026)

*RCTs with published articles (n=101)

**RCTs without published articles and (planned) completion date available in CT record (n=21)

DISCUSSION

We conducted an evaluation study based on a case study of cardiovascular RCTs to determine whether CTR searches can accurately identify a greater number of completed RCTs than EBD searches and quantified the number of eligible ongoing RCTs. Our results indeed showed that CTR searches identify a greater number of completed RCTs compared to EBD searches (n=83 vs. n=79). However, this was mainly due to completed RCTs for which results were not published. Moreover, we found a high agreement of completed RCTs with published results between the two searches (81%). We identified some completed RCTs exclusively by one source. Furthermore, we identified all ongoing trials either by both sources or exclusively by CTR search. Our findings support methodological guidelines recommending the inclusion of CTR searches in systematic review search strategies and align with previous study results showing that CTR searches identify additional RCTs for systematic reviews of interventions (2, 4, 40).

In recent years, a priori CT registration has significantly improved and has contributed to the increased accuracy of CTR searches. In our study, 96% of the included RCTs were registered in a CTR. These results differ from previous findings by Glanville et al. in 2014, which indicated low compliance with CT registration but align with findings by Knelangen et al. in 2018 which indicate high adherence to CT registration, particularly in RCTs that focus on new drugs (19). Moreover, our study confirms the observations of Glanville et al. and Knelangen et al. that the CTR search primarily overlooks unregistered trials (18, 19). Nevertheless, it is worth noting that in Glanville et al.'s study, RCTs were still missed even after 2005, whereas Knelangen et al. found that the non-registered studies were either pre-2005 or non-RCTs. In contrast, our study found that all five RCTs lacking CT registration were conducted after 2005. While Knelangen et al. predominately analysed pivotal trials conducted within the context of drug approval in Europe, Glanville et al. and our study included RCTs from a broader geographical context. This may explain the difference in findings between our study and the more recent findings by Knelangen et al.

Lacking CT registration presents a major challenge to exclusively performing CTR searches. Universal coverage of a CT registration mandate and better adherence to the already existing ones are important pillars for exclusively performing CTR searches in the future (3, 41, 42). Furthermore, the CT search missed a few trials in our study, despite being available in ClinicalTrials.gov, simply because they were not included in CENTRAL. A search directly through ClinicalTrials.gov or the ICTRP would have likely led to the identification of these trials.

Retrieving results from CT records requires additional effort. Most CT records do not have any results posted or provide links to published articles (20, 43). In our study, this was especially the case for ICTRP records. Results reporting items in the CT record were nearly always incomplete, making manual searches to obtain the published article or contacting study authors for information necessary. ClinicalTrials.gov-registered trials, on the other hand, provided CT data as well as links to the published articles more frequently.

CTR searches remain the most important tool to explore the presence of publication bias on systematic review findings. In our study, we found that searching CTRs leads to the identification of additional RCTs that are not included in EBDs yet (Table 2-4). Nearly half of the unpublished trials had results posted in their CT record. Systematic review findings may be greatly impacted by publication bias if only EBDs are searched and unpublished trials not identified (11, 12).

The screening effort for eligible RCTs is significantly reduced with a CTR search compared to that of an EBD search. In our study, we needed to screen one twentieth of records identified by CTR search compared to the number of records identified by EBD search (n=523 vs. n=9,395), while identifying more eligible completed RCTs with the CTR search (n=83 vs. n=79).

Implications

CTR searches are valuable for identifying completed RCTs without published articles and ongoing RCTs. In our study, the CTR search identified more completed trials than the EBD search and identified all eligible ongoing trials. Nevertheless, the CTR search did not identify all completed RCTs with available results. Therefore, researchers (i.e. librarians/information specialists and systematic reviewers) aiming to identify all RCTs with available results should continue to search both CTRs and EBDs.

CTR searches have the advantage of significantly reducing the screening effort. For systematic searches where the absence of a few RCTs (i.e. unregistered RCTs) is acceptable, performing only CTR searches is a good alternative to the combined search approach. Nevertheless, an important disadvantage of performing CTR searches is the current infrastructure of most CTRs. In order to make CTR searches more user-friendly for researchers (i.e. librarians/information specialists, systematic reviewers), some important issues need to be addressed. First, results reporting in the CT records needs to be improved significantly. Most completed RCTs did not report any results in their CT record. Second, links to the published articles need to be added to the CT records. These were not provided in most cases, especially for ICTRP records. The manual search for results and published articles contributed significantly to the workload of our project. Finally, we also found that it was impossible to download all published articles from the CT record at once or to load them directly into our systematic review software. Addressing these issues will greatly improve the efficiency of CTR searches and make it more feasible to search CTRs alone for RCTs for systematic reviews.

Strengths and Limitations

Our study has several strengths. We followed recommended methodological guidelines to identify RCTs for systematic reviews of interventions (2, 44). Moreover, we searched multiple databases with different combinations of search terms (23, 25, 29-31). Also, we contacted study authors for information on missing trial information. Besides, our search represented a research question from clinical practice. Nevertheless, our study has some important limitations. First, CT registration has only been mandated in major

medical journals since 2005 (3). CTR searches may not identify older trials since these may have fewer registrations available (3, 18). However, we specifically focused on DOACs since they emerged relatively recently, and we expected a larger amount of RCTs identifiable by both sources than with any (older) medication group included in the systematic review on which the case study was based upon (22). Second, we performed our evaluation based on a specific subset of cardiovascular RCTs. Although the proof-of-concept nature of our study allowed for this, our results may differ from other clinical fields and may require confirmation. Third, although CTRs include all types of intervention studies, we specifically focused on RCTs. This design limitation means that our conclusions are not applicable to CTR searches of other study designs. Nonetheless, other study designs are less frequently registered than RCTs and therefore less relevant for CTR searches (3, 45-47). Fourth, we only screened some records in duplicate. However, we held a calibration session before full-text screening, and the reviewers achieved our pre-specified interrater reliability threshold of 0.7. In addition, we checked for the presence of unmatched RCTs in the original datasets to ensure that incorrectly excluded RCTs were rightfully included in the final analysis.

CONCLUSION

Our study findings show that CTR searches identify a greater number of completed RCTs compared to EBD searches. In addition, CTR searches allowed us to identify all eligible ongoing RCTs. Of note, CTR searches also identify RCTs that have been completed but not yet published. Nevertheless, we found that some RCTs were identified exclusively by either CTR or EBD search. Hence, if the goal is to identify all RCTs with available results, both CTRs and EBDs should continue to be searched. Once the barriers to conducting CTR searches alone are removed, CTR searches will be a more suitable alternative to the combined approach of searching both EBDs and CTRs. Therefore, improving the usability and infrastructure of CTRs would be beneficial to facilitate the review process and enable the assessment of critical information.

Acknowledgements

None.

Authors' contributions

Tabea Kaul: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – Original draft, Visualization, Project administration.

Julia MT Colombijn: Conceptualization, Methodology, Investigation, Writing – Review & Editing. **Robin WM Vernooij:** Conceptualization, Writing – Review & Editing.

Rene Spijker: Methodology, Software, Investigation, Writing – Review & Editing.

Demy L Idema: Investigation, Writing – Review & Editing.

Linde F Huis in 't Veld: Investigation, Writing – Review & Editing.

Johanna AA Damen: Conceptualization, Methodology, Investigation, Writing – Review & Editing. **Lotty Hooft:** Conceptualization, Methodology, Supervision, Writing – Review & Editing.

Funding

None.

Competing interests

The work of JMT Colombijn & RWM Vernooij was supported by the Dutch CardioVascular Alliance, an initiative with the support of the Dutch Heart Foundation, 2020B008 RECONNECT. The other 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.

Ethical approval

Not required for this work.

Registration and protocol

The protocol of this study was registered on the Open Science Framework and is available at: <https://doi.org/10.17605/OSF.IO/WJ9BA>.

Data availability

Data are available upon reasonable request to the corresponding author.

Supplementary data

Supplementary data to this article can be found online.

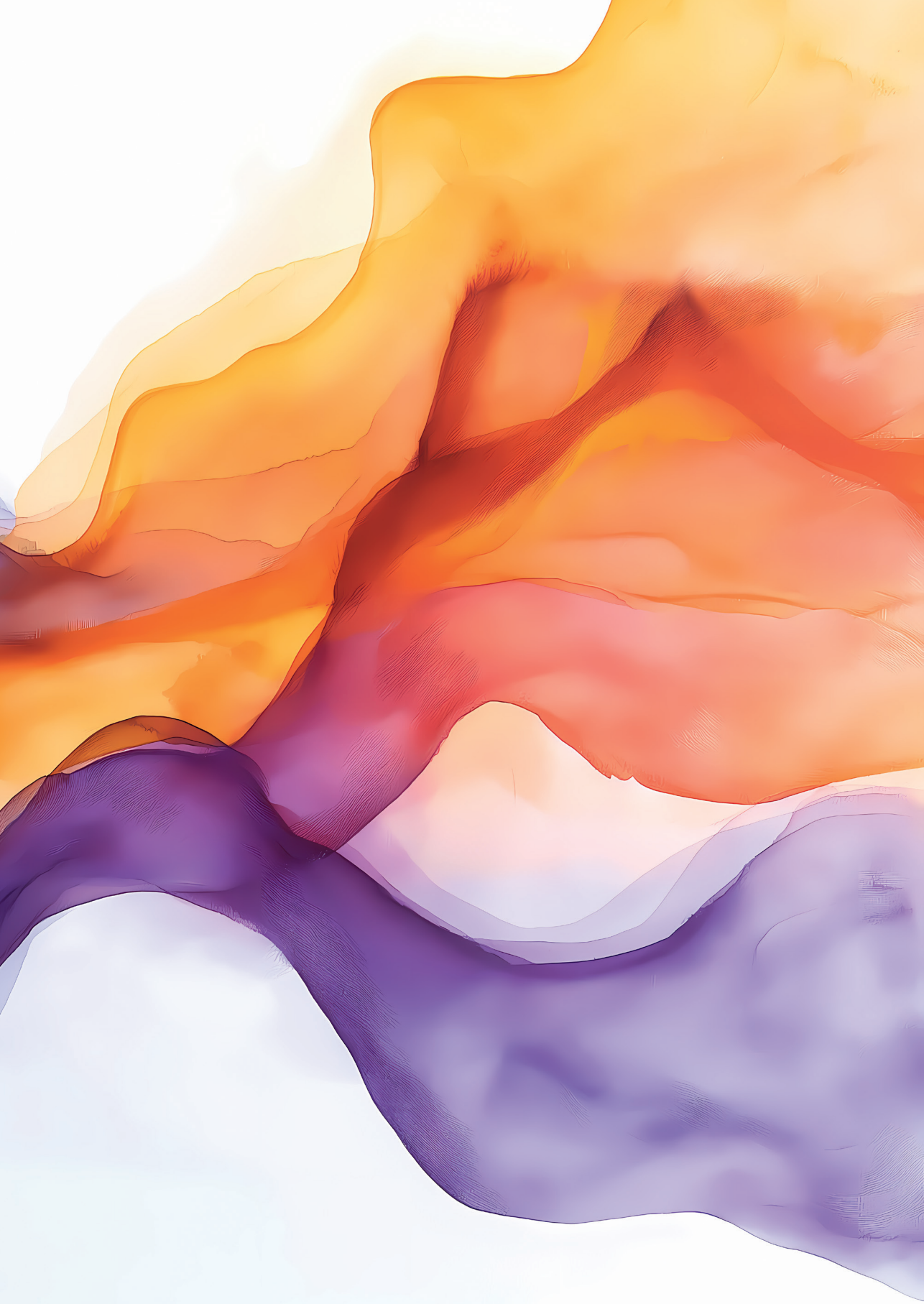


REFERENCES

- (1) Hariton E, Locascio JJ. Randomised controlled trials – the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2018;125:1716-. <https://doi.org/10.1111/1471-0528.15199>
- (2) Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook/;2022[accessed 22 April 2023].
- (3) De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med*. 2004;351:1250-1. <https://doi.org/10.1056/NEJMe048225>
- (4) Baudard M, Yavchitz A, Ravaut P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. *BMJ*. 2017;356:j448. <https://doi.org/10.1136/bmj.j448>.
- (5) Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009. <https://doi.org/10.1002/14651858.MR000006.pub3>.
- (6) Tan AC, Jiang I, Askie L, Hunter K, Simes RJ, Seidler AL. Prevalence of trial registration varies by study characteristics and risk of bias. *J Clin Epidemiol*. 2019;113:64-74. <https://doi.org/10.1016/j.jclinepi.2019.05.009>.
- (7) Lefebvre C GJ, Briscoe S, Featherstone R, Littlewood A, Marshall C, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland LS. Chapter 4: Searching for and selecting studies. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA . editor. *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3. Cochrane: Cochrane; 2022. Available from www.training.cochrane.org/handbook/;2022 [accessed 22 April 2023]
- (8) Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol*. 2011;64:1277-82. <https://doi.org/10.1016/j.jclinepi.2011.01.011>
- (9) Schmucker C, Schell LK, Portalupi S, Oeller P, Cabrera L, Bassler D, et al. Extent of non-publication in cohorts of studies approved by research ethics committees or included in trial registries. *PLoS One*. 2014;9:e114023. <https://doi.org/10.1371/journal.pone.0114023>.
- (10) McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research - a narrative review. *Trials*. 2010;11:37.
- (11) Jefferson T, Jones M, Doshi P, Del Mar C. Neuraminidase inhibitors for preventing and treating influenza in healthy adults: systematic review and meta-analysis. *BMJ*. 2009;339:b5106. <https://doi.org/10.1186/1745-6215-11-37>
- (12) Jefferson T, Jones MA, Doshi P, Del Mar CB, Hama R, Thompson MJ, et al. Neuraminidase inhibitors for preventing and treating influenza in adults and children. *Cochrane Database of Systematic Reviews*. 2014. <https://doi.org/10.1002/14651858.CD008965.pub4>.
- (13) Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 2008;3:e3081. <https://doi.org/10.1371/journal.pone.0003081>.
- (14) Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867-72. [https://doi.org/10.1016/0140-6736\(91\)90201-Y](https://doi.org/10.1016/0140-6736(91)90201-Y).
- (15) Abbot NE, Ernst E. Publication bias: direction of outcome less important than scientific quality. *Perfusion* 1998;11:182-4.

- (16) Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*1990;263:1385-9.
- (17) Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess* 2000;4:1-115. <https://doi.org/10.3310/hta4100>.
- (18) Glanville JM, Duffy S, McCool R, Varley D. Searching ClinicalTrials.gov and the International Clinical Trials Registry Platform to inform systematic reviews: what are the optimal search approaches? *J Med Libr Assoc.* 2014;102:177-83. <https://doi.org/10.3163/1536-5050.102.3.007>.
- (19) Knelangen M, Hausner E, Metzendorf M-I, Sturtz S, Waffenschmidt S. Trial registry searches for randomized controlled trials of new drugs required registry-specific adaptation to achieve adequate sensitivity. *J Clin Epidemiol.* 2018;94:69-75. <https://doi.org/10.1016/j.jclinepi.2017.11.003>.
- (20) Goldacre B, DeVito NJ, Heneghan C, Irving F, Bacon S, Fleminger J, et al. Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource. *BMJ.* 2018;362:k3218.<https://doi.org/10.1136/bmj.k3218>.
- (21) Kaul T, Colombijn, JMT; Vernooij, RWM; Spijker, R; Damen, JAA; Hooft, L. Protocol registration: Clinical Trial Register Searches for Systematic Reviews of Randomized Clinical Trials Involving Interventions. online 2022. <https://doi.org/10.17605/OSF.IO/E647J>.
- (22) Colombijn J ID, van de Braak K, Spijker R, Meijvis SCA, Bots M L, Hooft L, Verhaar MC, Vernooij RWM, Evidence for pharmacological interventions to reduce cardiovascular risk for patients with chronic kidney disease: A study protocol of an evidence map. *BMC Systematic Reviews* 2022. <https://doi.org/10.1186/s13643-022-02108-x>
- (23) Cochrane Library. Cochrane Central Register of Controlled Trials (CENTRAL), <https://www.cochranelibrary.com/central/about-central>; 2022 [accessed 29 August 2022].
- (24) Cochrane Library. How CENTRAL is created, <https://www.cochranelibrary.com/central/central-creation>; 2023 [accessed 30 May 2023].
- (25) ClinicalTrials.gov. ClinicalTrials.gov Homepage, <https://clinicaltrials.gov/>; 2022 [accessed 21 March 2022].
- (26) Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, et al. Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO). *Kidney Int.* 2005;67:2089-100. <https://doi.org/10.1111/j.1523-1755.2005.00365.x>.
- (27) Gansevoort RT, Correa-Rotter R, Hemmelgarn BR, Jafar TH, Heerspink HJ, Mann JF, et al. Chronic kidney disease and cardiovascular risk: epidemiology, mechanisms, and prevention. *Lancet.* 2013;382:339-52. [https://doi.org/10.1016/s0140-6736\(13\)60595-4](https://doi.org/10.1016/s0140-6736(13)60595-4).
- (28) International Diabetes Federation. The IDF consensus world wide definition of the METABOLIC SYNDROME Brussels, Belgium 2006, <https://www.idf.org/e-library/consensus-statements/60-idfconsensus-worldwide-definitionof-the-metabolic-syndrome.html#:~:text=IDF%20Consensus%20Worldwide%20Definition%20of%20the%20Metabolic%20Syndrome&text=The%20metabolic%20syndrome%20is%20a,cholesterol%20and%20high%20blood%20pressure/>; 2023 [accessed 18 May 2003].
- (29) World Health Organization. International Clinical Trials Registry Platform (ICTRP), <https://trialsearch.who.int/>; 2022 [accessed 29 August 2022].
- (30) National Library of Medicine. MEDLINE. National Center for Biotechnology Information (NCBI), <https://www.nlm.nih.gov/medline/index.html/>; 2021 [accessed 29 August 2022].
- (31) Elsevier. Embase: The comprehensive biomedical research database, <https://www.elsevier.com/solutions/embase-biomedical-research/>; 2022 [accessed 17 May 2022].
- (32) McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012;22:276-82.
- (33) Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement.* 1960;20:37-46. <https://doi.org/10.1177/001316446002000104>.

- (34) Microsoft Corporation. Microsoft Excel (Version 16.68), <https://office.microsoft.com/excel/>; 2022 [accessed 26 January 2023].
- (35) Department of Health and Human Services. Clinical Trials Registration and Results Information Submission. In: Federal Register: The Daily Journal of the United States Government; 2016. p.177.
- (36) 110 U.S. Congress. Food and Drug Administration Amendments Act of 2007. Title VIII, Section 801. United States of America: U.S. Government; 2007.
- (37) World Health Organization. WHO Statement on public disclosure of clinical trial results, <https://www.who.int/news/item/09-04-2015-japan-primary-registries-network/>; 2022 [accessed 18 October 2022].
- (38) Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*. 1927;22:209-12. <https://doi.org/10.1080/01621459.1927.10502953>.
- (39) R Core Team. R: A language and environment for statistica computing. R version 2022.07.2, Vienna, Austria: R Foundation for Statistical Computing, <https://www.R-project.org/>; 2022 [accessed 21 December 2022].
- (40) Alqaideem Z, Nguyen P-Y, Awadh M, Page MJ. Impact of searching clinical trials registers in systematic reviews of pharmaceutical and non-pharmaceutical interventions: Reanalysis of meta-analyses. *Research Synthesis Methods*. 2023;14:52-67. <https://doi.org/10.1002/jrsm.1583>.
- (41) Isojarvi J, Wood H, Lefebvre C, Glanville J. Challenges of identifying unpublished data from clinical trials: Getting the best out of clinical trials registers and other novel sources. *Research Synthesis Methods*. 2018;9:561-78. <https://doi.org/10.1002/jrsm.1294>.
- (42) Viergever RF, Li K. Trends in global clinical trial registration: an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. *BMJ Open*. 2015;5:e008932. <https://doi.org/10.1136/bmjopen-2015-008932>.
- (43) Huser V, Cimino JJ. Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. *PLoS One*. 2013;8:e68409. <https://doi.org/10.1371/journal.pone.0068409>.
- (44) Hunter KE, Webster AC, Page MJ, Willson M, McDonald S, Berber S, et al. Searching clinical trials registers: guide for systematic reviewers. *BMJ*. 2022;377:e068791. <https://doi.org/10.1136/bmj-2021-068791>.
- (45) Zarin DA, Fain KM, Dobbins HD, Tse T, Williams RJ. 10-Year Update on Study Results Submitted to ClinicalTrials.gov. *New England Journal of Medicine*. 2019;381:1966-74. <https://doi.org/10.1056/NEJMs1907644>.
- (46) ClinicalTrials.gov. Trends, Charts, and Maps, <https://clinicaltrials.gov/ct2/resources/trends#LocationsOfRegisteredStudies/>; 2021 [accessed 21 March 2022].
- (47) Gresham G, Meinert JL, Gresham AG, Piantadosi S, Meinert CL. Update on the clinical trial landscape: analysis of ClinicalTrials.gov registration data, 2000–2020. *Trials*. 2022;23:858. <https://doi.org/10.1186/s13063-022-06569-2>.



3

Improving the Reporting and Use of Trial Results in Clinical Trials Registries: Global Practices, Barriers, and Recommendations

Tabea Kaul, Johanna AA Damen, Anna Lene Seidler, Melina Willson, Ghassan Karam,
Demy Idema, Mike Kusters, Mary Ann Dowsett, Tala Ibrahim Hasan Abutahoun,
Lotty Hooft, Kylie E Hunter

Manuscript submitted for publication

ABSTRACT

Background and Objectives: Recent initiatives promoted results reporting on clinical trials registries to improve transparency and reduce publication bias. However, local reports suggest that results reporting on registries is often inadequate, limiting their usefulness for evidence synthesis. We aimed to i) provide an overview of results reporting practices across clinical trials registries globally, ii) identify barriers and facilitators to reporting and using results from trials registries, and iii) develop recommendations to improve reporting and usability of results in trials registries.

Study Design and Setting: Three-part mixed methods study. Part 1: Descriptive analysis of results reporting practices for randomised controlled trials (RCTs) starting between 2010 and 2022 across six trials registries (one from each WHO region), with an in-depth analysis focusing on reporting formats and accessibility. Part 2: Two separate online surveys targeting trial registrants and evidence users. Part 3: Discussion amongst author group to generate recommendations.

Results: Part 1: Our sample included 201,265 RCTs, with 17% (33,163 trials) reporting some form of results on a registry. A subset showed 63% of posted results accessible in the registry record, with 64 to 98% of results data available in a reusable format. Part 2: 86% (194/225) of registrants were aware of registry results reporting possibilities, but time, effort, and fear of publishing interference were barriers. For evidence users, 51% (36/70) had used registry results, with barriers including mistrust of non-peer-reviewed data and difficulty locating results. Part 3: Recommendations include standardizing registry interfaces, addressing misconceptions, and fostering trust in registry-reported results.

Conclusion: Results reporting practices on registries are increasing. Improving these requires better infrastructure, policies, training, and funding. With adequate support, registries can become essential for transparent and efficient evidence dissemination, enhancing research quality, and reducing duplication.

INTRODUCTION

Currently, only about half of all research is ever published and outcomes are often underreported (1, 2). Consequently, medical decisions frequently rely on incomplete or biased evidence (3). Unfortunately, this issue mainly affects research with ‘negative’ or null results, potentially leading to distorted findings in systematic reviews and other scientific works (4). Timely reporting of trial results is important for many reasons. It avoids unnecessary duplication of research, which wastes valuable resources and constitutes unnecessary participant burden, mitigates publication bias, and ensures regulatory and public health decision-making are based on all available evidence (5, 6). A proposed solution to address the delayed or lack of publication of clinical trial results is to make them available on trial registration records (6, 7).

Clinical trials registries are a useful resource to identify unpublished trials, helping to mitigate or assess the extent of publication bias (4). Another benefit is that reporting results on registries takes less time than going through the process of publishing a manuscript in a medical journal. Therefore, disseminating trial results through registry records may even be a preferred option, particularly if up-to-date results are urgently required, e.g., to align with open-access policies and time frames required by funders, or during a pandemic.

Since 2004, policies, legislation, and agreements around reporting of trial results on registries have been developed, mainly focussing on clinical trials involving drugs and/or medical devices (8-11). In 2015, the World Health Organization (WHO) addressed a broader audience calling for disclosure of findings of key outcome measures on a clinical trials registry for all clinical trials (5, 7). Many funders of medical research have supported this statement (12). Aligning with the WHO’s 2015 Joint Statement, the May 2022 World Health Assembly (WHA) resolution (75.8) mandates trial registration and promotes timely disclosure of trial results in registries (7, 13). Following this, the WHO’s International Clinical Trials Registry Platform (ICTRP) refined international standards for the reporting of trial results on registries (including an open consultation in 2022) and has recently published updated guidance, including three key recommendations to implement the reporting of summary results (11, 14).

Local reviews suggest that results reporting rates in registries may be low (15-18). A recent study found that only about half (53%) of the 500 included EU-CTR-registered trials had results available in their registry record (18). Alayche et. al reported even less for Canadian ClinicalTrials.gov-registered trials (38% of 6720 included trials)(17). Furthermore, differences between reported results in journal publications and those

included in registry records may jeopardize the use of registry results varying from differences in which outcomes are reported (i.e., selective reporting) to differences in reported adverse events (19-22). To date, no comprehensive overview exists of results reporting practices across registries globally.

Given the need for action per WHO and WHA agreement and the essential role of registries in evidence-based medicine, we aimed to 1) provide an overview of results reporting practices of randomised controlled trials (RCTs) across registries, 2) identify barriers and facilitators to reporting results on and using results from registries, and 3) develop recommendations to increase the reporting and usability of results available on registries.

METHODS

We performed a three-part mixed-methods study including a quantitative and a survey-based approach. The project protocol was prospectively posted on Open Science Framework before any analyses were conducted (DOI 10.17605/OSF.IO/3NDHK).

Part 1: Overview of results reporting practices across selected clinical trials registries

Eligible trial records

We included RCT records starting between January 1st, 2010, and December 31st, 2022 (since we did not expect many trials registered past 2022 to have results available (7, 13)) from a selection of registries, comprising one registry per region within the WHO's Registry Network (23). Based on their size (i.e., number of registered records) and the possibility to report results, we selected the following registries:

- African Region: Pan African Clinical Trials Registry (PACTR);
- Eastern Mediterranean Region (EMR): Iranian Registry of Clinical Trials (IRCT);
- European Region (EUR): ISRCTN Registry;
- Southeast Asia Region (SEAR): Thai Clinical Trials Registry (TCTR);
- Region of the Americas (AMR): ClinicalTrials.gov;
- Western Pacific Region (WPR): Australian New Zealand Clinical Trials Registry (ANZCTR).

We were unable to select trials based on their completion status, completion date, or results posting date, as this information was not available for a subset of registered

trials. In case the actual trial start date was not available, we opted for the anticipated start date. If that date was not available, we included records based on registration date.

For an in-depth analysis, we randomly sampled 50 records per registry with results posted between April 2023 and April 2024. We chose this sample size because we considered it sufficient to provide an overview of results reporting practices, while maintaining feasibility based on available resourcing. If fewer than 50 records were available, we sampled backwards to reach 50 records per registry.

Retrieval of trial records

We downloaded trial records from ANZCTR (April 17th, 2024), ClinicalTrials.gov (April 11th, 2024), and ISRCTN (April 11th, 2024) from the registry websites in CSV format. We received data on request from PACTR in CSV format (April 15th, 2024) and from TCTR in XML format (April 30th, 2024). Upon inquiry, IRCT asked us to download records from the WHO's ICTRP website.

Data preparation and screening

TK performed data cleaning per included registry dataset by identifying data items corresponding to our pre-defined data extraction items (Supplement, Section A). Data items were harmonised and merged into one large dataset including all registry datasets. Then, TK screened records based on their eligibility by filtering the dataset on study type (i.e., RCTs) and trial start date (i.e., January 1st, 2010, and December 31st, 2022). To verify the included data, TK checked 10 records per registry based on the files provided per registry.

Data extraction and analysis

We defined *trial results data* as any numeric data relating to the final sample or outcomes of an RCT, i.e., any of the following: baseline characteristics, participant flow, primary outcome, secondary outcomes, or adverse events (24). We defined *reporting of results* as any trial results data that were available via the registry record, i.e., available via added link to a results publication in a medical journal, via a link to a data repository (e.g., Open Science Framework), as attachment (e.g., as a PDF) or embedded within the registry (e.g., as tables). The data extractions for the overview of results reporting practices were performed by TK.

For the in-depth analysis, we evaluated whether results were accessible and usable. We defined *usable* as including results data for each comparison group (i.e., numbers, effect estimates and descriptions, as applicable). Results were *partially usable* if data

were described only for part of the study sample (e.g., only for one comparator group, certain participants, or intervention centres). Results were *not usable*, if description and/or numbers were lacking. Also, for primary and secondary outcomes, we assessed any discrepancies between planned and reported outcomes (based on trial record). The data extractions for the in-depth analysis were performed by MK, DI, MAD, and TIHA. MK, DI, and TK verified the extractions (i.e., MK and DI did not verify their own extractions but those of others). The supplement (Sections A & B) includes further details on data extraction and outcome definitions (i.e., what can be considered ‘usable’ per data item).

We calculated frequencies (counts) and proportions (percentages) of all categorical outcomes, as well as median with interquartile ranges of the months between time of trial completion and reporting of results in the registry record. All statistical analyses were performed using R studio version 2023.12.1+402 (25).

Part 2: Cross-sectional web-based surveys

Part 2 included two online surveys conducted with the data capturing software Castor (26) (Supplement, Section E). The first survey targeted registrants who had registered a trial on a registry and focused on results *reporting*. The second survey aimed to identify barriers to *using* results from registry records and therefore included evidence users (i.e., anyone that may be using the results of trials from registries for their purpose, e.g., systematic reviews).

We developed the surveys informed by literature review and expert discussion (within the author group). The surveys comprised an informed consent procedure and multiple-choice items accompanied by free-text boxes for comments. Some questions were contingent on previous replies and/or trial status. Three volunteers from outside the author group (with relevant expertise) piloted the surveys before dissemination.

Recruitment of participants

For the registrant survey, we randomly sampled 1000 records per registry of trials included in Part 1 starting between 2015 and 2020 using simple randomisation with fixed seed, and invited researchers via the email address linked to the publicly available trial record. We chose this time frame since it takes about five years from first participant enrolment to journal publication (27).

For the evidence user survey, we invited participants via Cochrane platforms, LinkedIn, X, our networks, and by emailing authors of Cochrane systematic reviews of interventions published between 2023 and 2024 (n=305).

For both surveys, we asked the recipients of the invitation to forward the invitation to the correct person in case they did not align with the target group.

Analysis

We analysed the survey responses as frequencies and proportions where appropriate, accompanied by a narrative summary of the comments. We used all available data in case of incomplete responses. Data from Castor were exported to Microsoft Excel (28) and then to the open-source software R (25) for data cleaning, coding, and analysis.

Ethical approval

Although this study is outside of the scope of the *Dutch Medical Research Involving Human Subjects Act* (WMO)(29), UMC Utrecht independently reviewed it (case number: 24U-1501) and confirmed that no formal ethics approval was needed.

Part 3: Development of recommendations to improve the reporting and usability of results available in registries

TK and KEH drafted recommendations to improve the reporting and usability of results from registries, informed by the findings of study Parts 1 and 2. These recommendations were discussed and refined using a consensus process within the author group, comprising global experts in trial registration and evidence synthesis (TK, KEH, ALS, MW, LH, JAAGD, GK). Recommendations were revised (TK, KEH) and final consensus was agreed via email.

RESULTS

Part 1: Overview of results reporting practices across selected registries

We included a total of 201,265 records in our analysis. Of these, 17% (33,163 trials) included some form of results reporting (Table 1): 58% were drug trials, 40% had industry involvement (either as primary sponsor or funder), and most recruited participants in the region of the Americas (30%) (Supplement, Section C). Trials with an earlier start or completion date and drug interventions were more likely to report results (Supplement, Section C). Of 3638 COVID-19 trials, 521 reported results. Most results were posted in 2019 (Figure 1). The supplement includes further stratifications (Section C). Contrary to initial plans, we could only categorize interventions as “drug”

or “non-drug” due to labelling inconsistencies across registries preventing further granularity.

Table 1 *Availability of results in trials registries of trials starting between 2010 and 2022*

Registries	Total records (N)	Records reporting any results n (%)*
Overall	201,265	33,163 (17)
ANZCTR	11,745	1,144 (10)
ClinicalTrials.gov	169,621	26,266 (15)
IRCT	3,494	12 (0.03)
ISRCTN	9,920	5,172 (52)
PACTR	2,105	170 (8)
TCTR	4,380	399 (9)

*Numerator: Number of records reporting any results per registry, Denominator: Total records per registry;

ANZCTR: Australian New Zealand Clinical Trials Registry;

IRCT: Iranian Registry of Clinical Trials;

ISRCTN: ISRCTN Registry;

PACTR: Pan African Clinicals Trials Registry;

TCTR: Thai Clinical Trials Registry.

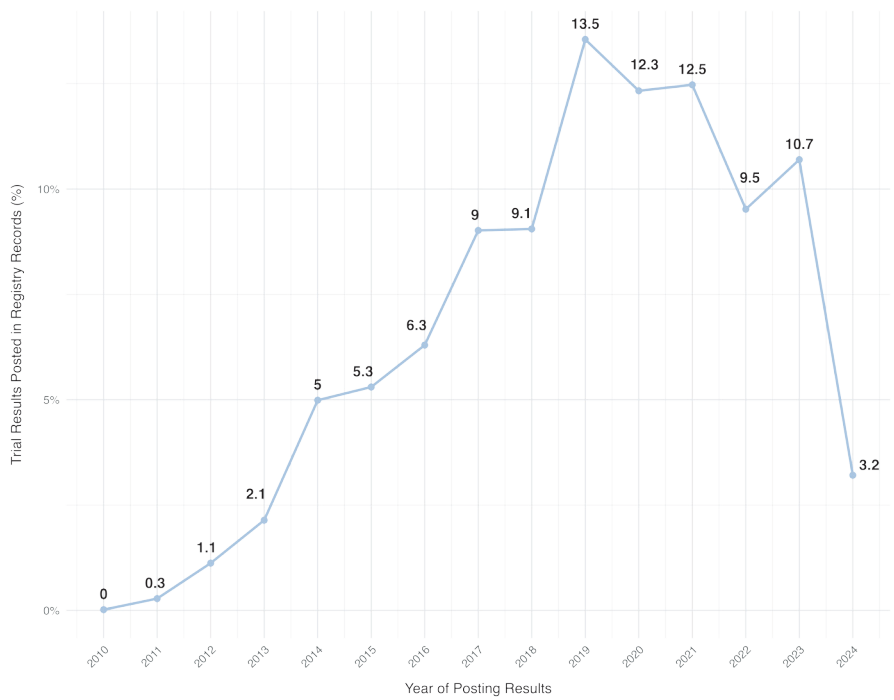


Figure 1 Trends in results reporting in clinical trials registries by year of posting results (Note: This analysis only includes records with an available results posting date (n=33,159). Four records included results posting dates outside of the above range (likely an automation mistake, all had trial start dates between 2010 and 2022). Therefore, Figure 1 contains a smaller number of records, although 33,163 trials had some form of results available.

In-depth analysis

Overall, we included 262 registry records in our in-depth analysis, i.e., 50 each from ANZCTR, ClinicalTrials.gov, ISRCTN, PACTR, and TCTR, and 12 from IRCT (owing to limited availability of results data) (Supplement, Section D). The median time between study completion and first posting of results in registry record was 16 months (IQR: 9-29 months) (Table 2). Most results were available in the registry record (63%), followed by links to journal publications (21%), and a combination of both (16%) (Figure 2).

Table 2 Median time in months between completion date and results posting date

Registry	N	Median (IQR range) in months*
Overall	262	16 (9-29)
ANZCTR	50	20 (10-34)
ClinicalTrials.gov	50	15 (9-26)
IRCT**	12	-
ISRCTN	50	19 (12-31)
PACTR	50	18 (12-30)
TCTR	50	8 (5-13)

*Only if completion date available (n=250)

**Does not include completion date item

We included registry records with manually entered or uploaded results and excluded those with results linked to published reports (78% [n=204]) (Supplement, Section D). Of these, 36% (n=74) included a link to a protocol, and 98% (n=199) reported results as aggregated data (i.e., summary results data). The results were mainly embedded (50%), rather than included as an attachment (47%). Most reported results data on primary outcome (96%) and least on race/ethnicity (32%). Sex data were the most usable (98%), while participant flow data (64%) were the least usable, largely due to missing sample sizes for each group. Primary outcomes were usable for 78%, while secondary outcomes were usable for 76%. Discrepancies (i.e., planned and reported outcomes did not match) in primary (16%) and secondary (50%) outcomes were primarily due to omission/non-reporting of the planned outcome. The supplement (Section D) presents further results.

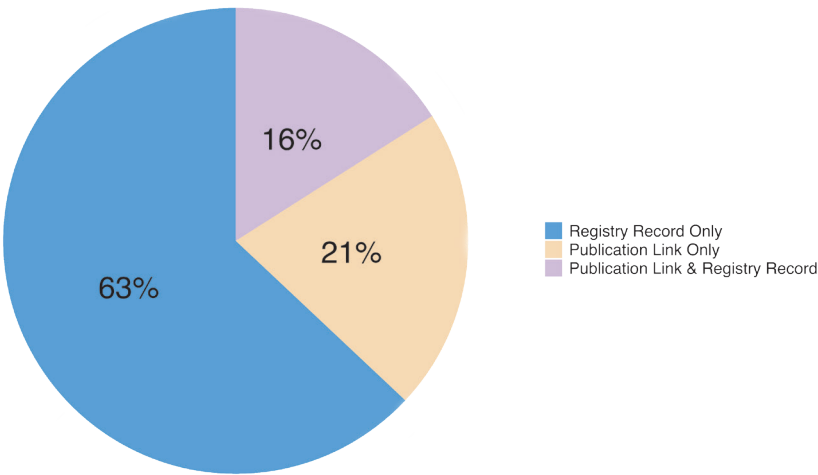


Figure 2 Availability of trial results data (n=262)

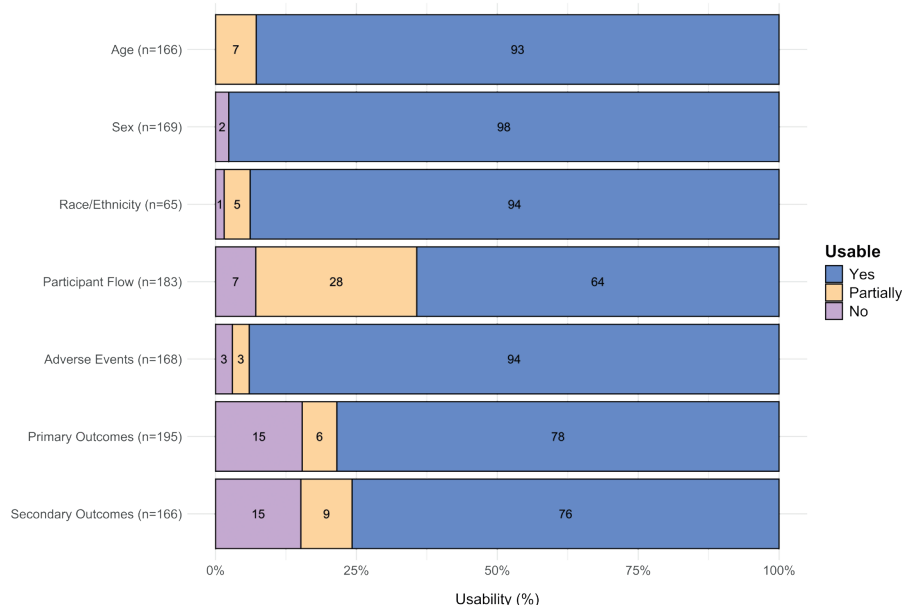


Figure 3 Usability of trial results data from registry records (n=204 [Note: This excludes records that only reported results as a link to a journal publication.]

Part 2: Cross-sectional web-based surveys

The supplement (Section E) includes summarized responses to surveys and free-text comments.

Registrant survey

This survey was open between July 12th to August 30th, 2024. We sent 6000 invitations (1000 per registry) via email, of which 1198 were undeliverable. Of the remaining 4802, 264 responded to the survey (208 complete; 56 partial). Participants were mostly principal investigators (70%) based in the Western Pacific Region (25%). 86% (194/225) were aware that registries include options to report results.

Registrants who had not reported results on the registry for trials completed more than 12 months ago, often prioritized publishing manuscripts over reporting results to registries (61%, 46/75). More than half of all registrants who reported their results (53%, 41/78) were not required by a research funder to do so.

According to free-text comments, the two main concerns for reporting results on registries include the fear of interference with a publishable journal manuscript and time and effort required for reporting. Also, some recommended the implementation

of reminders to update trial records, as well as the development of a standardized reporting format, with options for less common RCT designs (e.g., cluster RCTs).

Evidence user survey

This survey was open between July 8th and October 21st, 2024. Overall, 77 participants responded to the survey (69 complete; 8 partial) (Supplement, Section E). Respondents were primarily systematic reviewers (52%), followed by clinical trialists (23%), other professions (21%) (e.g., clinicians, academic staff, students), and guideline developers (4%). Most participants were based in the European region (51 %).

About half of participants had previously used registry results (51%; 36/70), primarily for systematic reviews or evidence synthesis (86%; 31/36). Among those who had not, the main barrier was inability to locate results (65%, 22/34). Most evidence users would be encouraged to use registry results data (85%; 61/72) if all trial information was centralized. Additionally, most preferred downloading the data as a file (66%; 46/70), ideally in Excel format (76%; 53/70).

According to free-text comments, the greatest barrier to using registry results is the lack of trust in non-peer reviewed data. Suggested improvements include enhancing registry searchability, standardizing reporting formats, enabling comments under records, and developing best practice guidelines for searching and using registry results for evidence synthesis.

Overall, the common ground between trial registrants and evidence users is the need for better systems to facilitate results reporting and access. They share the need for improved structure that streamlines processes with standardized formats, easier searchability, and downloadable files. For both stakeholder groups, trust in registry-reported results is required.

Part 3: Recommendations for improving reporting and usability of results included in registries

On October 22nd, 2024, the author group discussed and finalized recommendations to improve the reporting and usability of results available on registries. Boxes 1 and 2 include key concerns identified, proposed actions to address these concerns, and key implementers.

Box 1. Recommendations to improve reporting of trial results data in registries*

Identified concerns	Proposed actions	Stakeholders/ Implementers
Reporting results in a registry may interfere with the opportunity to publish results in a journal	• Promote and broaden dissemination of the ICMJE policy on results reporting and its implementation, including addressing misinterpretations • Clearly state in journal policy that there is no interference with publication if key outcome results are posted in registry record	• ICMJE • WHO • ICJME Journals • Collaborative groups (e.g., clinical trial networks, industry, investigator-initiated trial networks, clinical trial research offices, ethics committees)
	• Address registrants’ misconception about conflicts with journal publications	• Clinical trial institutes • Industry • ICMJE • WHO • ICMJE Journals • Collaborative groups (e.g., clinical trial networks, industry, investigator-initiated trial networks, clinical trial research offices, ethics committees)
Time and effort required to provide trial results	• Registries to offer multiple options for reporting results, such as automated indexing of journal publications and uploading Word documents of published or unpublished manuscripts • Develop a standard reporting template to guide trialists when reporting key results • Implement AI tools that extract data from study reports into the registry format	• Registries
Reporting for different types of trials	• Develop guidance on how best to report results for different study designs (e.g., registries can provide links to CONSORT for different trial designs as part of this guidance) • Include flexible templates for reporting of different types of studies	• Registries • WHO ICTRP

Box 1. Recommendations to improve reporting of trial results data in registries* (*continued*).

Identified concerns	Proposed actions	Stakeholders/Implementers
Incentives for registrants	Promote and broaden dissemination that trials difficult to publish in journals can still share key outcomes on registries (e.g., by incorporation into standard operating procedures (SOPS) on trial conduct, reporting, and dissemination)	- Research funders - Clinical trial institutes - Industry - Chief and associate trial investigators - Trial operational staff
Lack of awareness of the role of registries	Educate medical researchers and students on the importance of reporting trial results and incorporating training on how to effectively report these results in registries, embedding this knowledge as a core component of medical degree curricula	- Clinical trial governance and support offices at universities and hospitals - Collaborative groups (e.g., clinical trial networks, industry, investigator-initiated trial networks, clinical trial research offices, ethics committees [prompt researchers to update their trial record if trial completed]) - WHO
	Registry websites to include a webpage describing the reason and value in providing trial results (published or unpublished) in registries	Registries
Lack of awareness among trialists in updating registry records	- Send email reminders to principal investigators to update trial records with results - Include timeframes around updating registry records - Ethical standards and guidance to include updating registry records as part of their best standards in research practice (i.e., publication and dissemination sections)	- Registries - Ethics committees (during annual and final reporting cycles) - Research funders
Research funders do not require reporting results on registries	- Enforce existing recommendations - Report within 12 months' time frame after study completion in accordance with e.g., WHO policy/agreement, FDA	- Regulatory bodies (e.g., FDA, EMA) - Research funder - Public/consumer advocacy groups

*informed by Part 1 and 2 after discussion within the author group; AI: artificial intelligence

Box 2. Recommendations to improve use of trial results data from registries*

Identified concerns	Proposed action	Stakeholders/ Implementers
Accessibility of registries	• Enhance search functionality to identify trial protocols and results in registries (dependent on funding) • Enable automatic downloads of data from records • Standardise reporting format	• Registries supported financially by research funders (governments and not for profits)
Lack of awareness of the role of registries in logging results and protocols	• Registries work with key organisations in evidence synthesis to further promote their value • Educate medical researchers and students on the importance of registry data for evidence synthesis, incorporating training on how to effectively use results from registries (e.g., by developing best practice guidance), and by embedding this knowledge as a core component of medical degree curricula	• Registries • WHO • Cochrane • Joanna Briggs Institute (JBI) • Guidelines International Network (GIN) • Academic institutions • Collaborative groups (e.g., Cochrane)
Lack of trust in results	• Build trust in results reported on registries by facilitating engagement between primary study authors and other researchers (e.g., by enabling comment functions) • Introduce optional peer review within registries or quality checks to build trust • Encourage trustworthiness of information on registries by ensuring supportive information that help interpret the data are provided in the registry (e.g., protocols, SAPs, etc)	• Registries • Researchers

*informed by part 1 and 2 after discussion within the author group

DISCUSSION

Our analysis found substantial variation in results reporting (0.3 to 53%) across registries (Table 1), with most reported results (64 to 98%) being usable based on a subset (Figure 3, Supplement, Section D). Our findings corroborate local assessments that results reporting rates across registries vary and most are low (17, 18). Barriers to report and use registry results include the need for better systems to facilitate results reporting and access, insufficient incentives to report results, misconceptions about interference with publishable manuscripts, and lack of trust in registry-reported results. We propose recommendations to address these, building on previously developed recommendations (30-32). The updated WHO statement on the reporting of summary results in registries has recently been published (11). Our proposed actions for the key stakeholder groups extend these with more detail and identify issues specific to data re-use.

How can we promote the reporting of results in registries?

Cultural shift needed

Despite WHO and regulatory requirements, a key barrier remains the misconception that registry reporting counts as prior publication (33). However, major journals—including those following ICMJE guidelines—explicitly state that posting trial results in registries does not prevent journal publication (34-36). Further, the ICMJE website states (30):

“The ICMJE will not consider as prior publication the posting of trial results in any registry that meets the above criteria if results are limited to a brief structured abstract or tables (to include trial participants enrolled, baseline characteristics, primary and secondary outcomes, and adverse events).”

To counter this misconception, ICMJE, WHO, national clinical trial networks and advocacy groups should actively promote journal policies that allow registry reporting (11).

Another barrier is that medical researchers are sometimes not aware of registries and their role for research transparency. To prepare medical researchers and students to adequately perform trial reporting, we advocate for comprehensive training on research transparency as a core component of relevant health/science/medical degrees and during on-boarding processes of research-based work. Where already done,

research institutes and registries should thus continue to educate study investigators on the importance of reporting results directly on registries.

A perceived limited benefit for registrants to report results on registries may further affect reporting rates. Currently, the traditional academic system recognizes journal publications more than sharing results via other means (e.g., registries). Thus, a cultural shift is needed to provide incentives for researchers to share their results via registries, such as recognition of reported results for promotions and grant applications (similar to publication lists) or via shared authorship in case of secondary data use (37). Furthermore, by monitoring and quantifying the reporting adherence rates (e.g., as suggested by the updated WHO summary results reporting statement), there may be a further incentive to report results (11).

Simplifying processes and increasing efficiency

Technical enhancements are needed to improve results reporting rates (11). Registrants frequently criticised the time and effort involved with reporting results on registries (e.g., due to clunky interfaces, limited flexibility for less common RCT designs). The implementation of multiple options to report results, such as automated indexing of journal publications or enabling uploads of attachments (e.g., published or unpublished manuscripts), may increase reporting rates. For example, we identified great heterogeneity between results reporting rates, with one registry having as much as 53% of records including some form of results reporting (Table 1). Registries enabling automated links to journal publications have higher reporting rates, suggesting that integrating registry records with published literature could improve compliance (38). Also, we recommend registries develop guidance on how to best report results for different study designs.

Besides, the implementation of reminders to update registry records by registries, ethics committees, and research funders may be a useful tool to increase reporting rates, especially since the time between prospective trial registration and last participant visit can take years (27).

How can we enhance the usability of results from registries?

Improving accessibility for evidence synthesis

Technical enhancements to improve accessibility of trial results will promote their use. We identified a need for improvements in the search functionality of registries and the retrieval mechanisms. Funders should support registries in enhancing search interfaces, standardizing reporting formats, and implementing artificial intelligence

(AI)- driven data extraction to improve usability. While costly upfront, these changes will ultimately reduce research duplication and lower patient burden (39, 40). Therefore, similar to Chan et. al, we propose research funders to financially support registries in implementing infrastructure changes which will result in increased usability of results from registries (11).

Standardized (structured) reporting formats on trials registries can support manual evidence synthesis and simplify AI-driven data extraction (41, 42). While included registries adhered to reporting items proposed by WHO (32), they were displayed differently across registries, e.g., item 23. *Summary Results* of the WHO's minimum dataset could include reporting embedded as free-text (e.g., TCTR) or as tabular format (e.g., ClinicalTrials.gov), while others included results in attachments (e.g., PACTR). Thus, we advocate for the development of a standard reporting format to facilitate reuse. The updated WHO statement may be used to facilitate this process (11).

Promoting trust in registry results

Results from registry records are largely usable for evidence synthesis. Nevertheless, we identified a lack of trust in the reported results due to the absence of a formal peer-review process. Often trial information can be scattered across several different sources, prompting Cochrane authors to develop methods to deal with this in evidence synthesis (43). However, when all trial information is centralized in the registry record (e.g., protocols, statistical analysis plans, etc.), the results can be interpreted in a more reliable manner, thereby fostering trust in the results.

Where possible, systematic reviews incorporate registry results in their evidence synthesis, and guidelines on how to assess registry-reported data can help evidence users build confidence in their use (44).

Previous studies identified a higher percentage of discrepancies between planned and reported primary outcomes in main publications (~30%) compared to what we identified between planned and reported primary outcomes included in registry records (16%) (45, 46). Though our sample was relatively small, this may suggest that registries are less prone to selective outcome reporting, perhaps because both planned and reported outcomes are registered in the same location (i.e., the registry record). Future work could investigate this further.

Lastly, quality checks in registries can build trust in results data. However, implementation may strain already limited resources and lack feasibility, especially in low resource settings. Another angle of implementation may be via journals: differences between trial information included in the registry record and the publication do not

affect publication acceptance rates (22). But, if sporadically, journal editors would validate research findings with the available trial information in the registry record, this may encourage primary study authors to comprehensively register their results and help evidence users trust registry record findings.

Implications

We have outlined several recommendations to improve reporting and use of results available on registries and identified key stakeholders for implementation. The effort necessary to implement our proposed actions varies across identified concerns and stakeholder groups. While there is increasing awareness of the importance of trial results access via registries, enhancements are needed to maximise their value (7, 8, 10-13). In the future, we envisage efficient (re)use of data from registries will reduce duplication, improve timely access to results, and lower burden on researchers and patients. To achieve this, investment in registry infrastructure, including (but not limited to) user-friendly search interfaces, automatic data uploads (perhaps using AI tools) from different resources into standardised formats, and/or an increase in research staff capacity per registry, will require significant investment from funders (11).

Strengths and limitations

We evaluated the results reporting practices of RCTs on a large number of registry records across several international registries and surveyed global registrants and evidence users. Further, we developed key recommendations according to the results of the descriptive analysis, the surveys, and with input of international experts.

However, the descriptive analysis was restricted by missing or potentially outdated information on trial status and/or date of trial completion in the registries. Consequently, we calculated results availability by completion date where available (Supplement, Section C). Second, while we successfully obtained data directly from five registries (via direct download or email inquiry), data from IRCT could not be accessed directly (as confirmed via email consultation with IRCT) and was therefore retrieved from the ICTRP instead. To address this, for the in-depth analysis, we checked the registry records for all registries on the registries' websites (including for IRCT both the ICTRP record as well as the IRCT website), to ensure that the trial information was complete.

Third, for the in-depth analysis, we had to sample back to previous years for some registries to obtain our targeted 50 records including results per registries. This was the case for ANZCTR (2022: n=26, 2023-2024: n=24), IRCT (2018: n=8, 2019: n=1, 2020: n=1, 2022: n=1, 2023-2024: n=1), and PACTR (2022: n=16, 2023-2024: n=34). While for ANZCTR

and PACTR sampling up to 2022 sufficed to reach our target, we could only include 12 records from IRCT, despite sampling back up to 2018. While we do not expect much change in reporting practices between the period of 2022 and 2023/2024, it can be expected that results reporting practices (including infrastructures and requirements) have changed between 2018 and 2024. Thus, IRCT's older and limited data contributed less to our in-depth analysis and thus the findings of the in-depth analysis cannot be generalized.

Fourth, due to limited resourcing, we were unable to obtain higher response rates for the surveys. Nevertheless, our goal was to identify barriers and potential solutions to using results from registries. Thus, the results of the surveys, in combination with the information obtained during the descriptive analyses, and discussions amongst the author group comprising international experts, only contributed to developing potential solutions, rather than to draw concrete conclusions solely based on the survey responses.

CONCLUSION

The reporting and use of clinical trials registry results is increasing, but key barriers remain. Addressing these barriers will require a concerted effort from stakeholders, particularly funders, to prioritize investments in registry infrastructure and education. With adequate support, registries have the potential to become essential tools for improving research transparency and efficient evidence dissemination, ultimately improving research quality and efficiency.

Acknowledgements

None.

Authors' contributions

Tabea Kaul: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing - Original draft, Writing - Review & Editing, Visualization, Project administration.

Johanna AA Damen: Conceptualization, Methodology, Writing - Review & Editing.

Anna Lene Seidler: Conceptualization, Investigation, Methodology, Writing - Review & Editing.

Melina Willson: Conceptualization, Investigation, Methodology, Writing - Review & Editing.

Ghassan Karam: Conceptualization, Methodology, Writing - Review & Editing.

Demy Idema: Investigation, Writing - Review & Editing.

Mike Kusters: Investigation, Writing - Review & Editing.

Mary Ann Dowsett: Investigation, Writing - Review & Editing.

Tala Ibrahim Hasan Abutahoun: Investigation, Writing - Review & Editing.

Lotty Hooft: Conceptualization, Methodology, Writing - Review & Editing.

Kylie E Hunter: Conceptualization, Investigation, Methodology, Writing - Review & Editing, Supervision.

Funding

TK received a travel stipend from the Strategic Network Development grant of the University Medical Center Utrecht to support her research visit to the University of Sydney, which led to this project.

Competing interests

MW is the manager of the ANZCTR which is a clinical trials registry included in this research paper. MW was not involved in the review or analyses of ANZCTR data. KEH worked at the ANZCTR from 2009 to 2020 in a project officer/research role; she remains peripherally involved in ANZCTR related research. ALS was affiliated with ANZCTR from 2017 to 2024 as a research affiliate. ALS was not involved in the review or analyses of ANZCTR data. LH was the manager of the Netherlands Trial Register (NTR) until April 2022. However, this registry was not included in this research paper. All other 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.

Ethical approval

Although this study is outside of the scope of the *Dutch Medical Research Involving Human Subjects Act* (WMO), UMC Utrecht independently reviewed this study (case number: 24U-1501) and confirmed that no formal ethics approval was needed.

Registration and protocol

The protocol of this study was registered on the Open Science Framework and is available at: <https://doi.org/10.17605/OSF.IO/3NDHK>.

Data availability

Data are available upon reasonable request to the corresponding author.

Supplementary data

Supplementary data to this article can be found online.

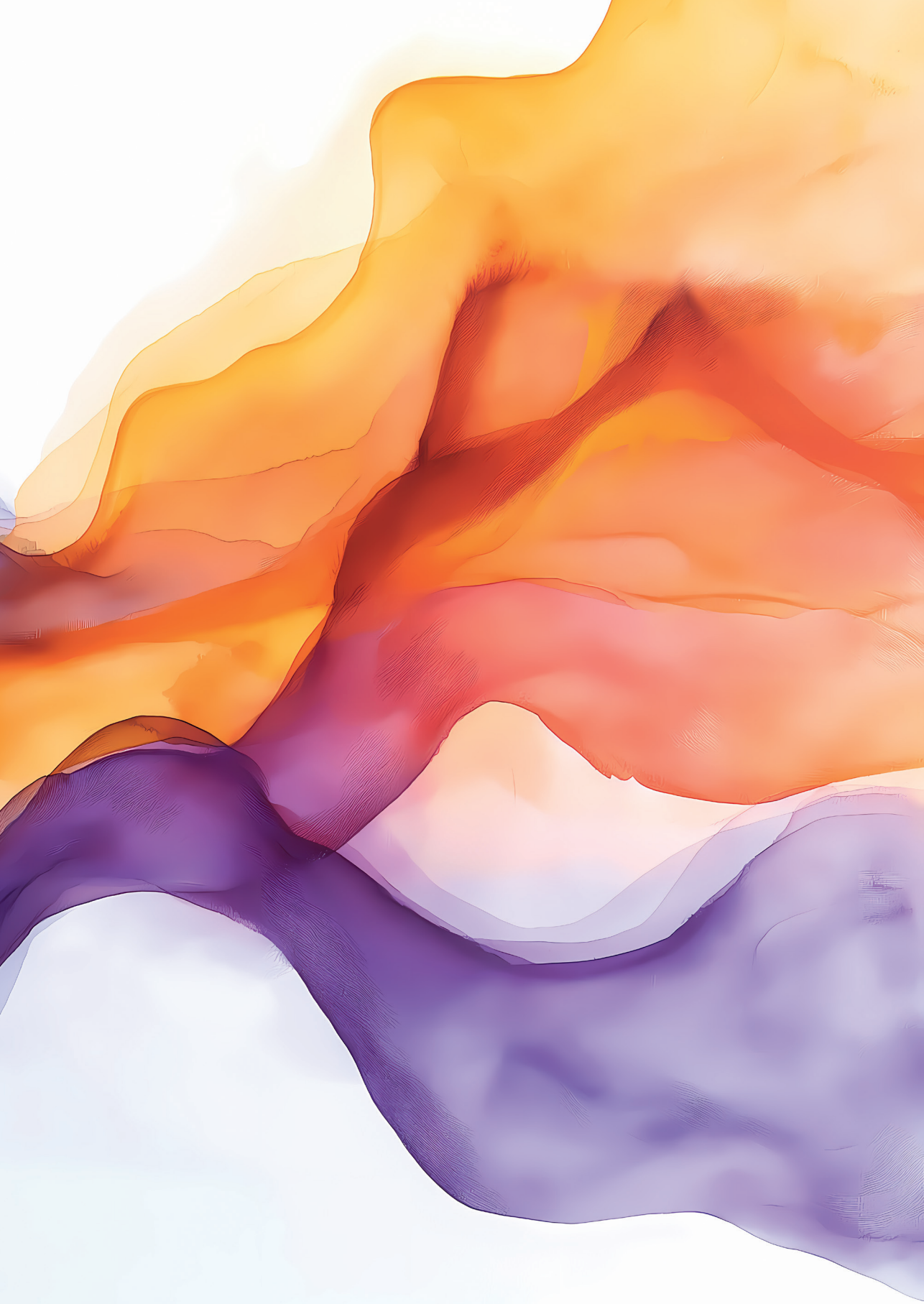


REFERENCES

- (1) Schmucker C, Schell LK, Portalupi S, Oeller P, Cabrera L, Bassler D, et al. Extent of Non-Publication in Cohorts of Studies Approved by Research Ethics Committees or Included in Trial Registries. *PLOS ONE*. 2014;9(12):e114023.
- (2) Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*. 2010;340:c365.
- (3) Liberati A. An unfinished trip through uncertainties. *BMJ*. 2004;328(7438):531.
- (4) Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev*. 2009;2009(1):Mr000006.
- (5) Moorthy VS, Karam G, Vannice KS, Kieny M-P. Rationale for WHO's New Position Calling for Prompt Reporting and Public Disclosure of Interventional Clinical Trial Results. *PLOS Medicine*. 2015;12(4):e1001819.
- (6) WMA Declaration of Helsinki - Ethical principles for medical research involving human subjects. Sect. 35.-36. (2013).
- (7) World Health Organization. WHO Statement on public disclosure of clinical trial results, 2015 [updated 9 April 2015 Available from: <https://www.who.int/news/item/09-04-2015-japan-primary-registries-network>.
- (8) U.S. Government. Food and Drug Administration Amendments Act of 2007: US Government Printing Office; 2007.
- (9) Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov — The Final Rule. *New England Journal of Medicine*. 2016;375(20):1998-2004.
- (10) European Parliament and the Council of the European Union. Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC, [Regulation]. *Official Journal of the European Union*; 2014 [updated May 27,2014; cited 2024 April 5]. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0536>.
- (11) Chan A-W, Karam G, Pymiento J, Askie LM, da Silva LR, Aymé S, et al. Reporting summary results in clinical trial registries: updated guidance from WHO. *The Lancet Global Health*. 2025;13(4):e759-e68.
- (12) International Clinical Trials Registry Plattform. Joint statement on public disclosure of results from clinical trials,. In: Research for Health, editor.: World Health Organisation; 2017.
- (13) Strengthening clinical trials to provide high-quality evidence on health interventions and to improve research quality and coordination,, WHA 75.8. Sect. 75.8 (2022).
- (14) International Clinical Trials Registry Plattform. Open consultation(Closed): WHO ICTRP guidance for reporting summary results in clinical trial registries.; World Health Organization;; 2022 [cited 2024 April 5]. Available from: <https://www.who.int/clinical-trials-registry-platform/reporting-on-findings>.
- (15) Chen R, Desai NR, Ross JS, Zhang W, Chau KH, Wayda B, et al. Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers. *BMJ*. 2016;352:i637.
- (16) DeVito NJ, Bacon S, Goldacre B. Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. *The Lancet*. 2020;395(10221):361-9.
- (17) Alayche M, Cobey KD, Ng JY, Ardern CL, Khan KM, Chan A-W, et al. Evaluating prospective study registration and result reporting of trials conducted in Canada from 2009 to 2019. *FACETS*. 2023;8:1-10.

- (18) DeVito NJ, Morley J, Smith JA, Drysdale H, Goldacre B, Heneghan C. Availability of results of clinical trials registered on EU Clinical Trials Register: cross sectional audit study. *BMJ Medicine*. 2024;3(1):e000738.
- (19) Hartung DM, Zarin DA, Guise JM, McDonagh M, Paynter R, Helfand M. Reporting discrepancies between the ClinicalTrials.gov results database and peer-reviewed publications. *Ann Intern Med*. 2014;160(7):477-83.
- (20) Tang E, Ravaud P, Riveros C, Perrodeau E, Dechartres A. Comparison of serious adverse events posted at ClinicalTrials.gov and published in corresponding journal articles. *BMC Med*. 2015;13:189.
- (21) Sun LW, Lee DJ, Collins JA, Carll TC, Ramahi K, Sandy SJ, et al. Assessment of Consistency Between Peer-Reviewed Publications and Clinical Trial Registries. *JAMA Ophthalmol*. 2019;137(5):552-6.
- (22) van Lent M, Int'Hout J, Out HJ. Differences between information in registries and articles did not influence publication acceptance. *Journal of Clinical Epidemiology*. 2015;68(9):1059-67.
- (23) World Health Organization. International Clinical Trials Registry Platform (ICTRP), 2022 [cited 2022 29 August]. Available from: <https://trialsearch.who.int/>.
- (24) Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Bmj*. 2010;340:c332.
- (25) Posit team. RStudio: Integrated Development Environment for R, Boston, MA, USA: Posit Software; 2024 [Available from: <http://www.posit.co/>].
- (26) Castor. Castor Software, 2024 [cited 2024 12 March]. Available from: <https://www.castoredc.com/>.
- (27) Showell M, Buckman S, Berber S, Ata Allah N, Patterson B, Cole S, et al. Publication bias in trials registered in the Australian New Zealand Clinical Trials Registry: Is it a problem? A cross-sectional study. *PLOS ONE*. 2023;18(1):e0279926.
- (28) Microsoft Corporation. Microsoft Excel (Version 16.68) [Software]. 2022 [cited 2023 26 January]. Available from: <https://office.microsoft.com/excel>.
- (29) Overheid.nl Wettenbank. Medical Research Involving Human Subjects Act (WMO) 2024 [cited 2024 November 13].
- (30) International Committee of Medical Journal Editors (ICMJE). Clinical Trials [Available from: <https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html>].
- (31) World Health Organization. WHO Registry criteria: Registry criteria (Version 2.1, April 2009), 2009 [cited 2024 November 14]. Available from: <https://www.who.int/clinical-trials-registry-platform/network/registry-criteria>.
- (32) World Health Organization. WHO Trial Registration Data Set (Version 1.3.1) 2024 [cited 2024 12 March]. Available from: <https://www.who.int/clinical-trials-registry-platform/network/who-data-set>.
- (33) Altman LK. The Ingelfinger rule, embargoes, and journal peer review--Part 1. *Lancet*. 1996;347(9012):1382-6.
- (34) Journal of Clinical Epidemiology. Clinical trial results, 2024 [Available from: <https://www.jclinepi.com/content/authorinfo#clinical-trial-results-1-15>].
- (35) The Lancet. Publishing excellence 2024 [Available from: <https://www.thelancet.com/publishing-excellence>].
- (36) BMJ Journals. Policies 2024 [Available from: <https://authors.bmj.com/policies/>].
- (37) Lo B, DeMets DL. Incentives for Clinical Trialists to Share Data. *New England Journal of Medicine*. 2016;375(12):1112-5.
- (38) ISRCTN Registry. FAQs: About ISRCTN 2025 [cited 2025 11 February]. Available from: <https://www.isrctn.com/page/faqs#finding-study-information>.

- (39) Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet*. 2009;374(9683):86-9.
- (40) Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ*. 2018;363:k4645.
- (41) Shah RF, Bini S, Vail T. Data for registry and quality review can be retrospectively collected using natural language processing from unstructured charts of arthroplasty patients. *The Bone & Joint Journal*. 2020;102-B(7 Supple B):99-104.
- (42) Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Res Synth Methods*. 2024;15(4):576-89.
- (43) McKenzie JE, Brennan SE. Synthesizing and presenting findings using other methods. *Cochrane Handbook for Systematic Reviews of Interventions* 2019. p. 321-47.
- (44) Lefebvre C GJ, Briscoe S, Featherstone R, Littlewood A, Marshall C, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland LS. Chapter 4: Searching for and selecting studies. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA . editor. *Cochrane Handbook for Systematic Reviews of Interventions* version 63. Cochrane: Cochrane; 2022.
- (45) Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA*. 2009;302(9):977-84.
- (46) Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. *BMC Medicine*. 2015;13(1):282.



4

Development of Prognostic Models for In-Hospital Mortality and ICU Admission in Confirmed COVID-19 Patients: An IPD Meta-Analysis from LMICs

Tabea Kaul, Jasper van den Brink, Valentijn MT de Jong, Silvia Bertagnolio,
Janet V Diaz, Ronaldo Silva, Soe Soe Thwin, Laure Wynants, Maarten van Smeden,
Karel GM Moons, Johanna AA Damen

Manuscript in preparation

ABSTRACT

Background: The COVID-19 pandemic especially burdened low- and lower middle-income countries (LMICs). Clinical prediction models support clinical decision-making by calculating an individual risk estimate for a COVID-19 patient to fall critically ill (i.e., to die in hospital or to need admission to the intensive care unit [ICU]). However, none of the currently developed prognostic models perform well in LMICs.

Study Design and Objective: Model development study using individual participant data to develop two clinical prediction models for adult confirmed COVID-19 patients admitted to the hospital to assess the risk of dying in hospital or the need for ICU admission. We used data from the WHO COVID-19 database, including data from LMICs. Models were developed using multiple logistic regression. Candidate predictors comprised demographics, comorbidities, and symptoms. We evaluated model generalizability by performing internal-external cross validation (IECV), assessing calibration (slope, intercept, plot), and discrimination (c-statistic). Lastly, we refitted the models in individual countries.

Results: The development data included 12 countries (n=15,876 patients), including 1,938 (12.2%) mortality events, and 275 (1.7%) ICU admissions. Both models included the following predictors: age, sex, HIV, heart rate, respiratory rate, confusion, and body temperature. The in-hospital mortality model additionally included presence of tuberculosis or pneumonia, chronic kidney disease, diabetes, dyspnea, requiring oxygen therapy and AVPU score; the ICU model also included smoking, presence of cough, sore throat, and systolic blood pressure. We identified heterogeneity across countries in model performance measures. Both models performed well but showed some signs of miscalibration and overfitting (in-hospital mortality model c-statistic: 0.86 [95%-CI: 0.82;0.89], ICU admission model: c-statistic: 0.74 [95%-CI: 0.69;0.79]).

Conclusion: The two developed models provide ideal starting points for further validation in specific LMIC settings. We strongly advise externally validating them before implementation into the clinical care pathway.

INTRODUCTION

On May 5th, 2023, the COVID-19 pandemic officially ended (1). Nevertheless, COVID-19 continues to impact global health and will likely continue to do so in the future (2, 3). Low- and lower-middle income- countries (LMICs) were severely impacted by the COVID-19 pandemic (4-6). Between January 2020 and November 2024, LMICs alone reported over 1 million COVID-19 related deaths to the World Health Organization (WHO) compared to over 3 million in high-income countries (6, 7). This number is likely an underestimation, as many countries suspended the reporting of COVID-19 deaths to the WHO (2). Furthermore, the difference in the number of COVID-19 related deaths between LMICs and high-income countries might be due to several reasons such as variations in comorbidities and environmental factors (4, 8). Besides, healthcare systems of LMICs and high-income countries differ in resources, capacity availability, vaccination, and disease management (4, 9-11).

Several measures can reduce the burden of COVID-19 on healthcare systems. These include the expansion of capacities (e.g., staff) or the implementation of public health measures (e.g., travel restrictions) (12, 13). Also, clinical prediction models that predict a patient's risk of disease progression and/or mortality (i.e., their prognosis) may be introduced as decision-aids within the clinical care pathway to support medical staff during the decision-making process regarding the treatment of a patient (14).

Clinical prediction models have been developed for COVID-19 to predict infection, mortality, or disease progression (14-16). Clinical prediction models combine several different individual characteristics to estimate the risk of an individual of a certain health outcome, either being present (i.e., diagnosis) or developing in the future (i.e., prognosis). Unfortunately, most clinical prediction models for COVID-19 lack methodological quality, for example due to small sample size, overfitting, or incomplete evaluation of performance measures (14, 16). Furthermore, many of these models have not been validated in data other than the data they were developed on, and it is thus unclear whether they can be applied to different populations or settings (14-16).

Few prognostic clinical prediction models exist that have been developed and/or validated for COVID-19 patients in the LMIC setting (14, 16). In their validation study, Damen et al. identified 144 prognostic models of hospitalized COVID-19 patients in the LMIC setting. However, only four of the 144 models could be validated due to unavailability of predictors, lacking description of the full model or poor methodological quality (i.e., high risk of bias). Furthermore, the four validated models showed issues of miscalibration, overfitting, and heterogeneity between countries and thus cannot be

directly applied in practice without appropriate model updating (16). This shows the need for better performing models applicable to the LMIC setting.

To support the decision-making process of healthcare providers in LMICs in the management of adult patients with confirmed COVID-19 admitted to the hospital at first point of contact, we developed two prognostic prediction models to 1) predict in-hospital mortality and 2) predict admission to the intensive care unit (ICU) due to disease progression during hospital stay. We combined data from several different countries in an individual participant data (IPD) meta-analysis.

METHODS

This study is reported in accordance with the TRIPOD-Cluster and the TRIPOD+AI reporting guidelines (17, 18). We did not prepare a publishable protocol.

Data

This study used IPD from 34,216 patients obtained from the WHO Global Clinical Platform database for COVID-19 based on a standardized Case Report Form (CRF) (19, 20). This platform provides WHO member states with a platform to collect patient and clinical data of all COVID-19 patients worldwide. The WHO CRF has four modules, the first was completed on admission to the health centre (within 24 hours), the second was completed on ICU admission or ICU transfer (within 24 hours), the third was completed at discharge or death, and the fourth was completed if pregnant at admission or less than/equal to 21 days from pregnancy outcome (20). The WHO Core CRF contains sets of variables such as demographics, vaccination status, co-morbidities (e.g., hypertension), symptoms, medication use, vital signs, and outcomes (e.g., in-hospital mortality, transfer to other facility, palliative discharge, neonatal outcomes: discharged with complications/sequelae) (20).

Participants

We included adults (age 18 or older) admitted to a hospital with confirmed COVID-19 between January 3rd, 2020, and September 30th, 2021, in low- and lower-middle income countries (as defined by The World Bank (7)). Therefore, countries included: Bangladesh, Bolivia (Plurinational State of), Burkina Faso, Cameroon, Democratic Republic of Congo, Gambia, Ghana, Guinea, India, Indonesia, Malawi, Niger, Nigeria, Pakistan, Philippines, Sudan, Ukraine, Zambia, and Zimbabwe. We excluded participants in case of an unknown or negative COVID-19 test result. Also, we excluded participants from both

model development outcomes (i.e., in-hospital mortality and ICU admission) if they were directly admitted to the ICU upon hospital entry without being seen at the emergency department or at triage due to a significantly different prognosis (i.e., higher disease severity) compared to other participants of the sample. Finally, if countries included less than 100 participants, we removed them based on sample size.

Outcomes

We developed one model for the outcome in-hospital mortality and another for ICU admission in adults admitted to the hospital for COVID-19. COVID-19 mortality was defined as any patient death after admission to the hospital due to COVID-19. Patients discharged or transferred to another facility were classified as 'Alive'.

The need for ICU admission was assessed after patient admission to the hospital for COVID-19. This outcome definition does not include any patients that were moved to the ICU directly upon admission to the hospital (i.e., at T0, without screening at emergency department or triage).

Both outcomes did not require subjective interpretation or blind assessment.

Predictors

Candidate predictors were chosen according to previous models, literature, and availability in the dataset. We excluded predictors that were not included in any published COVID-19 prediction model (14, 16, 21). The predictors consisted of demographic variables such as age and sex, physiological variables such as blood pressure, comorbidities, symptoms, and blood values that (if available) were measured upon admission (first point of contact with hospital). Since predictor data includes data collected from over 800 participating centres (22), we cannot confirm blind assessment or have information of the demographic characteristics of the predictor assessors. All 96 candidate predictors are shown in the supplement (Section A).

We reduced the number of predictors before model building. We excluded potential predictors with more than 50% missing values and based on prevalence (i.e., <100 participants for one category in categorical variables). If predictors were related (e.g., fever, temperature), we selected one of these. In this situation, continuous variables with low percentages of missing values were preferred over dichotomous variables and variables with higher percentages of missing values. We combined the variables 'loss of smell' and 'loss of taste' since they are reported together in the literature (23). The supplement (Section A) includes the reason for exclusion per predictor.

Lastly, we performed data transformations on the continuous variables based on histograms, Q-Q plots, and comparing models with transformed and non-transformed variables based on Bayesian Information Criterion (BIC, i.e. lower BIC indicating better fit). Based on this, a log transformation was applied to age, respiratory rate, and body temperature for the outcome in-hospital mortality, and for age and respiratory rate for the outcome ICU admission.

All data preparatory steps were similar across sociodemographic groups.

Sample size

To estimate the required sample size, we used the method described by Riley et al. (24), assuming a shrinkage factor of 0.9. For the sample size calculation of the outcome in-hospital mortality, we used the lowest c-statistic (0.73) of the studies identified in a validation study (16, 25), assuming an outcome occurrence of in-hospital mortality of 12.2% (as found in our dataset). The sample size required, based on a broad estimation of 26 predictors (28 predictor parameters), was then calculated to be 3,282 with 401 events corresponding to an Events Per Parameter (EPP) of 14.3.

For ICU admission, we used a c-statistic of 0.78 (16, 26), assuming an outcome occurrence of 1.7% (as found in our dataset) and a broad estimation of 26 predictors (28 predictor parameters). Thus, the sample size required for the ICU admission outcome was 13,123, including 224 events, and an EPP of 7.97.

Missing data

We assumed missing data were missing at random (MAR), and imputed these 50 times using joint modelling imputation (JOMO) with a burn-in phase of 20,000 iterations and 1,000 iterations between imputations, on 10 imputations each (thus, we performed 5 separate imputation cycles using different seeds). The imputation models included all candidate predictors (except for those previously excluded), non-linear effects and outcomes. Finally, clustering was addressed in the imputation models by using country as a cluster variable.

Analytical methods

All descriptive and statistical analyses were conducted using R Studio (R version 4.4.1 [2024-06-14]) (27). Baseline characteristics were reported for categorical variables as counts and percentages (%), and as median and interquartile range (IQR) for continuous variables.

Model development and performance

We developed two prognostic models (one per outcome) using multivariable logistic regression. Initially, a full model containing all previously selected predictors was fitted. To account for clustering in the model, the country variable was used as a cluster variable. Subsequently, predictors were removed from the model using backwards selection based on the BIC. The model with the lowest BIC was selected until no further improvements could be made. Predictor selection was performed in each imputed dataset separately and predictors were included in the final model if they were selected in the majority (>25) of imputed datasets.

We used internal-external cross validation (IECV) to explore generalizability of the models, which involves fitting them on all clusters except one and externally validating them on the excluded cluster (28, 29). This process was continued until each cluster had been excluded once. We used country as a cluster variable and performed IECV on each imputed dataset. Rubin's rule was used to pool performance measures, after applying a logit transformation to the c-statistic (30). Model performance in each country was presented as forest plots. We evaluated discrimination using the c-statistic and assessed calibration using the calibration intercept and slope. We stacked the individual probabilities per patient to create the calibration plot.

Models fitted per country

For both outcomes, we refitted the developed models separately for each country. For this sample, we only included countries with a specific sample size. We used the c-statistics, and the predictor parameters derived from the IECV of the model developments for this calculation. For in-hospital mortality, this included 16 predictor parameters, outcome occurrence of 12.2%, and a c-statistic of 0.872, necessitating a sample size of 608 with 75 events and an EPP of 4.64. For ICU admission, we based our calculation on 12 predictor parameters, an outcome occurrence of 1.7%, and a c-statistic of 0.751, amounting to a minimum sample size required of 7338, including 125 events, and an EPP of 10.4.

For countries meeting these sample size requirements, we trained the developed models to their specific setting and evaluated their apparent performance using discrimination (c-statistic) and calibration (calibration plot).

Fairness

The developed prediction models can help to predict in-hospital mortality and ICU admission in an objective manner. We did not take any additional steps to address fairness.

Model output

The model output describes the individual probability of a confirmed COVID-19 patient to either die in hospital (model 1) or be admitted to ICU (model 2).

Patient and public involvement

This study did not include any patient and public involvement.

Ethical approval

This study followed the Declaration of Helsinki and used data from the WHO Global Clinical Platform for COVID-19, comprising over 800 centres. Each centre obtained ethics approval and participant consent. Data collection complied with all relevant legal, ethical, and medical standards. More details are available in the WHO terms of use (22).

RESULTS

Patient flow and characteristics

The initial dataset included 34,216 participants from 19 countries. After the exclusion of patients based on age, COVID-19 status, admission to ICU, country sample size, and number of outcome events per country, the final dataset encompassed 15,876 patients from 12 countries with 1938 in-hospital mortality and 275 ICU admissions events (Figure 1, Table 1). Therefore, our dataset met the sample size requirements for both global models.

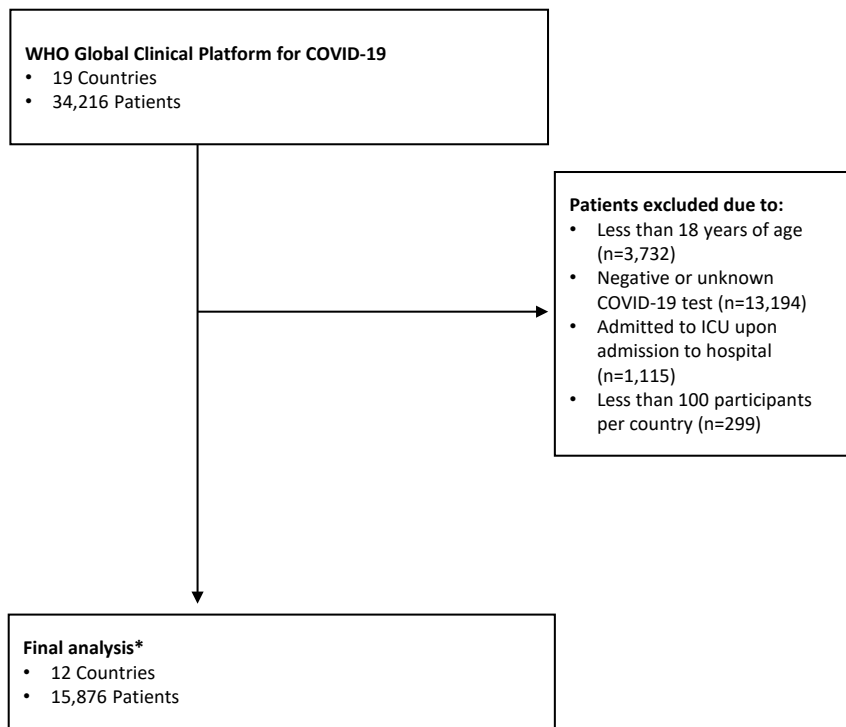


Figure 1 Participant flow

*Data from Burkina Faso did not include any ICU admission events. Thus, Burkina Faso was excluded for model development of the outcome ICU admission.

ICU, Intensive care unit

Table 1 describes the baseline characteristics of the sample. Most patients were from Nigeria (23.4%) and the least from Bolivia (0.9%). Participants were primarily male (53.6%), with a median age of 45 years (IQR: 32.0, 61.0). However, five out of 12 country samples included more females than males; participants from India were mostly females (84.0%) and of younger age (median age, years: 28.0, IQR: 24.0, 34.0) (Supplement, Section B). A total of 1938 (12.2%) patients died in-hospital, while 275 (1.7%) patients were admitted to the ICU. The most prevalent comorbidity was diabetes (13.5%). The most prevalent symptoms included cough (48.4%) and dyspnea (29.1%). Most predictors had more than 15% missing values except for sex (0.4%), diabetes (13.5%), cough (11.3%), and dyspnea (13.3%). Across countries, percentages of missing values in candidate predictors varied, with some, such as Cameroon and Guinea including systematic missingness (Supplement, Section B). Data from Burkina Faso was

excluded from the ICU model development, due to lack of ICU admissions throughout that sample. The data did not allow for evaluation of follow-up time.

Table 1 Baseline characteristics

Variables	Number (%)	Missing n (%)
Total sample, N	15,876	
<i>Country (%)</i>		
Bangladesh	720 (4.5)	
Bolivia (Plurinational State of)	140 (0.9)	
Burkina Faso	351 (2.2)	
Cameroon	2354 (14.8)	
Democratic Republic of Congo	195 (1.2)	
Guinea	1264 (8.0)	
India	413 (2.6)	
Malawi	2794 (17.6)	
Nigeria	3714 (23.4)	
Philippines	362 (2.3)	
Ukraine	210 (1.3)	
Zimbabwe	3359 (21.2)	
<i>General characteristics</i>		
Age ^a , years	45.0 (32.0, 61.0)	
Sex (%)		56 (0.4)
Male	8506 (53.6)	
Female	7314 (46.1)	
<i>Comorbidities</i>		
Coronary Heart Disease (%)	614 (3.9)	3307 (20.8)
Diabetes (%)	2136 (13.5)	2151 (13.5)
Smoking (%)	367 (2.3)	3843 (24.2)
Chronic Pulmonary Disease (%)	261 (1.6)	2675 (16.8)
Tuberculosis or Pneumonia (%)	280 (1.8)	3272 (20.6)
Asthma (%)	532 (3.4)	2516 (15.8)
Chronic Kidney Disease (%)	331 (2.1)	3152 (19.9)
HIV (%)	1001 (6.3)	4851 (30.6)
<i>Symptoms/clinical presentation at admission</i>		
Body Temperature ^a , °C	36.7 (36.3, 37.2)	3122 (19.7)
Heart Rate ^a , beats/min	90.0 (80.0, 101.0)	4053 (25.5)
Respiratory Rate ^a , breaths/min	21.0 (18.0, 24.0)	6040 (38.0)
Systolic Blood Pressure ^a , mmHg	125.0 (113.0, 140.0)	3692 (23.3)

Table 1 Baseline characteristics (*continued*).

Variables	Number (%)	Missing n (%)
AVPU Score (%)		5988 (37.7)
1	8782 (55.3)	
2	545 (3.4)	
3	444 (2.8)	
4	117 (0.7)	
Cough (%)	7684 (48.4)	1796 (11.3)
Confusion (%)	622 (3.9)	3118 (19.6)
Sore Throat (%)	2139 (13.5)	2657 (16.7)
Vomiting (%)	1144 (7.2)	2707 (17.1)
Diarrhea (%)	832 (5.2)	2683 (16.9)
Chest Pain (%)	2967 (18.7)	2931 (18.5)
Myalgia (%)	1820 (11.5)	2768 (17.4)
Fatigue (%)	4039 (25.4)	2709 (17.1)
Dyspnea (%)	4624 (29.1)	2111 (13.3)
Requiring Oxygen Therapy (%)	3460 (21.8)	5017 (31.6)
Loss of Taste or Smell (%)	2423 (15.3)	3282 (20.7)
<i>Outcomes</i>		
Deceased (%)	1938 (12.2)	752 (4.7)
Need for ICU admission (%)	275 (1.7)	5797 (36.5)

^a Median (Interquartile range [IQR])

AVPU, Alert Verbal Pain Unresponsive; assessment scale to measure a patient's level of consciousness; Score = 1 Alert, Score = 2 Voice responsive, Score 3= Pain responsive, Score 4 = Unresponsive.

ICU, intensive care unit

Final model

The supplement includes full model equations for predicting in-hospital mortality and ICU admission to enable external validation (Section C). The final model for in-hospital mortality included 13 predictors: age, sex, HIV, tuberculosis or pneumonia, chronic kidney disease, diabetes, AVPU scores, heart rate, respiratory rate, dyspnea, requiring oxygen therapy, confusion, and body temperature (Table 2). Results of the model for ICU admission are part of the supplement (Section D).

Table 2 Beta-coefficients and odds ratios with 95% confidence interval of each predictor included in the final model with the outcome in-hospital mortality.

	Beta coefficients (95% CI)	Odds ratio (95%-CI)
(Intercept)	-11.37 (-17.07--5.68)	-
Age*	1.62 (1.26–1.98)	5.04 (3.51–7.23)
Sex	-0.28 (-0.40--0.15)	0.76 (0.67–0.86)
HIV	0.35 (0.11–0.59)	1.43 (1.12–1.81)
Tuberculosis or Pneumonia	0.37 (-0.00–0.73)	1.44 (1.00–2.09)
Chronic Kidney Disease	0.71 (0.34–1.07)	2.03 (1.41–2.92)
Diabetes	0.32 (0.14–0.51)	1.38 (1.15–1.66)
AVPU Score = 2	0.76 (0.48–1.03)	2.13 (1.62–2.80)
AVPU Score = 3	1.54 (1.08–2.00)	4.67 (2.96–7.36)
AVPU Score = 4	1.43 (0.92–1.95)	4.20 (2.50–7.05)
Heart Rate ^a , beats/min	0.01 (0.00–0.01)	1.01 (1.00–1.01)
Respiratory Rate*, breaths/min	0.45 (-0.25–1.16)	1.58 (0.78–3.18)
Dyspnea	0.83 (0.65–1.00)	2.29 (1.92–2.72)
Requiring Oxygen Therapy	1.02 (0.80–1.24)	2.77 (2.23–3.44)
Confusion	1.07 (0.82–1.32)	2.91 (2.26–3.73)
Body Temperature, °C*	-0.31 (-1.66–1.04)	0.73 (0.19–2.84)

*log transformed before model development

Internal-external cross validation

Model discrimination (c-statistic), calibration slope, and intercept from the IECV of the in-hospital mortality model are shown in Figures 2 and 3 (29). Results of the ICU admission model are part of the supplement (Section D).

Overall, for the model for in-hospital mortality (Figure 3), the calibration slope and intercept showed values close to the ideal values (calibration intercept: -0.004 (95%-CI: -0.92; 0.92); calibration slope: 1.10 (95%-CI: 0.88;1.33)). For risk values > 0.1, the model showed moderate underestimation of the risk. There was heterogeneity in calibration between countries. Calibration slope estimates ranged from 0.71 to 2.21 and show heterogeneity and wider confidence intervals compared to the c-statistics. The calibration intercepts ranged from -2.98 to 2.73, probably showing some variation in baseline risks between regions. Overall risk was underestimated in Bolivia (Calibration intercept: 2.73 [2.36;3.11]) and overestimated in Bangladesh (Calibration intercept: -2.98 [-3.60;2.37]). The c-statistic was 0.86 (95%-CI: 0.82;0.89), but there was heterogeneity across countries, ranging from 0.77 to 0.92. India showed the largest amount of uncertainty around the point estimate (0.11;0.99) because of the low number of events.

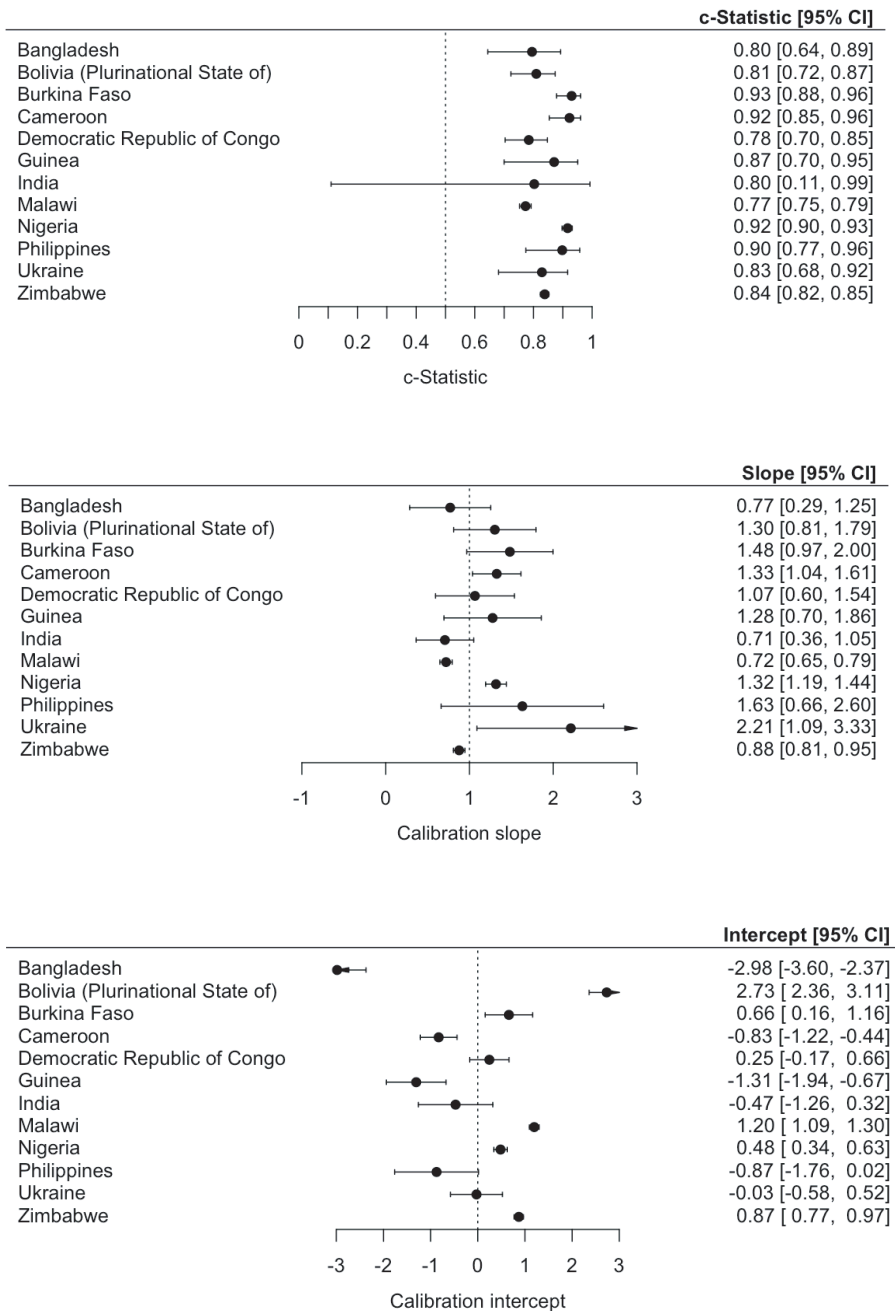


Figure 2. Forest plots of internal-external cross validation across countries (in-hospital mortality).

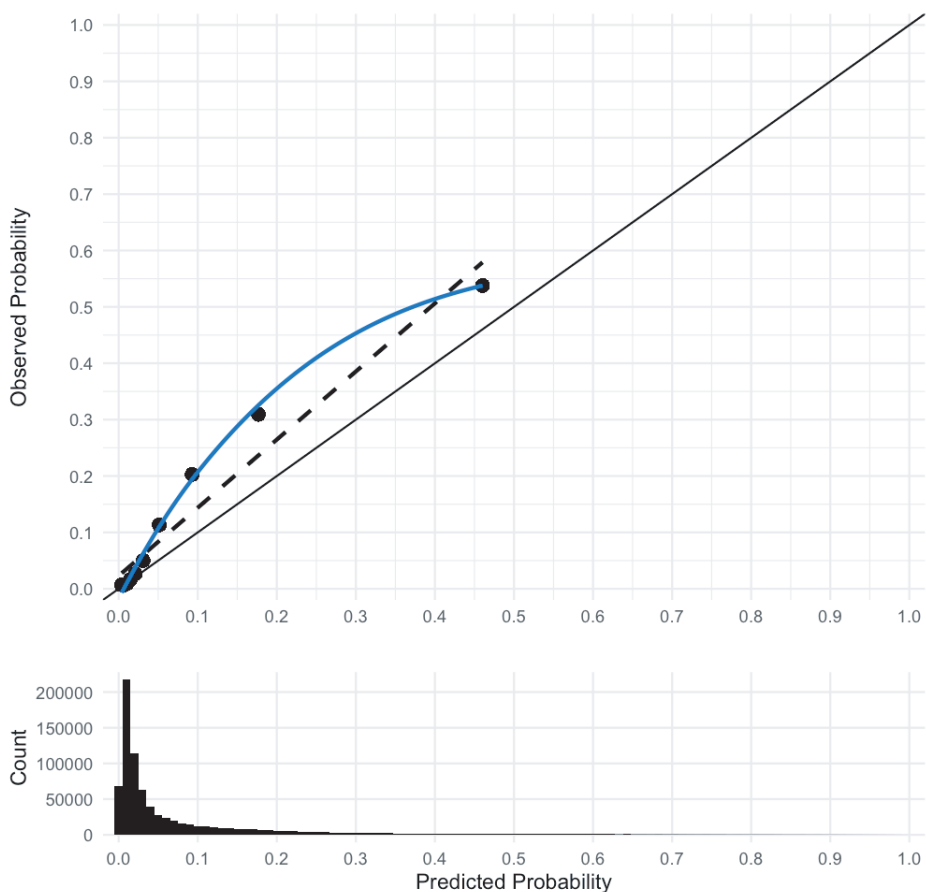


Figure 3. Calibration plot internal-external cross validation across countries (in-hospital mortality). Discrimination: c-statistic: 0.86 (95%-CI: 0.82;0.89); Calibration intercept: -0.004(95%-CI: -0.92; 0.92); Calibration slope: 1.10 (95%-CI: 0.88;1.33)

Models fitted per country

Because of the high heterogeneity between countries, we refitted the in-hospital mortality model on data from Malawi, Nigeria, and Zimbabwe, which had a sufficient number of participants and events according to sample size requirements (Supplement, Section E). The c-statistic ranged from 0.789 (95%-CI: 0.786;0.792) in Malawi to 0.942 (95%-CI: 0.922; 0.926) in Nigeria, showing good discrimination. The models fit per countries and their respective calibration plots are part of the supplement (Section E). The number of ICU admission events was insufficient to enable refitting of the ICU model for any given country.

DISCUSSION

To our knowledge, this is the largest model development study to date on data from multiple low- and lower-middle income countries (LMICs) for predicting in-hospital mortality and ICU admission in adult COVID-19-confirmed patients. In our individual participant data meta-analysis (IPDMA) of 15,876 patients from 12 LMICs, we developed two prognostic models: one for in-hospital mortality, and one for ICU admission.

Internal-external cross validation (IECV) showed both models performed well on average but had signs of miscalibration for higher risks. The model for in-hospital mortality performed better in terms of calibration and discrimination compared to the model for ICU admission. Both models showed substantial heterogeneity across countries. We refitted the in-hospital mortality models to the three largest samples (Malawi, Nigeria, Zimbabwe) so that prognostic models tailored to patient characteristics are available in these countries. Due to sample size requirements, this was not possible for the ICU admission model. Both models are for use at the first point of hospital admission. However, external validation is strongly advised before integrating any of the developed models into the clinical care pathway.

Comparison to previous research

Previously developed COVID-19 prediction models cannot be adequately applied to the LMIC setting. While performing well, most require blood biomarker measurements as predictors which we found largely unavailable across LMICs, thus indicating that they are not routinely collected (25, 31-33). The LMIC-PRIEST study developed a clinical severity score for suspected COVID-19 patients in emergency departments across Western Cape, South Africa (34). This study of over 400,000 patients aimed to predict a 30-day composite outcome, including death and ICU admission. While targeting a different patient population (i.e. patients *suspected* of COVID-19), the selected variables largely overlap with ours. Unlike our models, LMIC-PRIEST includes oxygen saturation and inspired oxygen, which had high missing values in our dataset.

For in-hospital mortality, Zhang et al.'s model in the Chinese setting included a total of 12 predictors, aligning with predictors in our model with age, sex, chronic kidney disease, diabetes, and dyspnea (33). Zhang et al.'s model also included hypertension, chronic lung disease, cancer, cough, immunocompromised status, heart disease, and diarrhea (33). In our development study, we excluded cancer due to a low number of observations and hypertension due to comparability to other variables (i.e., systolic blood pressure) before model building (Supplement, Section A). The predictors chronic

pulmonary disease, heart disease, cough, and diarrhea were excluded during model building as these did not improve prediction of the outcome in our data, while the predictor immunocompromised status was not available in our dataset (Supplement, Section A). Another Chinese model by Wang et al. included only three predictors for the individual risk prediction of in-hospital mortality (32). It comprised age, hypertension, and coronary heart disease (32). The models by Zhang et al. and Wang et al. models both include fewer predictors; however, age is a predictor identified in all of our models and thus seems a highly relevant predictor in the context of COVID-19 related in-hospital mortality.

Our ICU admission model overestimated the risk of ICU admission need, likely due to the low number of outcome events. The model by Zhou et al. for ICU admission included fewer predictors than ours did in the Chinese setting, however, predictors largely overlap with ours (26). Furthermore, their model included chronic kidney disease and fever, while our model includes sex, HIV status, heart rate, cough, sore throat, confusion, and body temperature.

Interestingly, while both our models included the predictor HIV status, none of the previously discussed models did (26, 32-34). At least for countries with a known high prevalence of HIV (e.g., Malawi, Zimbabwe) (Supplement, Section B), this could be a possible explanation why previously developed models performed less well in the LMIC setting.

Developing clinical prediction models that are applicable to various populations remains challenging. We identified great amounts of heterogeneity between the different countries for both models. Similar observations have been described by others when validating COVID-19 prognostic models across different clusters (16, 35). This highlights the need to tailor and fit developed models to the intended use setting. Our developed models present ideal starting points for external validation. Their full model equations (including coefficients) are part of the supplement (Section C) to support validation and subsequent implementation efforts.

Strengths and limitations

A major strength of our study is the large sample size: we developed the models on data of over 15,000 participants from across 12 countries and various geographical settings. Further, this was the first multi-country study to develop COVID-19 prognostic models for the LMIC setting. We adhered to TRIPOD-Cluster and TRIPOD+AI standards, followed guidance specific to the IPDMA setting, and performed multiple imputation to account for any missing data (17, 18, 29).

Nevertheless, our study had several limitations. First, we identified heterogeneity for both models across various countries, which affects the generalizability of the models. The application of the models in different populations could provide significantly different results and should therefore be externally validated before implementation of the models into practice. To address this heterogeneity, we fitted the in-hospital mortality model to the three largest samples (i.e. Malawi, Nigeria, Zimbabwe) and showed excellent apparent performance (as expected). However, we did not perform internal validation, and an excellent apparent performance does not circumvent the need for external validation.

Second, outcome occurrence varied between countries. This was particularly problematic for the ICU admission outcome, which was infrequent in the dataset and did not meet the sample size requirements to allow for model fitting per country (24). Admission to the ICU is a challenging outcome: access to critical care lacks data for LMICs, and where the data are available, ICU availabilities are sparse (36). For example, Africa has an estimated 0.8 ICU beds per 100,000 population, compared to 11.5 per 100,000 population in Europe (37, 38). This further illustrates the disparities in healthcare between high-income countries and LMICs. Therefore, extra caution should be placed before implementing the ICU model into practice.

Besides, outcome occurrence variation was also an issue for the in-hospital mortality model, e.g., India and Cameroon showed low occurrence of in-hospital mortality. This may subsequently affect the model performance in these countries and their certainty around each performance estimate.

Further, a considerable proportion of the data had missing values. Before model development, we had to remove a large proportion of predictors, i.e., those with more than 50% missing values (e.g., blood measurements). This may indicate that these values are not routinely recorded in the setting of the included LMICs. Still, among the remaining predictors, many had more than 15% missing values. While we imputed these data with multiple imputation, the occurrence of missing values introduces uncertainty.

Lastly, the models do not yet include any information on recent advances in COVID-19 disease, such as virus variants and availability of vaccines. Nevertheless, the coverage of vaccination varies per country and may not be as accessible in the LMIC setting compared to other settings (10, 11).

Implications

Considering the severe impact the COVID-19 pandemic had on already constrained healthcare systems in LMICs, it is important to proactively strengthen systems to

prevent future disruptions (5, 39). Integrating clinical prediction models into the clinical care pathway may be an important step to support faster clinical decision-making. To facilitate this integration, our two developed models for in-hospital mortality and ICU admission serve as ideal starting points for further validation and integration of prediction models in the LMIC setting. However, the identified heterogeneity across different LMICs indicates that not one single model can be applied to the different settings. Therefore, we advise thorough validation and updating of our developed models to the specific populations and settings. Given the rapidly changing landscape of treatment advances (e.g., vaccine rollouts, drug developments), one might even consider integrating continuous updating and monitoring of clinical prediction models in such cases, e.g. pandemic situations, to ensure adequate risk estimation, especially with rapidly evolving viral variants, as in COVID-19 (40, 41).

CONCLUSION

We present two prognostic models for in-hospital mortality and ICU admission in LMICs for patients with confirmed COVID-19 at first hospital admission. The included predictors are routinely available. The model for in-hospital mortality has also been tailored to the available data from Malawi, Nigeria, and Zimbabwe. It is strongly recommended that all developed models are thoroughly externally validated prior to use and, if necessary, updated to fit the setting they are intended for.

Acknowledgements

None.

Authors' contributions

The study concept and design were conceived by TK, JvB, VdJ, SB, JVD, RS, SST, LW, MvS, KGMM, and JAAD. TK and JvB performed the analysis. JAAD verified the underlying data. TK and JvB wrote the original draft of the study. TK, JvB, VdJ, SB, JVD, RS, SST, LW, MvS, KGMM, and JAAD reviewed and revised the first draft critically for intellectual content. TK revised the draft which led to the final version included in this thesis. TK performed the project administration. JAAD supervised this project.

Funding

This research was funded by the World Health Organization.

Competing interests

The authors declare no competing interests.

Ethical approval

This study followed the Declaration of Helsinki and used data from the WHO Global Clinical Platform for COVID-19, comprising over 800 centres. Each centre obtained ethics approval and participant consent. Data collection complied with all relevant legal, ethical, and medical standards. More details are available in the WHO terms of use: https://cdn.who.int/media/docs/default-source/documents/emergencies/terms-of-use-global-covid19-data-platform.pdf?sfvrsn=d0a440c1_3.

Registration and protocol

This study was not registered and does not include a publishable protocol.

Data availability

Data will not be made available.

Supplementary data

Supplementary data to this article can be found online.

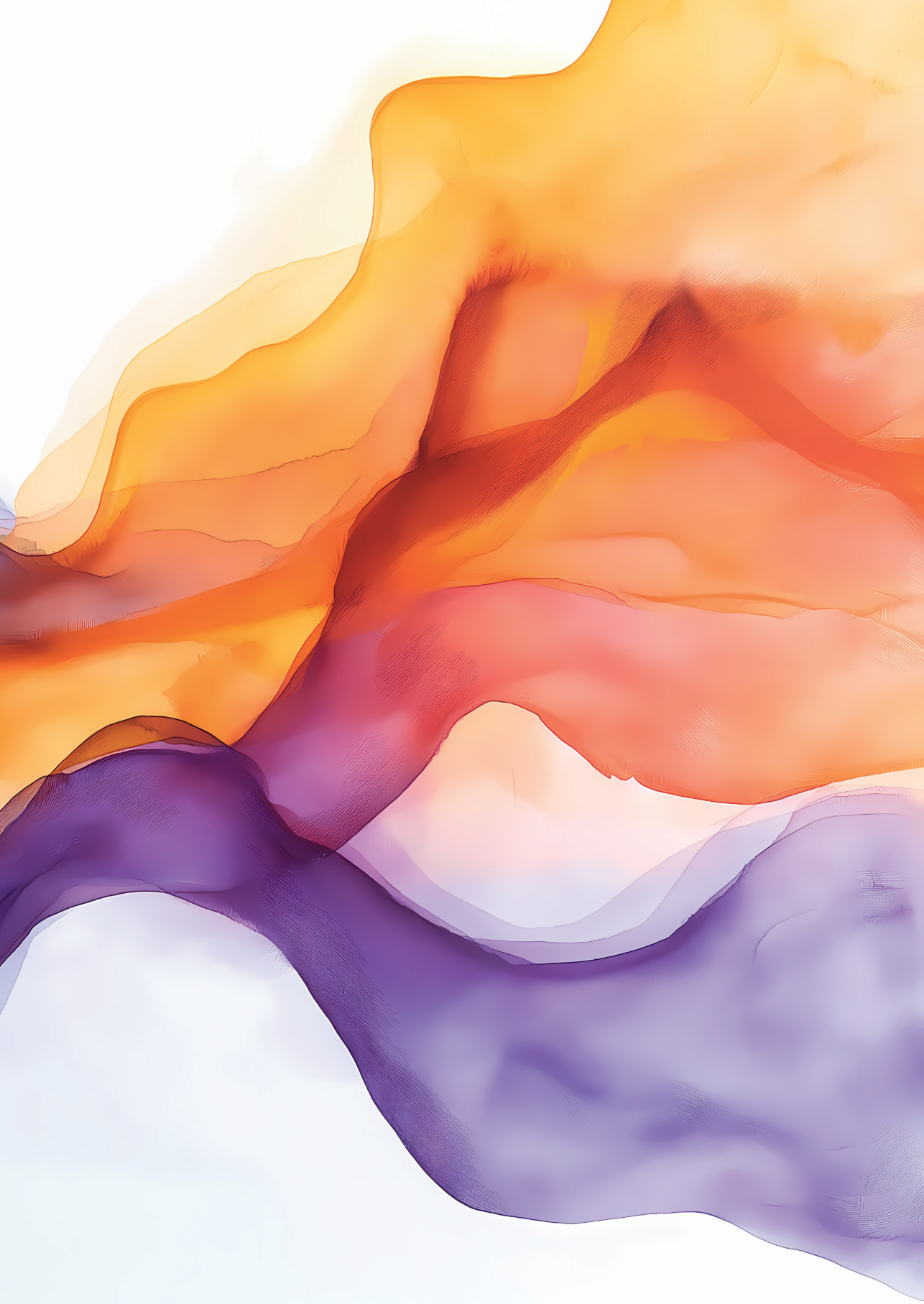


REFERENCES

- (1) WHO chief declares end to COVID-19 as a global health emergency [press release]. online: United Nations,, 05 May 2023 2023.
- (2) World Health Organization. COVID-19 epidemiological update – 6 November 2024 - Edition 173 Online: World Health Organization,; 2024 [updated 06 November 2024; cited 2024 02 December]. 173:[Available from: COVID-19 epidemiological update – 6 November 2024.
- (3) Telenti A, Arvin A, Corey L, Corti D, Diamond MS, García-Sastre A, et al. After the pandemic: perspectives on the future trajectory of COVID-19. *Nature*. 2021;596(7873):495-504.
- (4) Khedr A, Al Hennawi H, Rauf I, Khan MK, Mushtaq HA, Lodhi HS, et al. Differential mortality with COVID-19 and invasive mechanical ventilation between high-income and low-and middle-income countries: a systematic review, meta-analysis, and meta-regression. *Infez Med*. 2022;30(1):51-8.
- (5) Levin AT, Owusu-Boaitey N, Pugh S, Fosdick BK, Zwi AB, Malani A, et al. Assessing the burden of COVID-19 in developing countries: systematic review, meta-analysis and public policy implications. *BMJ Glob Health*. 2022;7(5).
- (6) World Health Organization. Number of weekly COVID-19 deaths reported to WHO (cumulative total) 2024 [updated 17 November 2024; cited 2024 03 December]. 7 days to 5 January 2020 – 7 days to 17 November 2024. Available from: <https://data.who.int/dashboards/covid19/deaths?n=o>.
- (7) The World Bank. World Bank Country and Lending Groups online2024 [cited 2024 02 December]. Available from: <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>.
- (8) Bajgain KT, Badal S, Bajgain BB, Santana MJ. Prevalence of comorbidities among individuals with COVID-19: A rapid review of current literature. *American Journal of Infection Control*. 2021;49(2):238-46.
- (9) Schultz MJ, Gebremariam TH, Park C, Pisani L, Sivakorn C, Taran S, et al. Pragmatic Recommendations for the Use of Diagnostic Testing and Prognostic Models in Hospitalized Patients with Severe COVID-19 in Low- and Middle-Income Countries. *Am J Trop Med Hyg*. 2021;104(3_Suppl):34-47.
- (10) Duroseau B, Kipshidze N, Limaye RJ. The impact of delayed access to COVID-19 vaccines in low- and lower-middle-income countries. *Front Public Health*. 2022;10:1087138.
- (11) Hotez PJ. SARS-CoV-2 variants offer a second chance to fix vaccine inequities. *Nature Reviews Microbiology*. 2023;21(3):127-8.
- (12) Ayouni I, Maatoug J, Dhoub W, Zammit N, Fredj SB, Ghammam R, et al. Effective public health measures to mitigate the spread of COVID-19: a systematic review. *BMC Public Health*. 2021;21(1):1015.
- (13) McCabe R, Schmit N, Christen P, D'Aeth JC, Løchen A, Rizmie D, et al. Adapting hospital capacity to meet changing demands during the COVID-19 pandemic. *BMC Medicine*. 2020;18(1):329.
- (14) 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.
- (15) Chandna A, Mahajan R, Gautam P, Mwandigha L, Gunasekaran K, Bhusan D, et al. Facilitating Safe Discharge Through Predicting Disease Progression in Moderate Coronavirus Disease 2019 (COVID-19): A Prospective Cohort Study to Develop and Validate a Clinical Prediction Model in Resource-Limited Settings. *Clinical Infectious Diseases*. 2022;75(1):e368-e79.
- (16) Damen JA. Validation of prognostic models predicting mortality or ICU admission in patients with COVID19 in low- and middle-income countries: a global individual participant data meta-analysis. . 2024.

- (17) Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ*. 2023;380:e071018.
- (18) Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
- (19) World Health Organization. The WHO Global Clinical Platform for COVID-19 2024 [cited 2024 02 December]. Available from: <https://www.who.int/teams/health-care-readiness-clinical-unit/covid-19/data-platform>.
- (20) World Health Organization. WHO Global Clinical Platform for COVID-19 - CORE CASE REPORT FORM (CRF) 2020 [cited 2024 02 December]. Available from: <https://iris.who.int/bitstream/handle/10665/350007/WHO-2019-nCoV-Clinical-CRF-2021.1-eng.pdf?sequence=1>.
- (21) Mishra S, Khare D. On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study. *Journal of Medical Statistics and Informatics*. 2014;2:9.
- (22) World Health Organization. TERMS OF USE for the Global Clinical Data Platform for Clinical Characterization and Management of Patients with Emerging Infectious Diseases (such as COVID-19, viral hemorrhagic fever (including Ebola virus disease), Mpox, Acute hepatitis of unknown etiology, Lassa Fever, Disease X), : World Health Organization; 2023 [updated 2 June 2023; cited 2025 29 January]. Available from: https://cdn.who.int/media/docs/default-source/documents/emergencies/terms-of-use-global-covid19-data-platform.pdf?sfvrsn=d0a440c1_3.
- (23) Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med*. 2020;26(7):1037-40.
- (24) Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*. 2019;38(7):1276-96.
- (25) Berzuini C, Hannan C, King A, Vail A, O'Leary C, Brough D, et al. Value of dynamic clinical and biomarker data for mortality risk prediction in COVID-19: a multicentre retrospective cohort study. *BMJ Open*. 2020;10(9):e041983.
- (26) Zhou Y, He Y, Yang H, Yu H, Wang T, Chen Z, et al. Development and validation a nomogram for predicting the risk of severe COVID-19: A multi-center study in Sichuan, China. *PLoS One*. 2020;15(5):e0233328.
- (27) R Core Team. R: A Language and Environment for Statistical Computing Vienna, Austria: R Foundation for Statistical Computing.; 2024 [cited 2024 29 January]. Available from: <https://www.R-project.org/>.
- (28) Collins GS, Dhiman P, Ma J, Schlusser MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. 2024;384:e074819.
- (29) Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM, Cochrane IPDM-aMg. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLOS Medicine*. 2015;12(10):e1001886.
- (30) Heymans MWE, Iris;. Chapter 9 Rubin's Rules. 2019 20 January 2019. In: *Applied Missing Data Analysis* [Internet]. Amsterdam: Heymans and Eekhout. Available from: <https://bookdown.org/mwheyman/bookmi/rubins-rules.html>.
- (31) Bennouar S, Bachir Cherif A, Kessira A, Bennouar DE, Abdi S. Development and validation of a laboratory risk score for the early prediction of COVID-19 severity and in-hospital mortality. *Intensive Crit Care Nurs*. 2021;64:103012.
- (32) Wang K, Zuo P, Liu Y, Zhang M, Zhao X, Xie S, et al. Clinical and Laboratory Predictors of In-hospital Mortality in Patients With Coronavirus Disease-2019: A Cohort Study in Wuhan, China. *Clin Infect Dis*. 2020;71(16):2079-88.

- (33) Zhang H, Shi T, Wu X, Zhang X, Wang K, Bean D, et al. Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. medRxiv. 2020:2020.04.28.20082222.
- (34) Marincowitz C, Hodkinson P, McAlpine D, Fuller G, Goodacre S, Bath PA, et al. LMIC-PRIEST: Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19 in a middle-income setting. PLoS One. 2023;18(6):e0287091.
- (35) de Jong VMT, Rousset RZ, Antonio-Villa NE, Buenen AG, Van Calster B, Bello-Chavolla OY, et al. Clinical prediction models for mortality in patients with covid-19: external validation and individual participant data meta-analysis. Bmj. 2022;378:e069881.
- (36) Murthy S, Leligdowicz A, Adhikari NK. Intensive care unit capacity in low-income countries: a systematic review. PLoS One. 2015;10(1):e0116949.
- (37) Ayebele E, Kassebaum N, Roche A, Biccard B. Africa's critical care capacity before COVID-19. Southern African Journal of Anaesthesia and Analgesia. 2020;26(3):162-4.
- (38) Rhodes A, Ferdinande P, Flaatten H, Guidet B, Metnitz PG, Moreno RP. The variability of critical care bed numbers in Europe. Intensive Care Medicine. 2012;38(10):1647-53.
- (39) Okereke M, Ukoro NA, Adebisi YA, Ogunkola IO, Favour Iyagbaye E, Adiola Owbor G, et al. Impact of COVID-19 on access to healthcare in low- and middle-income countries: Current evidence and future recommendations. The International Journal of Health Planning and Management. 2021;36(1):13-7.
- (40) Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagnostic and Prognostic Research. 2021;5(1):1.
- (41) van Klaveren D, Zanos TP, Nelson J, Levy TJ, Park JG, Retel Helmrich IRA, et al. Prognostic models for COVID-19 needed updating to warrant transportability over time and space. BMC Med. 2022;20(1):456.



5

Methodological Quality Assessment Tools for Diagnosis and Prognosis Research: Overview and Guidance

Tabea Kaul*, Bas E Kellerhuis*, Johanna AA Damen, Ewoud Schuit, Kevin Jenniskens,
Maarten van Smeden, Johannes B Reitsma, Lotty Hooft, Karel GM Moons, Bada Yang

**contributed equally and shared first authors*

J Clin Epidemiol. 2025 Jan;177:111609

ABSTRACT

Background and Objective: Multiple tools exist for assessing the methodological quality of diagnosis and prognosis research. It can be challenging to decide on when to use which tool. We aimed to provide an overview of existing methodological quality assessment (QA) tools for diagnosis and prognosis studies, highlight the overlap and differences among these tools, and to provide guidance for choosing the appropriate tool.

Study Design and Setting: We performed a methodological review of tools designed for assessing risk of bias, applicability, or other aspects related to methodological quality in studies investigating tests/factors/markers/models for classifying or predicting a current (diagnosis) and/or future (prognosis) health state. Tools focusing exclusively on causal research or on reporting quality were excluded. Guidance was subsequently developed to assist in choosing an appropriate QA tool.

Results: We identified 14 QA tools, eight of which were developed for assessment of diagnosis studies, four for prognosis studies, and two addressing both. We propose a set of five questions to help guide the process of choosing a QA tool based on the purpose or question of the user: whether the focus is on (1) diagnosis, prognosis, or another domain; (2) a prediction model versus a test/factor/marker; (3) evaluating simply the performance of a test/factor/marker versus assessing its added value over other variables; (4) comparing two or more tests/factors/markers/models; and (5) whether the user aims to assess only risk of bias or also other quality aspects.

Conclusion: Existing QA tools for appraising diagnosis and prognosis studies vary in purpose, scope, and contents. Our guidance may help researchers, systematic reviewers, health policy makers, and guideline developers in specifying their purpose and question to select the most appropriate QA tool for their assessment.

INTRODUCTION

Methodological quality assessment (QA) tools are designed to aid the identification of shortcomings in the design, conduct, analysis, and reporting of a study (1). Such tools are routinely used in systematic reviews to assess the credibility of included studies, while one may also use them to guide the design of new primary studies or to judge their inclusion in clinical guidelines or health policy guidance.

Methodological quality is not a singular concept and may encompass different constructs. One may be interested in different constructs for various reasons and aims, e.g. to assess internal validity (i.e. risk of bias), external validity (i.e. applicability and generalizability), or the reporting quality of a study(2). Different QA tools may cover these aspects to varying extents.

There is a rapidly growing landscape of QA tools for diagnosis (i.e. classifying or predicting the current presence or absence of a certain health outcome, disorder or disease) and prognosis (i.e. predicting the occurrence of a future health outcome) research (3-5). For example, searching the 'Quality Assessment and Risk of Bias Tool Repository' by Duke University (6) yields at least two tools for assessing the quality of diagnosis studies (CASP Diagnostic Study Checklist (7), and QUADAS (8)) and at least two for prognosis studies (QUAPAS (9) and QUIPS (10)). A brief literature search reveals many other potentially relevant tools (3, 11, 12).

As multiple QA tools are available for the same domain or type of studies, with potential overlap in purpose, scope, structure, and contents, choosing the right tool for the user's aim or question can be challenging. Moreover, research questions and studies about diagnosis and prognosis can be diverse, and it is not always clear which tool aligns best with a specific research question or study. This ambiguity may lead to the selection of a tool that does not match the user's intention. Conversely, more than one existing QA tool may be appropriate for the same diagnosis or prognosis question or study.

The aim of this paper is to provide an overview of existing QA tools for diagnosis and prognosis research, to highlight their differences and similarities, and to provide guidance in the selection of the most appropriate QA tool for specific contexts. While questions about diagnosis and prognosis are generally considered separate domains, this paper addresses them together. This is because of their shared similarities in study methodology and possible confusion between them, illustrated by the use of QA tools for diagnosis in systematic reviews with prognostic questions and vice versa (13-15).

METHODS

Study design

We performed a methodological review on existing QA tools for diagnosis and prognosis studies. We registered the protocol of this review on the Open Science Framework (Available at: <https://osf.io/UGVYA/>).

Eligibility criteria

Inclusion criteria

We included tools intended for studies investigating the *performance* of variables for classifying or predicting the presence or absence of a certain health outcome, e.g. a disease or disorder (referred to as diagnosis) or for predicting a future health outcome (referred to as prognosis). Here, we use the term *performance* to refer to a measure of how well a variable classifies or predicts the outcome; this can include diagnostic accuracy (in the context of diagnosis studies) and prognostic value or accuracy (in the context of prognosis studies). The index variable under study can be a single variable (e.g., a patient characteristic, (bio)marker, or imaging test) or a multivariable model. Diagnosis and prognosis studies are sometimes collectively referred to as ‘prediction research’ and exclude studies that aim to establish a causal relation between the index variable and outcome (e.g. etiological studies, randomized trials of test-treatment strategies), as well as studies assessing whether the index variable modifies a therapeutic or preventive intervention effect (16).

Eligible tools for this review were those that should help assess risk of bias, applicability, or other aspects of quality related to a diagnosis or prognosis study (2). We defined *risk of bias* as limitations in study design, data collection, and analysis that may lead to systematic over- or underestimation of study results and *applicability* as relevance of the primary study to the aim of the assessor or reviewer. In case multiple updates of the same QA tool were identified, we included the most recent one.

Exclusion criteria

We excluded tools that were not intended for prediction research as well as those for assessing studies on the prevalence (prevalence studies) and incidence (overall prognosis studies) of outcomes, as these do not address the association between an index variable and an outcome variable. We also excluded tools that exclusively contained items about quality of reporting. Finally, we only considered tools that were

reported in a format that is ready for use by an assessor without requiring adaptations in the tool's form or content. Therefore, we did not consider guidance papers that describe shortcomings in study design or papers that merely contain a list of potential quality items to be QA tools.

Searching for and selecting QA tools

We systematically searched for QA tools in electronic databases, on websites of health policy and guideline organizations, and by consulting methodologists in the field of prediction research. All records were assessed in duplicate for eligibility. More details on the search and selection procedure can be found in the Supplement Section A, B, and C.

Data extraction and analysis

From each QA tool, we extracted data about its scope, structure, and QA items (Supplement, Section D). To enable a comparison of QA items between tools, we grouped similar items together by assigning a label for each item type (Supplement, Section E), for example 'predictor blinding' for items associated with blinded execution of index tests or factors (17, 18). Some items addressed more than one quality issue and were therefore assigned multiple item type labels. We classified each item type as pertaining to *risk of bias*, *applicability*, *reporting*, *summary of risk of bias* (a summary rating across several risk of bias items), *summary of applicability* (a summary rating across several applicability items), or *other* (a miscellaneous category). We classified items as *risk of bias* and *applicability* using the definitions described under section *Eligibility criteria*; an item was classified as *reporting* if the item focused on a specific description being present (or missing) from the study report; an item not belonging to any of the other categories was classified as *other*. Two reviewers (TK & BK) independently extracted data and assigned item type labels, with a third reviewer (BY) resolving disagreements.

Development of guidance for selecting a QA tool

Based on the similarities and differences we identified between the tools during our analysis, we developed a set of guiding questions through discussions within our author group. TK, BK, and BY proposed the initial set of questions, and refined them through discussions with the rest of the author group.

RESULTS

Overview

Results of the search

We identified 1,433 records of which fourteen were included (Figure 1). Of these fourteen QA tools, four were identified by both database and manual searching, one only by database searching, eight only by manual searching and one uniquely identified by methodologist consultation. We identified eight tools for assessing diagnosis studies (Arrivé et al. (19), CASP Diagnostic Study Checklist (7), CEBM Diagnostic Accuracy Studies (17), DAQS(20), QUADAS-2 (11), QUADAS-C (12), QUADOMICS (21), RaMboMan [diagnostic test accuracy studies](22, 23)), four for assessing prognosis studies (CEBM Prognostic Studies (24), QUAPAS (9), QUIPS (10), RaMboMan [etiology/risk/prognosis studies] (23, 25)), and two for assessing both diagnosis studies and prognosis studies (CASP Clinical Prediction Rule Checklist (18), PROBAST (26, 27)). A detailed list of reasons for exclusions based on full text assessment is provided in Section F of the Supplement.

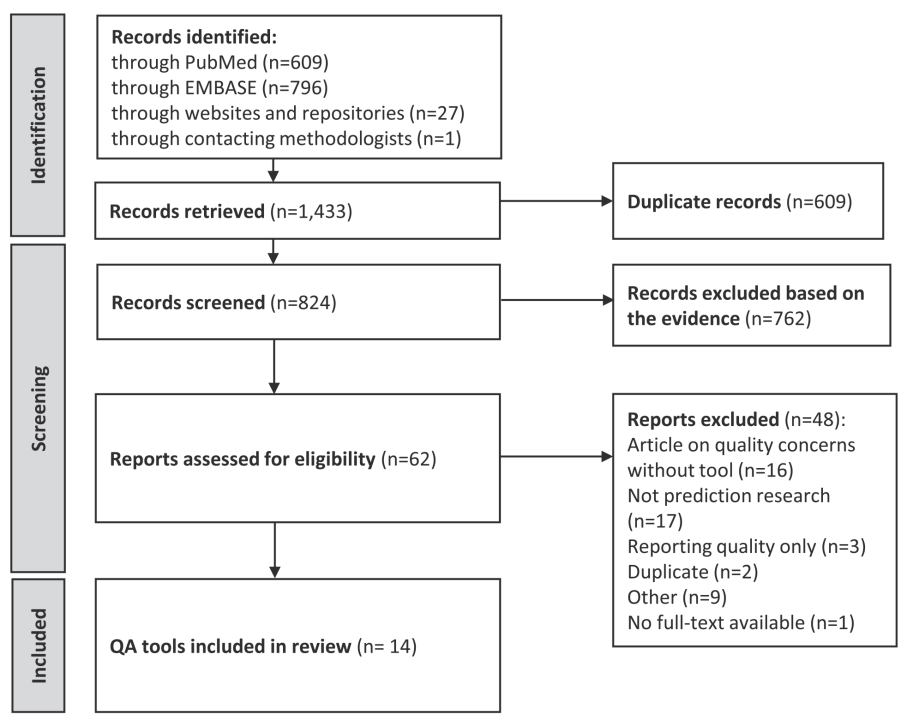


Figure 1 Flow diagram of included QA tools

Main characteristics of the included QA tools

Table 1 summarizes the main characteristics of the included QA tools. The tools contained a median of 18 items, ranging from 4 (CEBM Diagnostic Accuracy Studies (17)) to 38 (QUIPS (10)). All tools contained items about risk of bias and nine tools also addressed applicability. Nine tools contained items about reporting quality, with the proportion of reporting items varying from 0% (e.g., QUADAS-2 (11)) to 57% (DAQS (20)). Twelve tools were intended for assessing studies that evaluate the performance of univariable predictors (e.g., tests/factors/[bio]markers); the remaining two were intended for multivariable prediction models (CASP Clinical prediction rule checklist [Section A] (18); PROBAST (26, 27)). One tool addressed bias in the comparison of diagnostic performance (QUADAS-C (12)).

Table 1 Main characteristics of the included QA tools

QA tool	Diagnosis or prognosis?	Quality aspects assessed	Type of assessment	Number of items	Reporting quality items % (n/N)*
Arrivé et al.† (19)	Diagnosis	RoB + Rep.	Checklist with numerical score	15	47% (7/15)
CASP-Dx (7)	Diagnosis	RoB + App. + Rep.	Checklist	7	43% (3/7)
CEBM-Dx (17)	Diagnosis	RoB + Rep.	Checklist	4	25% (1/4)
DAQS (20)	Diagnosis	RoB + Rep.	Checklist with numerical score	21	57% (12/21)
QUADAS-2 (11)	Diagnosis	RoB + App.	SQs with summary judgment	18	0% (0/18)
QUADAS-C (12)	Diagnosis	RoB	SQs with summary judgment	18	0% (0/18)
QUADOMICSt (21)	Diagnosis	RoB + App. + Rep.	Checklist	16	38% (6/16)
RAMboMAN (Dx) (22, 23)	Diagnosis	RoB + App. + Rep.	SQs with summary judgment	21	14% (3/21)
CEBM-Px (24)	Prognosis	RoB + App.	Checklist	5	0% (0/5)
QUAPAS (9)	Prognosis	RoB + App.	SQs with summary judgment	27	0% (0/27)
QUIPS (10)	Prognosis	RoB + App. + Rep.	SQs with summary judgment	38	37% (14/38)
RAMboMAN (Px)(23, 25)	Prognosis	RoB + App. + Rep.	SQs with summary judgment	24	17% (4/24)

Table 1 Main characteristics of the included QA tools (*continued*).

QA tool	Diagnosis or prognosis?	Quality aspects assessed	Type of assessment	Number of items	Reporting quality items % (n/N)*
CASP-CPR (18)	Both	RoB + Rep.	Checklist	6	33% (2/6)
PROBAST (26, 27)	Both	RoB + App.	SQs with summary judgment	29	0% (0/29)

*Number of items per QA tool;†The tools are tailored to a specific clinical domain, namely studies of radiological tests (Arrivé et al.) and of omics-based tests (QUADOMICS); RoB: Risk of bias; App.: Applicability; Rep.: Reporting; Arrivé et al.: A scale of methodological quality for clinical studies of radiologic examination;CASP-Dx: CASP Diagnostic Study Checklist (Section A) ; CEBM-Dx: CEBM Diagnostic accuracy studies; RAMboMAN (Dx): RAMboMAN (diagnostic test accuracy studies); CEBM-Px: CEBM Prognostic Studies; RAMboMAN (Px): RAMboMAN (etiology/risk/prognosis studies); CASP-CPR: CASP Clinical Prediction rRule Checklist (Section A); SQ: signalling question.

Items included in QA tools

The Supplement (Section G) provides an overview of the types of items included in the 14 tools. We identified 44 unique item types within the included QA tools. We organized the item types in the following domains: Participants (items on participant selection), Predictors (items on the measurement of predictor variables), Outcome (items on the measurement of outcome variables), Covariates (items on other predictors variables were adjusted for), Analysis (items on the statistical analysis), Comparative (items about comparing tests/factors/[bio]markers/models), Overall (items on overall summary ratings on quality aspects), and Other (miscellaneous category: items that are not about risk of bias, applicability, or reporting). A detailed overview of item label definitions is provided in Section E of the Supplement. Some item types only appeared in one or a few tools, e.g., the item type ‘Covariate measurement’ (only in QUIPS (10)) and those in the Comparative domain (only in QUADAS-C (12)). Other item types were found in most tools, such as items about the study design and sources of data used in the study (item type ‘Data source’). The item matrix (Supplementary file 1) contains a full list of items from each QA tool in their original wording.

Guidance

By examining the differences between the 14 included QA tools, we developed five questions which can guide users in their choice of a QA tool that best corresponds to their question and purpose of assessment (Table 2). In the following sections, we list

the guiding questions followed by a description. Examples can be found in Table 2. The five questions may be used at any moment when difficulty arises when to use which QA tool. This may be during the protocol writing phase or when assessing a study. The questions invite careful thinking about the research question and purpose of the assessment from the user's (rather than the study authors') point of view. Table 3 shows which answers to these guiding questions correspond to which QA tools.

Question 1: Is your research question about diagnosis, prognosis, or another domain?

Users should first consider whether their aim or question is about diagnosis, prognosis, or some other aim or focus. Diagnosis questions are about determining the presence or absence of a health outcome (e.g. disease or disorder) of an individual at the current moment, whereas prognosis questions are about determining the occurrence of a health outcome at a later moment in time. Therefore, diagnostic study designs are typically cross-sectional, whereas prognostic designs are longitudinal, requiring a period of follow-up. QA tools for prognosis studies may include items relevant to the longitudinal aspect of the study, for example about competing risks and censoring. Diagnosis and prognosis questions are not about causal associations between the index variable(s) and the health outcomes (16, 28, 29). To assess the methodological quality of causal research, users should use QA tools specifically designed for those, such as Cochrane's Risk of Bias 2 tool for randomized studies (30) or ROBINS-I for nonrandomized studies (31).

Question 2: Is your research question about a prediction model?

While most of the identified QA tools are developed to assess the methodological quality of studies that investigate a single test, factor, or (bio)marker, users may be interested in assessing the quality of studies that developed or evaluated the performance of a prediction model, which uses multiple variables (predictors) jointly to estimate the probability that an individual has a certain health outcome (diagnostic prediction model) or will have in the future (prognostic prediction model). By combining multiple variables in a single model, one can answer the question 'What is the predicted individualized (diagnostic or prognostic) probability for a person based on a set of variables X1, X2, X3, et cetera?'. Well-known diagnostic and prognostic models include Wells' score for deep venous thrombosis (32) and CHA₂DS₂-VASc for stroke risk in people with atrial fibrillation(33), respectively. Studies developing or evaluating (i.e. testing or validating) diagnostic or prognostic models typically require additional considerations in the design

and analysis, such as evaluating the model in a different set of individuals than the one in which the model was originally developed.

Question 3: Is your research question about (a) simply the performance of a test, factor, or (bio)marker or (b) its added value over other variables?

Users may be interested in the predictive performance of a single diagnostic or prognostic test, factor, or (bio)marker: ‘What is the performance of test X in estimating or predicting outcome Y?’. But others may be interested in the performance of that test while accounting for other variables, to assess whether that test retains its diagnostic or prognostic performance when the value of these other variables (often referred to as covariates) is known: ‘What is the performance of test X in predicting outcome Y after taking covariates Z into account?’. The latter question may be relevant to those who want to know whether a new test has added value over existing clinical information, such as age and sex. If this added value is of key interest, the ideal study design is one in which test X and outcome Y as well as all relevant covariates Z have been measured (without error), and in which an appropriate method was used to account for these covariates in data analysis. An appropriate QA tool should contain items relating to these aspects of study design and analysis.

Note that a question regarding the added value of a test (or factor or [bio]marker), though it involves multiple variables, is different from a question regarding a prediction model. In the former, one is mainly interested in the performance of a single predictor variable only after accounting for covariates. In the latter, one is interested in the performance of all the variables combined (i.e., model performance).

Question 4: Is your research question about comparing the performance of two or more tests, factors, (bio)markers, or models?

Users may be interested in knowing the absolute or relative difference in performance between competing diagnostic or prognostic tests (factors, (bio)markers, or models), for example with the intention of possibly replacing an existing test with a new test: ‘What is the difference in performance between X1 and X2 in estimating or predicting Y?’. Comparisons between two tests involve more than just two evaluations of different tests, so it is insufficient to use QA tools designed for the evaluation of a single test, even if they are applied to each test individually. A study comparing the performance of two tests should ensure that the groups being compared are exchangeable, e.g., by having all participants undergo both tests or by randomizing participants to one test

or the other (34). In such a study, the reference standard for determining the health status or outcome is not one of the competing tests under investigation.

Question 5: Is your aim to assess only risk of bias?

While most users will be interested in assessing risk of bias, others may also be interested in assessing applicability of results to their intended use setting. Users may also consider whether other constructs are of interest, e.g. the quality of reporting or a broader definition of quality including, e.g. random error. When the focus is strictly on risk of bias, a tool with a high proportion of risk of bias items and a low proportion of items about reporting quality and other constructs is likely preferable. Table 1 shows, for each tool, the proportion of items about quality of reporting; Section G of the Supplement identifies tools with ‘other’ quality items that are neither about risk of bias, applicability, or reporting quality.

Table 2 Five questions about the user’s research question and the purpose of assessment.

Question	Possible answers:	Example
1. Is your research question about diagnosis, prognosis, or another domain?	• Diagnosis • Prognosis • Another domain (e.g., etiology)	You are interested in the predictive value of an imaging test in determining whether patients have Merkel cell carcinoma (35). In your ideal study, all study patients simultaneously undergo the imaging test and the reference standard (pathology) to determine the presence of Merkel cell carcinoma. Your question is about diagnosis, because the interest is in determining the current health state of patients using the imaging test. (Answer: <i>Diagnosis</i>)
2. Is your research question about a prediction model?	• Yes • No (not a prediction model, e.g., a test, factor, or (bio)marker)	You are interested in determining the risk of in-hospital mortality for COVID-19 patients using the 4C Mortality Score which uses eight predictors (e.g. age, sex, and C-reactive protein) to calculate an individual patient’s risk(36). The 4C Mortality Score is a prediction model aimed at calculating the probability of death within a certain timeframe for an individual patient using multiple predictors. (Answer: <i>Yes</i>)

Table 2 Five questions about the user's research question and the purpose of assessment. (continued).

Question	Possible answers:	Example
3. Is your research question about (a) simply the performance of a test, factor, or (bio)marker or (b) its added value over other variables?*	- Simply the performance of the test, factor, or (bio)marker - Added value over other variables	You are interested in studying a new (bio) marker for diagnosing Alzheimer's disease. An existing test is magnetic resonance imaging (MRI)(37). You wonder if the new (bio)marker adds information about the likelihood of Alzheimer's even when the results of MRI are known. Your ideal study assesses the performance of the (bio)marker conditional on MRI, e.g. by conducting separate analyses for MRI positive and negative patients. Will the (bio)marker still have value after accounting for another variable, i.e. MRI? <i>(Answer: Added value over other variables)</i>
4. Is your research question about comparing the performance of two or more tests, factors, (bio)markers, or models?	- Yes - No (focusing on the performance of a single test, factor, [bio] marker, or model)	You are interested in tests for predicting the occurrence of esophageal varices in patients with liver disease within the next ten years. You want to know if liver biopsy can be replaced by any of seven other non-invasive tests(38). Your ideal study compares the prognostic performance of liver biopsy and the non-invasive tests in a single study, in which each participant receives all eight tests and are followed up for the occurrence of varices. <i>(Answer: Yes)</i>
5. Is your aim to assess only risk of bias?	- Yes - No (e.g., both risk of bias and applicability)	You are interested in the prognostic accuracy of the Quick Sequential (Sepsis-related) Organ Failure Assessment (SOFA) prediction model in children with hospital-acquired urinary tract infection. You identified a study that assesses the prognostic accuracy of Quick SOFA in older adults with community-acquired urinary tract infection(39). The study is well designed and is at low risk of bias. But the study included only older adults, which prompts concerns regarding the applicability of the results to your population of interest. In this example, both risk of bias and applicability are assessed. <i>(Answer: No [both risk of bias and applicability])</i>

*Can be skipped if question 2 is answered 'Yes'

Table 3 QA tools with answers to the five guiding questions

QA tool name	Q1: Diagnosis, prognosis, or another domain?	Q2: Prediction model?	Q3: Performance or added value?*	Q4: Comparing tests/ markers/ models?	Q5: Aim is to assess only risk of bias?
Arrivé et al. (19)	Diagnosis	No	Performance	No	RoB + Rep.
CASP-Dx (7)	Diagnosis	No	Performance	No	RoB + App. + Rep.
CEBM-Dx (17)	Diagnosis	No	Performance	No	RoB + Rep.
DAQS (20)	Diagnosis	No	Performance	No	RoB + Rep.
QUADAS-2 (11)	Diagnosis	No	Performance	No	RoB + App.
QUADAS-C (12)	Diagnosis	No	Performance	Yes	RoB
QUADOMICS (21)	Diagnosis	No	Performance	No	RoB + App. + Rep.
RAMboMAN (Dx) (22, 23)	Diagnosis	No	Performance	No	RoB + App. + Rep.
CEBM-Px (24)	Prognosis	No	Performance	No	RoB + App.
QUAPAS (9)	Prognosis	No	Performance	No	RoB + App.
QUIPS (10)	Prognosis	No	Added value	No	RoB + App. + Rep.
RAMboMAN (Px) (23, 25)	Prognosis and etiology	No	Performance	No	RoB + App. + Rep.
CASP-CPR (18)	Both	Yes	N/A†	No	RoB + Rep.
PROBAST (26, 27)	Both	Yes	N/A†	No	RoB + App.

RoB: Risk of bias; App.: Applicability; Rep.: Reporting; N/A: Not applicable

*About simply the performance of the test, factor, or (bio)marker versus its added value over other variables.

†Usually not applicable to prediction models, as few studies will assess the added value of prediction models over other variables.

Arrivé et al.: A scale of methodological quality for clinical studies of radiologic examination

CASP-Dx: CASP Diagnostic Study Checklist (Section A)

CEBM-Dx: CEBM Diagnostic accuracy studies

RAMboMAN (Dx): RAMboMAN (diagnostic test accuracy studies)

CEBM-Px: CEBM Prognostic studies

RAMboMAN (Px): RAMboMAN (etiology/risk/prognosis studies)

CASP-CPR: CASP Clinical Prediction Rule Checklist (Section A)

Other considerations

Aside from these five guiding questions, there are other aspects users may consider when choosing the most appropriate QA tool for their specific diagnosis or prognosis question (Box 1). They could consider whether the tool contains all important quality issues relevant to their question or aim, and whether the number of redundant items is low. For example, if the question is about an outcome that is difficult to measure or detect, one may find it important to have a question about outcome misclassification. Users can use section G of the Supplement to gain an idea of the items included in each tool. Depending on available resources, systematic reviewers sometimes also consider the number of items and therefore time required to complete the tool (Table 1). Some may find it important to know whether a QA tool is simply a checklist or one that results in a summary judgment. Users may also be interested in knowing how the tools were developed, particularly if the items are based on the best available evidence and if a systematic and transparent approach was adopted to reach consensus on the development of the tool. The Supplement (Section H) provides a summary of methods used to develop the 14 tools. Some tools are designed or adapted for use in specific areas of research. For example, QUADOMICS (21) is a tailored version of the original QUADAS (8) for studies on omics-based tests. Moreover, some tools may be designed for multivariable prediction model development (e.g., PROBAST (26, 27); specifically, the items relating to model development), while others are designed for model evaluation (e.g., PROBAST (26, 27), CASP Clinical Prediction Rule Checklist (18)). For each of the 14 tools, we invite users to consider whether the tool satisfies all relevant aspects for their intended use.

Box 1. Other considerations when choosing a QA tool

- Whether the tool addresses the key methodological concerns relevant to the user's question (see Supplement [Section G] for an overview of items per tool)
 - Type of assessment: simple checklist, checklist leading to a numerical quality score, or signalling questions leading to a summary judgment (Table 1)
 - The number of items included in the tool (Table 1)
 - How the tool was developed (see Supplement [Section H])
 - Whether the tool is tailored for studies within a specific clinical domain
-

DISCUSSION

We conducted a methodological literature review to provide an overview of existing QA tools for diagnosis and prognosis studies and developed specific guidance to support

researchers in the selection of the QA tool appropriate for their specific question and purpose of assessment. We proposed five guiding questions to invite users to consider the precise nature of their research question and whether it matches with that of the QA tool.

We identified a variety of QA tools for diagnosis and prognosis studies. Our review distinguishes itself from others by specifically focusing on diagnosis and prognosis studies and by providing guidance in selecting a tool from the users' point of view. In contrast to a previous review by Whiting et al. (2005) on QA tools for diagnostic accuracy studies (n=91), our search yielded fewer eligible QA tools, despite including tools intended for both diagnosis and prognosis studies (n=14 of which n=8 about diagnosis) (3). This difference is mainly due to the fact that we excluded guidance papers and systematic reviews that merely described or included specific quality assessment items; we did not regard these as 'QA tools'. As we realize no universal definition of QA tools exists, we only included tools reported in a format that is ready for use without requiring adaptations in the tool's form or content. On the other hand, we identified more QA tools for diagnosis and prognosis studies than another review by Zeng et al. (2015) and the updated review by Ma et al. (2020) which provided an overview of QA tools for various study types and clinical guidelines (4, 5). We likely identified more eligible QA tools for diagnosis and prognosis studies as the inclusion criteria of these two reviews were restricted to QA tools 'also used or commonly used now', without providing a precise definition for this criterion (4, 5).

The LATITUDES network (www.latitudes-network.org), which maintains a library of validity assessment tools designed for use in evidence synthesis, has recently published a decision tree that is intended to help reviewers choose an adequate QA tool for different types of study questions, including diagnosis and prognosis studies(40). This decision tree also contains three guiding questions regarding QA tools for diagnosis and prognosis studies: *"Did the study compare the accuracy of a diagnostic test against a reference standard?"* ('Yes' leading to QUADAS-2), *"Did the study evaluate the prognostic value of one or more factors?"* ('Yes' leading to QUIPS), and *"Did the research develop, validate or update a general prediction model for diagnosis or prognosis?"* ('Yes' leading to PROBAST)(40). Though these questions are similar to ours, our guidance focusses on clarifying the question of the user (i.e. assessor, reviewer, or policy maker) rather than the intention of the primary study authors: the assessment of quality is based on the user's, not the study authors', perspective. We also deliberately refrained from recommending a particular tool, as the preference for a tool may depend on several

factors including, but not limited to, relevance of included quality items to the user's question and the time required to complete a tool.

Even after answering all five guiding questions, users may still face the choice between two or more QA tools. In such cases, they may wonder which of the tools is the best one for their question and purpose. To date, no consensus exists on how to determine the 'quality' of QA tools. One may prefer tools that appear to, at face value, contain all the necessary quality items (face validity), or ones that are developed following a systematic approach (2). One may examine studies of inter-rater reliability or reliability between reviewer pairs (41, 42). Users may prefer tools that are most supported by evidence from so-called meta-epidemiological studies (43-45), though such studies may not be without flaws(46-48). More research is needed on how to synthesize and appraise these types of data and which role they should play in a comparative assessment between multiple QA tools. The other considerations (Box 1) from our guidance can help users to choose a particular tool when evidence about tool validity and reliability are not directly available.

In the future, an (interactive) catalogue of items could be developed to help users with adequate expertise to tailor or compile a QA tool for specific assessment needs. Our matrix of extracted QA items from all 14 tools (Supplement) may help guide this development.

Limitations of this study

Though we attempted to provide general definitions of *risk of bias* and *applicability*, unambiguous consensus-based definitions are yet lacking. Furthermore, some QA tools may have been missed by our search, either because they were not yet included in the sources we searched or because their titles and abstracts lacked the specific terms we used in our search (e.g. 'methodological quality', 'critical appraisal', 'instrument', 'bias'). We supplemented our electronic search by manually searching websites of organizations specializing in systematic reviews and health technology assessment (Supplement, Section C) and consulted with methodologists to minimize the risk of overlooking relevant QA tools. As mentioned previously, we were not able to assess the quality of the included tools, as there is currently no agreement on what constitutes 'tool quality'. We extracted details on their development processes (Supplement, Section H), to allow users to assess the rigor with which these tools were developed.

CONCLUDING REMARKS

Existing methodological QA tools for diagnosis and prognosis studies vary in purpose, scope, structure, and composition, highlighting the need for careful consideration when selecting an appropriate tool. In our guidance on selecting appropriate tools, we emphasize the importance of formulating a well-defined diagnosis or prognosis research question and we invite users to consider which quality aspects are relevant for their specific situation or question. Using an appropriate tool for specific diagnosis or prognosis questions is crucial in ensuring that most aspects potentially compromising the quality of one or more studies of interest are identified. While the availability of a multitude of tools presents opportunities, it is also accompanied by challenges: users consistently encounter the need to invest significant time and effort to familiarize themselves with each tool, often compounded by a lack of clarity on how they differ from one another. We observed overlap in QA items and structural similarities across tools, suggesting potential duplication of efforts in tool development. Considering these issues, a collaborative effort across the field may be desirable to develop a single, interactive tool for diagnosis and prognosis studies that can be tailored to a specific research question. Such a tool, if sufficiently flexible to deal with diverse user needs, may improve the efficiency of quality assessment.

Acknowledgements

We are grateful to Linde Huis in 't Veld, MSc (Cochrane Netherlands) and Hayang Yang, MD, PhD (National University of Singapore) for their valuable comments on our manuscript.

Authors' contributions

Tabea Kaul: Writing - Review & Editing, Writing - Original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Bas E Kellerhuis: Writing - Review & Editing, Writing - Original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Johanna AA Damen: Writing - Review & Editing, Methodology, Conceptualization.

Ewoud Schuit: Writing - Review & Editing, Methodology, Conceptualization.

Kevin Jenniskens: Writing - Review & Editing, Methodology, Conceptualization.

Maarten van Smeden: Writing - Review & Editing, Methodology, Conceptualization.

Johannes B Reitsma: Writing - Review & Editing, Methodology, Conceptualization.

Lotty Hoof: Writing - Review & Editing, Methodology, Conceptualization.

Karel G.M. Moons: Writing - Review & Editing, Methodology, Conceptualization.

Bada Yang: Writing - Review & Editing, Writing - Original draft, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

None.

Competing interests

Specific authors contributed to the development of quality assessment tools included and described in this study. ES, JBR, LH, and KGMM were part of the development team of the PROBAST tool; JBR and BY were part of the development team of the QUADAS-C tool; and JBR was part of the development team of the QUADAS-2 tool. With the exception of BY, these authors did not participate in the phases of tool selection, data extraction, and data analysis. There are no competing interests for any other author.

Ethical approval

Not required for this work.

Registration and protocol

The protocol of this study was registered on the Open Science Framework and is available at: https://doi.org/10.17605/OSF.IO/UGVYA_.

Data availability

The item matrix including our data extractions can be found online in the supplementary data.

Supplementary data

Supplementary data to this article can be found online.

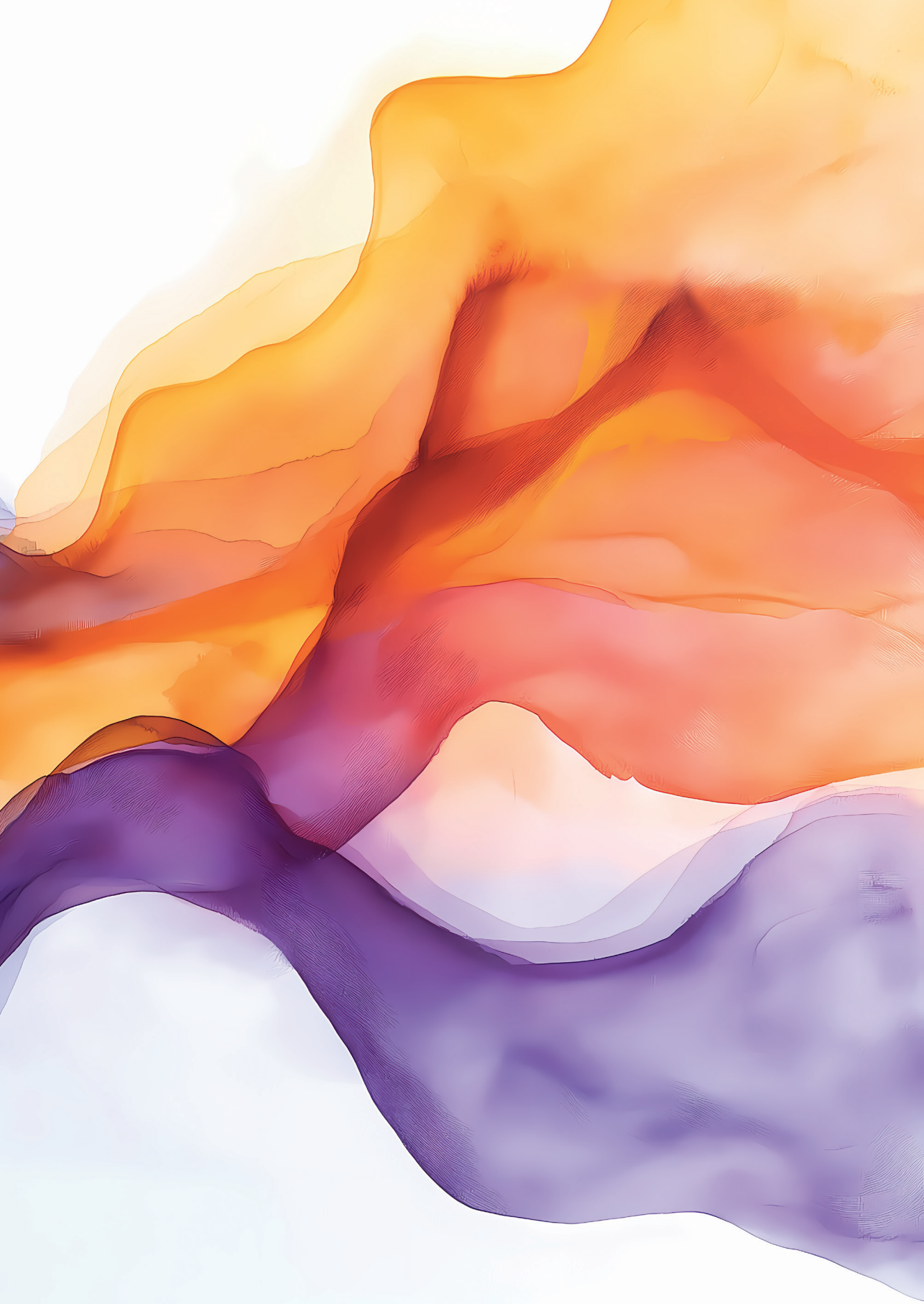


REFERENCES

- (1) Boutron I, Page MJ, Higgins JP, Altman DG, Lundh A, Hróbjartsson A, et al. Considering bias and conflicts of interest among the included studies. *Cochrane handbook for systematic reviews of interventions*. Hoboken, NJ: Wiley-Blackwell; 2019:177-204.
- (2) Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. *Systematic Reviews*. 2017;6(1):204.
- (3) Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol*. 2005;58(1):1-12.
- (4) Ma L-L, Wang Y-Y, Yang Z-H, Huang D, Weng H, Zeng X-T. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military Medical Research*. 2020;7(1):7.
- (5) Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. *J Evid Based Med*. 2015;8(1):2-10.
- (6) Duke University's Medical Center Library & Archives. Quality Assessment and Risk of Bias Tool Repository 2021 [cited 2024 April 24]. Available from: <https://doi.org/10.17605/OSF.IO/WS824>.
- (7) Critical Appraisal Skills Programme. CASP Diagnostic Study Checklist [online]. 2022 [cited 2022 December 19]. Available from: https://casp-uk.net/images/checklist/documents/CASP-Diagnostic-Study-Checklist/CASP-Diagnostic-Checklist-2018_fillable_form.pdf.
- (8) Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:25.
- (9) Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: An Adaptation of the QUADAS-2 Tool to Assess Prognostic Accuracy Studies. *Ann Intern Med*. 2022;175(7):1010-8.
- (10) Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med*. 2013;158(4):280-6.
- (11) Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-36.
- (12) Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, et al. QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. *Ann Intern Med*. 2021;174(11):1592-9.
- (13) Shen MH, Qiu GQ, Wu XM, Dong MJ. Utility of the DECAF score for predicting survival of patients with COPD: a meta-analysis of diagnostic accuracy studies. *Eur Rev Med Pharmacol Sci*. 2021;25(11):4037-50.
- (14) Huang Q, He C, Xiong H, Shuai T, Zhang C, Zhang M, et al. DECAF score as a mortality predictor for acute exacerbation of chronic obstructive pulmonary disease: a systematic review and meta-analysis. *BMJ Open*. 2020;10(10):e037923.
- (15) Adusumilli P, Ravikumar N, Hall G, Swift S, Orsi N, Scarsbrook A. Radiomics in the evaluation of ovarian masses - a systematic review. *Insights Imaging*. 2023;14(1):165.
- (16) Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE*. 2019;32(1):42-9.
- (17) Centre for Evidence-Based Medicine. CEBM Diagnostic Accuracy Studies University of Oxford; 2018 [cited 2024 April 24]. Available from: <https://www.cebm.net/wp-content/uploads/2018/11/Diagnostic-Accuracy-Studies.pdf>.

- (18) Critical Appraisal Skills Programme. CASP Clinical Prediction Rule Checklist 2018 [cited 2024 April 24]. Available from: <https://casp-uk.net/checklists/casp-clinical-prediction-rule-checklist.pdf>.
- (19) Arrivé L, Renard R, Carrat F, Belkacem A, Dahan H, Le Hir P, et al. A scale of methodological quality for clinical studies of radiologic examinations. *Radiology*. 2000;217(1):69-74.
- (20) Cook C, Cleland J, Hegedus E, Wright A, Hancock M. The creation of the diagnostic accuracy quality scale (DAQS). *J Man Manip Ther*. 2014;22(2):90-6.
- (21) Lumberras B, Porta M, Márquez S, Pollán M, Parker LA, Hernández-Aguado I. QUADOMICS: an adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of 'omics'-based technologies. *Clin Biochem*. 2008;41(16-17):1316-25.
- (22) University of Auckland. GATE CAT – Diagnostic Test Accuracy Studies (RAMboMAN) 2020 [cited 2024 April 24]. Available from: <https://www.fmhs.auckland.ac.nz/assets/fmhs/soph/epi/epiq/docs/GATE%20CAT%20Diagnostic%20Studies%20May%202014%202014%20V5.docx>.
- (23) Rod J, Shanthi A, Joanna B, Jennie C, Anne L, Gill R, et al. The GATE frame: critical appraisal with pictures. *Evidence Based Nursing*. 2006;9(3):68.
- (24) Centre for Evidence-Based Medicine. CEBM Critical appraisal of prognostic studies: University of Oxford Available at : <https://www.cebm.net/wp-content/uploads/2018/11/Prognosis.pdf>; 2018[Accessed 24 April 2024].
- (25) University of Auckland. GATE CAT – Risk Factor Cohort Studies (RAMboMAN) 2014 [cited 2024 April 24]. Available from: <https://www.fmhs.auckland.ac.nz/assets/fmhs/soph/epi/epiq/docs/GATE%20CAT%20Risk%20Factor%20Cohort%20Studies%20May%202014%20V3.docx>.
- (26) 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(1):51-8.
- (27) 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(1):W1-w33.
- (28) Shmueli G. To Explain or to Predict? *Statistical Science*. 2011;25.
- (29) Ramspek CL, Steyerberg EW, Riley RD, Rosendaal FR, Dekkers OM, Dekker FW, et al. Prediction or causality? A scoping review of their conflation within current observational research. *Eur J Epidemiol*. 2021;36(9):889-98.
- (30) 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.
- (31) Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
- (32) Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, et al. Accuracy of clinical assessment of deep-vein thrombosis. *Lancet*. 1995;345(8961):1326-30.
- (33) Lip GY, Nieuwlaar R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137(2):263-72.
- (34) Yang B, Olsen M, Vali Y, Langendam MW, Takwoingi Y, Hyde CJ, et al. Study designs for comparative diagnostic test accuracy: A methodological review and classification scheme. *J Clin Epidemiol*. 2021;138:128-38.
- (35) Colgan MB, Tarantola TI, Weaver AL, Wiseman GA, Roenigk RK, Brewer JD, et al. The predictive value of imaging studies in evaluating regional lymph node involvement in Merkel cell carcinoma. *J Am Acad Dermatol*. 2012;67(6):1250-6.
- (36) Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ*. 2020;370:m3339.

- (37) Frederiksen KS, Hasselbalch SG, Hejl AM, Law I, Højgaard L, Waldemar G. Added Diagnostic Value of (11)C-PIB-PET in Memory Clinic Patients with Uncertain Diagnosis. *Dement Geriatr Cogn Dis Extra*. 2012;2(1):610-21.
- (38) Rasmussen DN, Thiele M, Johansen S, Kjærgaard M, Lindvig KP, Israelsen M, et al. Prognostic performance of 7 biomarkers compared to liver biopsy in early alcohol-related liver disease. *J Hepatol*. 2021;75(5):1017-25.
- (39) Madrazo M, López-Cruz I, Zaragoza R, Piles L, Eiros JM, Alberola J, et al. Prognostic accuracy of Quick SOFA in older adults hospitalised with community acquired urinary tract infection. *Int J Clin Pract*. 2021;75(10):e14620.
- (40) LATITUDES network. LATITUDES risk of bias (ROB) tool Decision Tree Which tools are appropriate for my review? 2023 [cited 2023 15 December]. Available from: <https://www.latitudes-network.org/wp-content/uploads/2023/05/flowchart.jpg>.
- (41) Kaizik MA, Garcia AN, Hancock MJ, Herbert RD. Measurement properties of quality assessment tools for studies of diagnostic accuracy. *Brazilian Journal of Physical Therapy*. 2020;24(2):177-84.
- (42) Kaiser I, Pfahlberg AB, Mathes S, Uter W, Diehl K, Steeb T, 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).
- (43) 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.
- (44) Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *Cmaj*. 2006;174(4):469-76.
- (45) Damen J, Debray TPA, Pajouheshnia R, Reitsma JB, Scholten R, Moons KGM, et al. Empirical evidence of the impact of study characteristics on the performance of prediction models: a meta-epidemiological study. *BMJ Open*. 2019;9(4):e026160.
- (46) Moustgaard H, Jones HE, Savović J, Clayton GL, Sterne JA, Higgins JP, et al. Ten questions to consider when interpreting results of a meta-epidemiological study—the MetaBLIND study as a case. *Research Synthesis Methods*. 2020;11(2):260-74.
- (47) Herbert RD. Controversy and Debate on Meta-epidemiology. Paper 2: Meta-epidemiological studies of bias may themselves be biased. *J Clin Epidemiol*. 2020;123:127-30.
- (48) Hróbjartsson A, Boutron I, Turner L, Altman DG, Moher D. Assessing risk of bias in randomised clinical trials included in Cochrane Reviews: the why is easy, the how is a challenge. *Cochrane Database Syst Rev*. 2013;2013(4):Ed000058.



6

Assessing the Quality of Prediction Models in Healthcare Using the Prediction model Risk Of Bias ASsessment Tool (PROBAST): An Evaluation of its Use and Practical Application

Tabea Kaul, Johanna AA Damen, Laure Wynants, Ben Van Calster,
Maarten van Smeden, Lotty Hooft, Karel GM Moons

J Clin Epidemiol. 2025 Feb 181:111732;25

ABSTRACT

Background and Objectives: Since 2019, the Prediction model Risk Of Bias ASsessment Tool (PROBAST: www.probast.org) has supported methodological quality assessments of prediction model studies. Most prediction model studies are rated with a “High” risk of bias (ROB) and researchers report low inter-rater reliability (IRR) using PROBAST. We aimed to 1) assess the IRR of PROBAST ratings between assessors of the same study and understand reasons for discrepancies, 2) determine which items contribute most to domain-level ROB ratings, and 3) explore the impact of consensus meetings.

Study Design and Setting: We used PROBAST assessments from a systematic review of diagnostic and prognostic COVID-19 prediction models as a case study. Assessors included international experts in prediction model studies or their reviews. We assessed IRR using prevalence-adjusted bias-adjusted kappa (PABAK) before consensus meetings, examined bias ratings per domain-level ROB judgements, and evaluated the impact of consensus meetings by identifying rating changes after discussion.

Results: We analysed 2167 PROBAST assessments from 27 assessor pairs covering 760 prediction models: 384 developments, 242 validations, and 134 mixed assessments (including both). The IRR using PABAK was higher for Overall ROB judgements (development: 0.82 [0.76;0.89]; validation: 0.78 [0.68;0.88]) compared to domain- and item-level judgements. Some PROBAST items frequently contributed to domain-level ROB judgements, e.g. 3.5 Outcome blinding and 4.1 Sample size. Consensus discussions mainly led to item-level and never to Overall ROB rating changes.

Conclusion: Within this case study, PROBAST assessments received high IRR at the Overall ROB level, with some variation at item- and domain-level. To reduce variability, PROBAST assessors should standardize item- and domain-level judgements and hold well-structured consensus meetings between assessors of the same study.

INTRODUCTION

The Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) was launched in 2019 and was designed to support researchers in judging the risk of bias (ROB) and applicability of diagnostic and prognostic prediction models (1, 2). Bias refers to a systematic error in the estimators of the model's predictive performance for the target population of interest and may act in either direction such that there may be over – or under estimation of the true model performance. PROBAST consists of four domains with a total of 20 signalling questions concerning ROB and applicability (1, 2). These signalling questions are rated with “No” and “Probably No” to flag the potential for bias introduced by specific methodological aspects, “Yes” and “Probably Yes” indicating no evidence for bias, and “No Information” flagging insufficient information for adequate judgement (1, 2). Responses to signalling questions guide domain-level ROB judgements (“High”, “Low”, “Unclear”), which then inform the overall ROB rating of a study.

Most prediction model studies assessed in systematic reviews and other appraisals (e.g. for developing medical guidelines) receive “High” ROB ratings across different PROBAST domains (3-5). For example, this may be due to bias introduced by participant selection (i.e. PROBAST domain 1) or the analytical methods used to develop or validate the prediction model (i.e. PROBAST domain 4) (1, 2). A recent systematic review on prediction models regarding COVID-19 diagnosis and prognosis illustrates this issue: 545 of the 731 assessed prediction models received a “High” Overall ROB judgement using PROBAST (3).

Besides methodological concerns in prediction models and prediction model studies, other reasons for “High” ROB ratings using PROBAST have been explored. Langenhuijsen et al. described “ceiling effects”, i.e. the inability to distinguish between degrees of ROB ratings (5). That is, if a single signalling question receives a “No” rating, the domain may be rated as “High” ROB; similarly, if all signalling questions receive a “No” rating, the same applies (5). PROBAST guidance leaves whether a single “No” should lead to a “High” domain-level ROB rating up to the assessors and their specific context (1, 2). Therefore, determining if a “High” ROB judgement resulted from just a single signalling question is challenging without knowing how PROBAST assessments were operationalized (1, 2, 5). Conversely, PROBAST guidance for overall study ROB judgement are clear: if one ROB domain is rated “High”, the overall study ROB should be “High”; “Low” overall study ROB should only follow if all domains are rated “Low” (1, 2).

Several studies have reported poor inter-rater agreement and reliability amongst assessors using PROBAST (5, 6). Subjectivity and rater dependency are known pitfalls

of any ROB assessment (e.g. for randomized intervention trials, diagnostic test accuracy studies, and etiologic studies) and are not limited to PROBAST and prediction model studies (7, 8). Providing training sessions prior to using PROBAST showed improved agreement on domain- and overall prediction model study ROB judgements (6). As applies to all types and settings of ROB assessments, to prevent low inter-rater agreements and reliability, adequate guidance and trainings may be advisable before applying any ROB tool.

We conducted a comprehensive study using a case study of 2167 PROBAST assessments to (1) assess inter-rater reliability (IRR) and agreement in PROBAST assessments by identifying signalling questions with significant variation in ratings between different assessors of the same study and understanding the reasons for these discrepancies, (2) determine which PROBAST signalling questions contribute most significantly to domain-level ROB ratings, and (3) explore whether the use of consensus meetings between different assessors of the same study impacts PROBAST ratings.

METHODS

Case study

The case study comprised 2167 PROBAST assessments from a systematic review on prediction models for diagnosis and prognosis of COVID-19 (3). The review included the judgements of 27 pairs of assessors, who were international experts in prediction model studies or their reviews and had varying levels of experience using PROBAST (3). Each prediction model study was judged by two assessors using PROBAST. Details of the case study, as well as the instructions provided to assessors are part of the supplement (Section A & B).

Dataset description

The case study's data extraction sheet comprised two subgroups of data: the individual assessor ratings (where available) as well as the consensus ratings (i.e. the agreed ratings between the assessors after discussion). Each row in the data extraction sheet represented the assessment either of a single prediction model study (development or validation procedure), or of a single model from within a study including multiple models/procedures, or of both the model development and validation in one row if assessed together (i.e. hereafter referred to as 'mixed assessments').

We specifically focused on the ROB assessments using PROBAST, including responses to all 20 signalling questions, set into four domains *Participants*, *Predictors*,

Outcome, Analysis. PROBAST consists of two parts: 20 signalling questions for model development and 17 for model validation assessments. At the end, PROBAST includes an overall judgement on the ROB of the prediction model study (hereafter 'Overall ROB judgement') (1, 2). *Model development* refers to the actual development (or training) of the prediction model. *Model validation* refers to evaluating the predictive performance of a model using other data (external validation) (1, 2).

Abbreviations for the PROBAST signalling questions (referred to as items) can be found in Table 1.

Table 1 *Abbreviations of PROBAST items*

Domains	Abbreviations
<i>1. Participants</i>	
1. Were appropriate data sources used, e.g., cohort, RCT, or nested case-control study data?	1.1 Data source
2. Were all inclusions and exclusions of participants appropriate?	1.2 In/exclusion criteria
<i>2. Predictors</i>	
1. Were predictors defined and assessed in a similar way for all participants?	2.1 Predictor standardization
2. Were predictor assessments made without knowledge of outcome data?	2.2 Predictor blinding
3. Are all predictors available at the time the model is intended to be used?	2.3 Predictor availability
<i>3. Outcome</i>	
1. Was the outcome determined appropriately?	3.1 Outcome misclassification
2. Was a pre-specified or standard outcome definition used?	3.2 Outcome pre-specification
3. Were predictors excluded from the outcome definition?	3.3 Incorporation
4. Was the outcome defined and determined in a similar way for all participants?	3.4 Outcome standardization
5. Was the outcome determined without knowledge of predictor information?	3.5 Outcome blinding
6. Was the time interval between predictor assessment and outcome determination appropriate?	3.6 Time interval
<i>4. Analysis</i>	
1. Were there a reasonable number of participants with the outcome?	4.1 Sample size
2. Were continuous and categorical predictors handled appropriately?	4.2 Handling predictors

Table 1 *Abbreviations of PROBAST items (continued).*

Domains	Abbreviations
3. Were all enrolled participants included in the analysis?	4.3 Missing data
4. Were participants with missing data handled appropriately?	4.4 Missing data handling
5. Was selection of predictors based on univariable analysis avoided?*	4.5 Univariable predictor selection
6. Were complexities in the data (e.g. censoring, competing risks, sampling of control participants) accounted for appropriately?	4.6 Data complexity
7. Were relevant model performance measures evaluated appropriately?	4.7 Performance measures
8. Were model overfitting and optimism in model performance accounted for?*	4.8 Overfitting
9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?*	4.9 Predictor weights

*Model development assessments only

Statistical Analysis

Prior to all analyses and in accordance with the original guidance of PROBAST (1, 2), ratings of “Probably Yes” were merged with “Yes”, and “Probably No” with “No”, as they are treated similarly in PROBAST. All statistical analyses were performed using R Studio (Version 2023.12.0+369)(9).

Identifying variations in item- and domain-specific inter-rater reliability and agreement

We calculated the prevalence-adjusted bias-adjusted kappa (PABAK) to evaluate the IRR between different assessors of the same studies (10, 11). Though widely used, Cohen’s kappa may underestimate agreement in case of prevalence imbalances (10). Given the high proportion of “High” ROB ratings identified in the case study, we focused our IRR assessments on PABAK but calculated the Cohen’s kappa for reference (11).

We conducted these analyses for item-level ratings, as well as domain-level and Overall ROB judgements. PABAK and Cohen’s kappa values were classified as less than 0.21 indicating ‘slight’, 0.21 to 0.40 indicating ‘fair’, 0.41 to 0.60 indicating ‘moderate’, 0.61 to 0.80 indicating ‘substantial’, and more than 0.81 indicating ‘almost perfect’ agreement. For the absolute agreement between the assessors, we calculated counts

and proportions. Incomplete sets of assessments (i.e. where the rating of both or one of the assessors was missing) were excluded from the analyses.

Item-level ratings and domain-level ROB judgements

To evaluate how item-level ratings contribute to domain-level ROB judgements, we analysed “No” responses to the PROBAST items. We specifically examined “No” responses since these indicate the potential for bias and thus contribute to “High” domain-level ROB ratings (Supplement, Section B) (1, 2).

For our analysis, we considered a single PROBAST item to be “overruled” to inform domain-level ROB judgements if an item included a “No” response despite a “Low” or “Unclear” domain-level ROB judgement. Accordingly, an item contributed less to domain-level ROB judgements if it was frequently overruled (i.e. a high proportion of “No” responses despite a “Low” or “Unclear” domain-level ROB rating). Consequently, an item contributed more to domain-level ROB judgements if it had a low proportion of “No” responses in “Low” or “Unclear” domain-level ROB judgements but high proportions of “No” responses in “High” ROB domain-level judgements.

Besides, we assessed how many studies received at least one “No” response per domain-level ROB rating, to explore whether receiving a “No” response contributed to domain-level ROB judgements (giving an indication whether rating instructions were followed: *Did all domains rated with “High” ROB include at least one “No” response?*).

We calculated the counts and proportions of PROBAST items rated “No” per domain-level ROB rating. Consensus ratings were used for this analysis as they represented the most complete set and ensured a consistent interpretation of the scoring instructions.

Impact of consensus meetings on PROBAST ratings

To explore the impact of consensus meetings between different assessors of the same study on their PROBAST assessments, we evaluated 1) patterns of rating changes in PROBAST assessments from the assessor to the consensus rating and 2) whether an item-level rating change led to a domain-level ROB judgement change (i.e. domain-level shift). A *rating change* was present if both ratings of the assessors of the same study differed from the consensus rating. For example, if Assessor 1 scored item 1.1 Data source as “No Information” and Assessor 2 rated it as “No”, but after discussion during the consensus meeting, they agreed to rate it as “Yes”; this represents a rating change. Similarly, if both Assessor 1 and 2 initially rated the item as “No” but the consensus rating was “Yes”, a change in rating occurred. Furthermore, a domain-level shift occurred if, e.g. Assessor 1 rated item 1.1 Data source as “No Information” and Assessor 2 rated

1.1. as “No”, but the consensus rating included a “Yes”, this may result in a change on the domain-level. In this case, the domain-level ROB judgement may change from “Unclear” (Assessor 1) and “High” (Assessor 2) to a consensus rating of “Low” ROB.

We only considered complete assessment sets for the analysis, comprising two individual assessor ratings and one consensus rating per model. To identify patterns of rating changes, we calculated counts and proportions, with the numerator being the number of rating changes and the denominator being the number of models with complete rating sets. To evaluate domain-level shifts, we counted the number of item-level changes per domain and per domain-level changes.

RESULTS

Our analysis included 2167 assessments comprising 1407 individual assessor and 760 consensus ratings (for 113 consensus ratings, the individual assessments of one or both assessors were missing). The assessments comprised 760 different diagnostic or prognostic COVID-19 prediction models, including 384 model developments, 242 model validations, as well as 134 mixed assessments. Mixed assessments included model development and validation assessments in one data extraction row, which could not be separated out. These mixed assessments were analysed separately, and the supplement (Section C) presents their results.

Identifying variations in item- and domain-specific inter-rater reliability and agreement

Table 2 shows the IRR and agreement between the assessors of the same study using PABAK. Most model development items showed ‘moderate’ agreement, while most model validation items included ‘substantial’ agreement. The Overall ROB judgements for model development included ‘almost perfect’ agreement and ‘substantial’ agreement for model validations. In both model development and validation assessments, agreements were generally lower on domain-level compared to Overall ROB judgements. For model development, agreements ranged from ‘fair’ (Outcome) to ‘substantial’ (Analysis). In model validation assessments, agreements ranged from ‘moderate’ (Predictors) to ‘substantial’ (Participants & Analysis).

On item-level, three of the items included ‘slight’ agreement, i.e. item 4.6 Data complexity in both model development and validation, as well as item 3.5 Outcome blinding in model validation. Six items had ‘fair’ agreement, five in model development and one in model validation (i.e. 4.4 Missing data handling). ‘Moderate’ agreement

was identified for 12 items in model development and four in model validation (i.e. 2.1 Predictor standardization, 2.3 Predictor availability, 3.6 Time interval, 4.2 Handling predictors). Further, 'substantial' agreement was identified in a total of nine items (two in model development [i.e. 4.1 Sample size & 4.7 Univariable predictor selection]; seven in model validation). Only model validation items included 'Almost perfect' agreement: 1.1 Data source, 3.3. Incorporation bias, and 4.1 Sample size.

The supplement (Section D) includes the IRR based on Cohen's kappa. Cohen's kappa and PABAK showed similar IRR trends in model development assessments (both 'moderate') but lower for model validation assessments (PABAK: 'substantial'; Cohen's kappa: 'fair'). Furthermore, model developments included lower Overall ROB agreements using Cohen's kappa compared to PABAK (Cohen's kappa: 'fair'; PABAK: 'almost perfect'), but similar agreements in model validation assessments (both showing 'substantial' agreement).

Table 2 Inter-rater reliability (PABAK) and agreement between different assessors of the same prediction model study

Domains	Development				Validation			
	Models n*	Agreement %	PABAK	Interpretation	Missing ^a %	Agreement %	PABAK	Interpretation
<i>Domain 1: Participants</i>								
1.1 Were appropriate data sources used, e.g., cohort, RCT, or nested case-control study data?	284	74	0.47 (0.37; 0.57)	Moderate	26	91	0.81 (0.72; 0.91)	Almost perfect
1.2 Were all inclusions and exclusions of participants appropriate?	288	65	0.30 (0.19; 0.41)	Fair	25	89	0.77 (0.67; 0.87)	Substantial
Risk of bias introduced by participants or data sources	288	66	0.31 (0.2; 0.42)	Fair	25	88	0.76 (0.65; 0.86)	Substantial
<i>Domain 2: Predictors</i>								
2.1 Were predictors defined and assessed in a similar way for all participants?	285	75	0.51 (0.41; 0.61)	Moderate	26	78	0.55 (0.42; 0.69)	Moderate
2.2 Were predictor assessments made without knowledge of outcome data?	285	68	0.36 (0.25; 0.47)	Fair	26	88	0.76 (0.65; 0.86)	Substantial

Table 2 Inter-rater reliability (PABAK) and agreement between different assessors of the same prediction model study (continued).

Domains	Development				Validation			
	Models n*	Agreement %	PABAK	Interpretation	Missings ^a %	Agreement %	PABAK	Interpretation
2.3 Are all predictors available at the time the model is intended to be used?	285	76	0.53 (0.43; 0.63)	Moderate	26	78	0.56 (0.43; 0.70)	Moderate
Risk of bias introduced by predictors or their assessment	287	64	0.28 (0.16; 0.39)	Fair	25	73	0.47 (0.33; 0.61)	Moderate
Domain 3: Outcome								
3.1 Was the outcome determined appropriately?	284	77	0.55 (0.45; 0.65)	Moderate	26	84	0.69 (0.57; 0.81)	Substantial
3.2 Was a prespecified or standard outcome definition used?	287	79	0.57 (0.48; 0.67)	Moderate	25	89	0.77 (0.67; 0.87)	Substantial
3.3 Were predictors excluded from the outcome definition?	286	80	0.59 (0.50; 0.69)	Moderate	26	92	0.84 (0.75; 0.93)	Almost perfect
3.4 Was the outcome defined and determined in a similar way for all participants?	288	74	0.49 (0.39; 0.59)	Moderate	25	86	0.73 (0.62; 0.84)	Substantial

Table 2 Inter-rater reliability (PABAK) and agreement between different assessors of the same prediction model study (continued).

Development					Validation					
Domains	Models n*	Agreement %	PABAK	Interpretation	Missing ^a %	Models n*	Agreement %	PABAK	Interpretation	Missing ^a %
3.5 Was the outcome determined without knowledge of predictor information?	287	62	0.24 (0.13; 0.35)	Fair	25	147	58	0.16 (0; 0.32)	Slight	39
3.6 Was the time interval between predictor assessment and outcome determination appropriate?	287	72	0.44 (0.34; 0.55)	Moderate	25	148	78	0.55 (0.42; 0.69)	Moderate	39
Risk of bias introduced by the outcome or its determination	287	61	0.23 (0.11; 0.34)	Fair	25	148	84	0.69 (0.57; 0.81)	Substantial	39
Domain 4: Analysis										
4.1 Were there a reasonable number of participants with the outcome?	285	82	0.64 (0.55; 0.72)	Substantial	26	149	97	0.93 (0.88; 0.99)	Almost perfect	38
4.2 Were continuous and categorical predictors handled appropriately?	286	73	0.45 (0.35; 0.56)	Moderate	26	141	75	0.50 (0.36; 0.65)	Moderate	42
4.3 Were all enrolled participants included in the analysis?	284	68	0.37 (0.26; 0.47)	Fair	26	147	83	0.66 (0.54; 0.78)	Substantial	39

Table 2 Inter-rater reliability (PABAK) and agreement between different assessors of the same prediction model study (continued).

Domains	Development				Validation			
	Models <i>n</i> *	Agreement %	PABAK	Interpretation	Missing ^a %	Agreement %	PABAK	Interpretation
4.4 Were participants with missing data handled appropriately?	284	76	0.53 (0.43; 0.63)	Moderate	26	64	0.28 (0.13; 0.43)	Fair
4.5 Was selection of predictors based on univariable analysis avoided?	286	85	0.70 (0.62; 0.78)	Substantial	26	-	-	-
4.6 Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?	281	58	0.16 (0.04; 0.28)	Slight	27	52	0.04 (-0.13; 0.20)	Slight
4.7 Were relevant model performance measures evaluated appropriately?	286	74	0.48 (0.38; 0.58)	Moderate	26	84	0.68 (0.56; 0.80)	Substantial
4.8 Were model overfitting and optimism in model performance accounted for?	282	68	0.35 (0.25; 0.46)	Fair	27	-	-	-

Table 2 Inter-rater reliability (PABAK) and agreement between different assessors of the same prediction model study (continued).

Domains	Development				Validation					
	Models n*	Agreement %	PABAK	Interpretation	Missing ^a %	Models n*	Agreement %	PABAK	Interpretation	Missing ^a %
4.9 Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?	283	74	0.48 (0.38; 0.59)	Moderate	26	-	-	-	-	-
Risk of bias introduced by the analysis	287	84	0.68 (0.59; 0.76)	Substantial	25	141	88	0.76 (0.65; 0.87)	Substantial	42
Overall										
Reaching an overall judgment about risk of bias of the prediction model	281	91	0.82 (0.76; 0.89)	Almost perfect	27	146	89	0.78 (0.68; 0.88)	Substantial	40

*Number of models with individual and consensus ratings; IRR: Inter-rater reliability

PABAK: prevalence-adjusted bias-adjusted kappa (< 0.21: slight, 0.21 - 0.40: fair, 0.41 - 0.60: moderate, 0.61 - 0.80: substantial, 0.81 - 0.99: almost perfect, 1: perfect)
Prior to all analyses, the item-level rating "Probably No" was merged with "No", while "Probably Yes" was merged with "Yes".

^a Missing: Incomplete sets of assessments (i.e. where the rating of both or one of the assessors was missing)

Item-level ratings and domain-level ROB judgements

Table 3 shows the proportion of items scored with “No” per domain-level ROB judgement. Section E of the supplement includes the number of assessments per domain-level ROB rating with at least one item scored “No”.

The Participants domain had a “High” ROB in 29% (103/355) of model development and 15% (36/235) of model validation assessments. The highest “No” ratings in “High” ROB judgements were found for item 1.2 In/exclusion criteria (62% for model development; 83% for model validation). While none of the model development assessments included a “No” rating in the “Low” ROB ratings for item 1.2, one item received a “No” response among all model validation assessments (1/164). Additionally, 94% of “High” ROB model development assessments and all model validation assessments had at least one “No” response (Supplement, Section E).

The Predictors domain had a “High” ROB in 11% (41/360) of model development and 1% (2/233) of model validation assessments. For model developments, item 2.1 Predictor standardization received a “No” rating in 54% of assessments including a “High” ROB rating, while none did in “Low” ROB ratings. For model validation assessments, item 2.1 and item 2.2 Predictor blinding received “No” responses in 50% of the assessments (1 out of 2). Item 2.2 Predictor blinding never received a “No” response in assessments with “Low” ROB ratings. For “High” ROB ratings, 90% (37/41) of model developments included at least one “No” response, as did 50% of model validations (1/2) (Supplement, Section E).

The Outcome domain included “High” ROB in 19% (70/360) of the model development and 14% (34/237) of model validation assessments. The highest “No” ratings were found for item 3.5 Outcome blinding for the “High” ROB judgements (43% for model development; 76% for model validation). However, both included some “No” ratings in the “Low” ROB ratings. Interestingly, 86% of model development assessments judged with a “High” ROB included at least one item rated with “No”, while 97% of model validation assessments did (Supplement, Section E).

In the Analysis domain, 90% (314/351) of model development and 79% (178/226) of model validation assessments included “High” ROB judgements. Item 4.1 Sample size included the highest proportion of “No” ratings for assessments including a “High” ROB judgement (both 73%), while also including one “No” rating in the “Low” ROB judgements (14%; 4%). For “High” ROB ratings, 95% of model developments, and 81% of model validations included at least one item with a “No” response (Supplement, Section E).

Table 3 Number of items scored “No” per domain-level ROB judgement*

Domains	Development n(%) ^a			Validation n(%) ^a		
	High	Low	Unclear	High	Low	Unclear
<i>Participants</i>	<i>N=103</i>	<i>N=146</i>	<i>N=106</i>	<i>N=36</i>	<i>N=164</i>	<i>N=35</i>
1. Were appropriate data sources used, e.g., cohort, RCT, or nested case-control study data?	56 (54)	4 (3)	2 (2)	10 (28)	0 (0)	0 (0)
2. Were all inclusions and exclusions of participants appropriate?	64 (62)	0 (0)	5 (5)	30 (83)	1 (1)	10 (29)
<i>Predictors</i>	<i>N=41</i>	<i>N=188</i>	<i>N=131</i>	<i>N=2</i>	<i>N=160</i>	<i>N=71</i>
1. Were predictors defined and assessed in a similar way for all participants?	22 (54)	0 (0)	2 (2)	1 (50)	7 (4)	0 (0)
2. Were predictor assessments made without knowledge of outcome data?	13 (32)	4 (2)	3 (2)	1 (50)	0 (0)	0 (0)
3. Are all predictors available at the time the model is intended to be used?	5 (12)	4 (2)	1 (1)	0 (0)	0 (0)	0 (0)
<i>Outcome</i>	<i>N=70</i>	<i>N=140</i>	<i>N=150</i>	<i>N=34</i>	<i>N=173</i>	<i>N=30</i>
1. Was the outcome determined appropriately?	16 (23)	0 (0)	3 (2)	13 (38)	0 (0)	0 (0)
2. Was a prespecified or standard outcome definition used?	12 (17)	0 (0)	3 (2)	10 (29)	0 (0)	0 (0)
3. Were predictors excluded from the outcome definition?	18 (26)	4 (3)	1 (1)	16 (47)	1 (1)	3 (10)
4. Was the outcome defined and determined in a similar way for all participants?	22 (31)	0 (0)	7 (5)	12 (35)	0 (0)	0 (0)
5. Was the outcome determined without knowledge of predictor information?	30 (43)	9 (6)	14 (9)	26 (76)	7 (4)	4 (13)
6. Was the time interval between predictor assessment and outcome determination appropriate?	14 (20)	3 (2)	2 (1)	11 (32)	0 (0)	4 (13)
<i>Analysis</i>	<i>N=314</i>	<i>N=7</i>	<i>N=30</i>	<i>N=178</i>	<i>N=26</i>	<i>N=22</i>
1. Were there a reasonable number of participants with the outcome?	229 (73)	1 (14)	11 (37)	130 (73)	1 (4)	5 (23)
2. Were continuous and categorical predictors handled appropriately?	114 (36)	0 (0)	2 (7)	19 (11)	0 (0)	0 (0)
3. Were all enrolled participants included in the analysis?	80 (25)	3 (43)	1 (3)	80 (45)	3 (12)	10 (45)

Table 3 Number of items scored “No” per domain-level ROB judgement* (*continued*).

Domains	Development n(%) ^a			Validation n(%) ^a		
	High	Low	Unclear	High	Low	Unclear
4. Were participants with missing data handled appropriately?	121 (39)	2 (29)	2 (7)	87 (49)	3 (12)	8 (36)
5. Was selection of predictors based on univariable analysis avoided?	105 (33)	0 (0)	3 (10)	-	-	-
6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?	103 (33)	0 (0)	0 (0)	35 (20)	2 (8)	6 (27)
7. Were relevant model performance measures evaluated appropriately?	191 (61)	1 (14)	9 (30)	92 (52)	1 (4)	10 (45)
8. Were model overfitting and optimism in model performance accounted for?	204 (65)	1 (14)	7 (23)	-	-	-
9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?	22 (7)	0 (0)	1 (3)	-	-	-

Prior to all analyses, the item-level rating “Probably No” was merged with “No”, while “Probably Yes” was merged with “Yes”.

*The first row for each domain includes the number of assessments with a specific domain ROB rating (horizontally), following the number of items per domain ROB rating judged with “No” (vertically).

^aProportion: numerator: number of items scored “No” per domain-level ROB judgement; denominator: number of models per domain-level ROB judgement.

Impact of consensus meetings on PROBAST ratings

Patterns of rating changes

Table 4 shows the rating changes from the two individual assessor to the consensus rating. Rating changes ranged from 0 to 6%. The model development assessments included more rating changes compared to the model validation assessments. The Analysis domain included most rating changes, particularly in item 4.6 Data complexity (6%). However, fewer changes were identified on domain-level and none in the Overall ROB judgements.

Table 4 Rating changes from individual to consensus ratings

Domains	Development		Validation	
	<i>n</i> (%) ^a	Models <i>n</i> ^b	<i>n</i> (%) ^a	Models <i>n</i> ^b
<i>Participants</i>				
1. Were appropriate data sources used, e.g., cohort, RCT, or nested case-control study data?	2 (1)	252	0 (0)	149
2. Were all inclusions and exclusions of participants appropriate?	9 (2)	255	3 (1)	149
Risk of bias introduced by participants or data sources	15 (4)	265	4 (2)	149
<i>Predictors</i>				
1. Were predictors defined and assessed in a similar way for all participants?	6 (2)	253	2 (1)	147
2. Were predictor assessments made without knowledge of outcome data?	2 (1)	253	0 (0)	147
3. Are all predictors available at the time the model is intended to be used?	1 (0)	252	2 (1)	147
Risk of bias introduced by predictors or their assessment	13 (3)	270	4 (2)	147
<i>Outcome</i>				
1. Was the outcome determined appropriately?	2 (1)	252	0 (0)	148
2. Was a prespecified or standard outcome definition used?	3 (1)	254	1 (0)	148
3. Were predictors excluded from the outcome definition?	3 (1)	252	0 (0)	148
4. Was the outcome defined and determined in a similar way for all participants?	6 (2)	256	1 (0)	148
5. Was the outcome determined without knowledge of predictor information?	3 (1)	256	1 (0)	147
6. Was the time interval between predictor assessment and outcome determination appropriate?	4 (1)	255	0 (0)	148
Risk of bias introduced by the outcome or its determination	6 (2)	268	1 (0)	148
<i>Analysis</i>				
1. Were there a reasonable number of participants with the outcome?	20 (5)	250	1 (0)	149
2. Were continuous and categorical predictors handled appropriately?	21 (5)	252	2 (1)	140
3. Were all enrolled participants included in the analysis?	18 (5)	250	1 (0)	146
4. Were participants with missing data handled appropriately?	13 (3)	250	1 (0)	149
5. Was selection of predictors based on univariable analysis avoided?	16 (4)	250	-	-

Table 4 Rating changes from individual to consensus ratings (*continued*).

Domains	Development		Validation	
	<i>n</i> (%) ^a	Models <i>n</i> ^b	<i>n</i> (%) ^a	Models <i>n</i> ^b
6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?	22 (6)	247	6 (2)	141
7. Were relevant model performance measures evaluated appropriately?	17 (4)	251	1 (0)	150
8. Were model overfitting and optimism in model performance accounted for?	16 (4)	245	-	-
9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?	15 (4)	250	-	-
Risk of bias introduced by the analysis	9 (2)	261	3 (1)	141
<i>Overall</i>				
Reaching an overall judgment about risk of bias of the prediction model evaluation	0 (0)	254	0 (0)	143

^a Numerator: number of identified rating changes from individual to consensus rating (i.e. both individual model assessments had to differ from the consensus rating); denominator: Number of complete assessment sets

^b Number of models with a complete assessment set (i.e. two individual assessor ratings and one consensus rating)

Prior to all analyses, the item-level rating “Probably No” was merged with “No”, while “Probably Yes” was merged with “Yes”.

Domain-level shifts

The number of domain-level shifts ranged from none (Analysis domain in model validation assessments) to 10 (Participants domain in model development assessments) (Supplement, Section F). Most item-level rating changes occurred in model development assessments; however, relative to the number of item-level rating changes, most domain-level shifts were observed in model validation assessments (Supplement, Section F).

DISCUSSION

We performed a comprehensive study on the use and practical application of the Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) (1, 2) to assess diagnostic and prognostic prediction models using a case study of a large systematic review on COVID-19 prediction models (3). Our results show that the inter-

rater reliability (IRR) using PROBAST, based on the prevalence-adjusted bias-adjusted kappa (PABAK), was high for Overall ROB judgements but lacks agreement on some item-level ratings and domain-level ROB judgements. Furthermore, IRR was generally higher in prediction model validation assessments compared to prediction model development assessments. We identified certain PROBAST items that frequently contributed to domain-level ROB judgements. The use of consensus meetings between different assessors of the same study impacted item-level ratings to a larger extent than domain-level ROB judgements, while not showing any impact on Overall ROB judgements.

Comparison to other studies

Our results partly reflect findings reported by others (5, 6, 12). We observed varying IRR across PROBAST assessments but cannot confirm findings of consistently low IRR (5, 6). Use of the PABAK as opposed to the Cohen's kappa and differences in PROBAST operationalization (i.e. how domains should be rated) may be possible explanations for these variations (4, 10, 13).

We identified varying trends in IRR between model development and validation assessments. This distinction is often overlooked in prior research, where IRR calculations frequently combine data from various systematic reviews without distinguishing between these evaluation types (5, 6). Furthermore, due to the difference in assessors and research teams, the assessments may have been conducted with a different set of instructions (i.e. operationalizations), leading to greater variability in responses (i.e. lower IRR) (13). In our study, the assessors received the same set of instructions and consensus meetings were planned (per protocol) throughout the five assessment rounds to finalize ratings. Thus, subjectivity could be limited.

We found higher IRR between different assessors of the same study compared to previous reports (5, 6, 10). However, while we based our main IRR assessments on the PABAK, previous studies frequently used the Cohen's kappa. Cohen's kappa may underestimate IRR when a certain category is overrepresented (e.g. as in our case, a large proportion of "High" ROB ratings for most assessments of the Analysis domain) (3, 10). For reference, we included the IRR based on Cohen's kappa in the supplement (Section D). While the Cohen's kappa showed similar trends as by PABAK for some PROBAST items, others included lower IRR. Therefore, focusing on the PABAK to assess IRR seemed appropriate.

Adequate instructions and operationalization prior to any ROB assessments are key to ensuring consistency and reliability across assessments (7, 14, 15). In our study,

we identified certain PROBAST items that frequently contributed to domain-level ROB judgements, while others were frequently “overruled”. However, the instructions given in our case study may differ from those of other PROBAST assessments. It is difficult to evaluate the differences in how domain-level ROB judgements were operationalized across different reports without understanding the specific criteria assessors used. To ensure comparability of methodological quality assessments, reporting checklists (e.g. Preferred Reporting Items for Systematic reviews and Meta-Analysis [PRISMA]) might consider including operationalizations as a reporting item (16).

PROBAST presents a minimal set of methodological quality criteria which merely *guide* the assessments of prediction model studies for diagnosis and prognosis (1, 2). In our study, some PROBAST items frequently contributed to domain-level ROB judgements. Venema et. al published a short form of PROBAST in 2021 consisting of six selectively summarized items (12). Using short forms of PROBAST should be approached with caution, as they may falsely encourage assessors to completely disregard certain aspects of methodological quality (e.g. 1.1 Data source or 1.2 In/exclusion criteria [Domain 1: Participants]), which may severely impact a study's validity. This may especially be the case for inexperienced assessors or when using PROBAST to help guide the design of a prediction model study. PROBAST was developed to thoroughly assess the ROB in the entire prediction model study, addressing all relevant methodological aspects of a prediction model study.

As applies to all ROB tools, consensus meetings between assessors of the same study have a critical role in improving methodological quality assessments. In our study, rating changes occurred on both item- and domain-level judgements after consensus discussion and even in case of previous agreement between the individual assessors. Therefore, holding actual consensus discussions proved beneficial. These findings highlight the need for structured, well-facilitated consensus meetings between assessors for the same study.

Implications

To ensure consistency and reliability in ROB assessments, adequate training and standardized assessment procedures are essential. Although assessors were trained experts with varying levels of experience in using PROBAST, our analyses showed that certain PROBAST items would still benefit from more concise explanations and clear operationalization by the users prior to performing PROBAST assessments. Moreover, it is crucial to hold actual consensus meetings during ROB assessments. Even when

assessors initially agree on their individual ratings, discussions may still lead to changes in the domain-level ROB judgements.

These improvements have been incorporated into the recent development of PROBAST+AI (17, 18) (www.probast.org) which is a continuation and update of PROBAST-2019.

Limitations

Our study has several limitations. First, our case study included prediction models which were developed and/or validated during the global pandemic COVID-19. Due to the pandemic's urgency, development, validation, and PROBAST assessments were performed in a short amount of time. Furthermore, specific aspects (e.g. predicted outcomes) of the case study may have affected our results more than others. For example, in prediction model studies with an objective outcome (case study), such as mortality, PROBAST assessments will differ compared to those done in subjective outcomes, such as quality of life (QoL) (i.e. see item 3.5 Outcome blinding). Therefore, our findings might not reflect PROBAST assessments performed in other medical fields or settings.

Second, for a few items, individual assessments had high percentages of missing values. Though reviewers were asked to submit their individual ratings, this was not always done. It is possible that reviewers preferred to wait for the consensus meeting or consult an expert for difficult items, which may also explain why the consensus rating set was more complete than the individual assessments, and why certain items were missing more frequently than others.

Third, due to the high number of assessments, we were unable to critically evaluate any free-text comments assessors made in detail. Nevertheless, we randomly sampled 10 consensus ratings to explore any information assessors noted down. Notes primarily included reasons for specific domain-level ROB judgements.

Fourth, our results may be influenced by a certain “learning effect” that occurred in the multi-stage review of the case study (3). However, while some assessors were involved from the beginning, others only joined in later rounds, with varying workloads. Therefore, the “learning effect” likely only had a limited impact.

CONCLUSION

For Overall ROB judgements, PROBAST assessments include high inter-rater agreement and reliability. Some variation exists on item- and domain-level assessments. To ensure PROBAST assessments can be performed adequately and coherently, assessors involved in prediction model reviews or guideline development should therefore operationalize on how to judge PROBAST items and domains prior to conducting PROBAST assessments. We also encourage PROBAST users to hold well-structured consensus meetings between assessors of the same study. This will help to ensure that the PROBAST assessments are less subjective to variability. The findings of this study have led to the recent development of PROBAST+AI as an extension and update of PROBAST-2019.

Acknowledgements

None.

Authors' contributions

Tabea Kaul: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing – Original draft, Visualization, Project administration.

Johanna AA Damen: Conceptualization, Methodology, Investigation, Formal analysis, Writing – Review& Editing.

Laure Wynants: Methodology, Investigation, Writing – Review& Editing

Ben Van Calster: Methodology, Investigation, Writing – Review& Editing.

Marteen van Smeden: Methodology, Investigation, Writing – Review& Editing.

Lotty Hooft: Conceptualization, Methodology, Investigation, Writing – Review& Editing, Supervision. **Karel GM Moons:** Conceptualization, Methodology, Investigation, Writing – Review& Editing, Supervision.

Funding

None.

Competing interests

KGMM was involved in the development of the quality assessment tool included and described in this study. LW, BVC, KGMM, LH, MvS, and JAAG were involved in the quality assessment ratings included and described in our study. All other co-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.

Ethical approval

Not required for this work.

Registration and protocol

This study was not registered and does not include a publishable protocol.

Data availability

Data are available upon reasonable request to the corresponding author.

Supplementary data

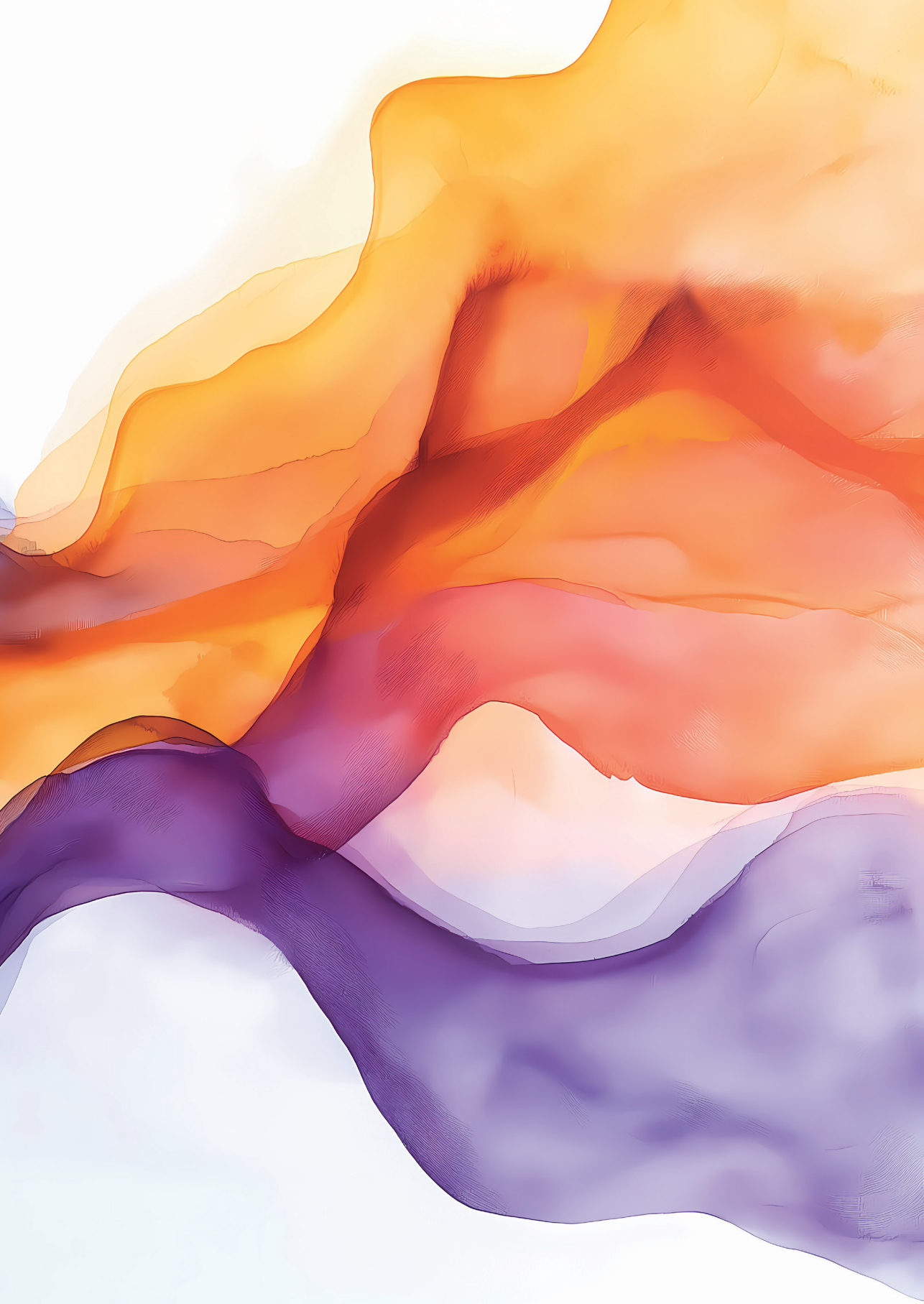
Supplementary data to this article can be found online.



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. *Annals of Internal Medicine*. 2019;170(1):W1-W33.
- (2) 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. *Annals of Internal Medicine*. 2019;170(1):51-8.
- (3) 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.
- (4) Kaiser I, Mathes S, Pfahlberg AB, Uter W, Berking C, Heppt MV, et al. Using the Prediction Model Risk of Bias Assessment Tool (PROBAST) to Evaluate Melanoma Prediction Studies. *Cancers*. 2022;14(12):3033.
- (5) Langenhuijsen LFS, Janse RJ, Venema E, Kent DM, van Diepen M, Dekker FW, et al. Systematic meta-review of prediction studies demonstrates stable trends in bias and low PROBAST inter-rater agreement. *Journal of Clinical Epidemiology*. 2023.
- (6) Kaiser I, Pfahlberg AB, Mathes S, Uter W, Diehl K, Steeb T, 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).
- (7) Tomlinson E, Cooper C, Davenport C, Rutjes AWS, Leeflang M, Mallett S, et al. Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies. *Journal of Clinical Epidemiology*. 2024;171.
- (8) Kanukula R, McKenzie JE, Cashin AG, Korevaar E, McDonald S, Mello AT, et al. Variation observed in consensus judgments between pairs of reviewers when assessing the risk of bias due to missing evidence in a sample of published meta-analyses of nutrition research. *Journal of Clinical Epidemiology*. 2024;166.
- (9) R Core Team. R: A language and environment for statistica computing. R version 2023.12.0+369, Vienna, Austria: R Foundation for Statistical Computing,; 2023 [cited 2024 31 January]. Available from: <https://www.R-project.org/>.
- (10) Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*. 1993;46(5):423-9.
- (11) Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960;20(1):37-46.
- (12) 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. *Journal of Clinical Epidemiology*. 2021;138:32-9.
- (13) Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *Journal of Clinical Epidemiology*. 2013;66(9):973-81.
- (14) Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. *Journal of Clinical Epidemiology*. 2022;141:99-105.
- (15) Crocker TF, Lam N, Jordão M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. *Journal of Clinical Epidemiology*. 2023;161:39-45.

- (16) 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. *Systematic Reviews*. 2021;10(1):89.
- (17) Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
- (18) Andaur Navarro CL, Damen JAA, Hooft L, Reitsma JB, van Smeden M, Riley RD, et al. Protocol for development of a risk of bias assessment tool for diagnostic and prognostic prediction studies based on artificial intelligence: PROBAST-AI 2022 [cited 25 August 2024]. Available from: <https://doi.org/10.17605/OSF.IO/4BJWT>.





PROBAST+AI: An Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods

Karel GM Moons, Johanna AA Damen, Tabea Kaul, Lotty Hooft,
Constanza Andaur Navarro, Paula Dhiman, Andrew L Beam, Ben Van Calster,
Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Marzyeh Ghassemi,
Georg Heinze, André Pascal Kengne, Lena Maier-Hein, Xiaoxuan Liu, Patricia Logullo,
Melissa D McCradden, Nan Liu, Lauren Oakden-Rayner, Karandeep Singh,
Daniel S Ting, Laure Wynants, Bada Yang, Johannes B Reitsma, Richard D Riley,
Gary S Collins, Maarten van Smeden

BMJ 388;2025:e082505

ABSTRACT

Background: Launched in 2019, PROBAST (Prediction model Risk Of Bias ASsessment Tool) is a tool to assess the risk of bias and applicability of prediction models and prediction model studies. Since then, there has been much progress on the methodology for prediction modelling in general and on the use of artificial intelligence (AI), including machine learning (ML) techniques, in particular.

Objectives: To develop PROBAST+AI, an updated tool for assessing quality, risk of bias, and applicability in all types of studies developing or evaluating (validating) multivariable diagnostic and prognostic prediction models using any data analytical (prevailing statistical or AI/ML) technique. Risk of bias refers to the potential for a systematic error (bias) in the estimators of a model's predictive performance for the target population(s) of interest.

Methods: The development of PROBAST+AI was informed by the original PROBAST tool and user experiences, by various systematic reviews and by recent methodological developments, followed by a 3-round Delphi survey among a large, global and diverse group of key stakeholders and experts on prediction modelling, AI/ML techniques and healthcare professionals, followed by a final consensus expert meeting. Participants gave their opinion on a large series of quality, risk of bias, and applicability items, using a 5-point Likert scale. Participants were also asked to suggest new items using free-text boxes.

Results: PROBAST+AI consists of two distinctive parts: model development and model evaluation. For model development, PROBAST+AI users assess the quality and applicability using 16 targeted signalling questions. For model evaluation, PROBAST+AI users assess the risk of bias and applicability using 18 targeted signalling questions. Both parts contain four domains: Participants and Data Sources, Predictors, Outcome, and Analysis. Applicability of the prediction model to the assessor's review question or to the assessor's intended use of the prediction model, is rated regarding the Participants and Data Sources, Predictors, and Outcome domains. PROBAST+AI also embeds signalling questions and aspects addressing prediction model fairness and algorithmic bias.

Conclusion: PROBAST+AI replaces the original PROBAST tool and provides a more focused tool that allows all key stakeholders (e.g., prediction model developers and companies, researchers, editors, reviewers, health care professionals, patients, institutional ethical review boards, healthcare guideline developers and health policy organisations) to examine the quality, risk of bias and applicability of any type of prediction model study in the healthcare sector, regardless which data analytical (prevailing statistical or AI/ML) technique is used.

INTRODUCTION

Prediction models or algorithms (hereafter referred to as prediction models) in healthcare estimate the probability of a diagnostic or prognostic outcome for individuals. In the diagnostic setting (including screening and monitoring), the model typically aims to predict or classify the presence of a particular outcome (e.g., disease or disorder). In the prognostic setting the aim is to predict a future (typically health-related) outcome, either in patients diagnosed with a certain disease or disorder, or in the general population. The primary use of a prediction model in healthcare is to support individual healthcare counselling and shared decision-making on, e.g., subsequent medical testing, referral to another healthcare professional or institute, treatment administration, discharge from hospitals, or lifestyle changes. For example, QRISK4 predicts the probability of developing a cardiovascular event over the next 10 years and informs whether an individual should undergo lifestyle changes or be prescribed medication (1). Prediction models are developed for and used in all healthcare settings and for all medical conditions to predict all types of outcomes. Thousands of models are published annually in the healthcare domain, often for the same target population, to predict the same health condition or the same types of outcomes (2-5). For example, within the first 15 months of the COVID-19 pandemic, 381 prognostic prediction models for COVID-19 were published (5).

For decades, traditional statistical modelling approaches, in particular regression modelling, have been the prevailing approaches when developing a prediction model. However, in more recent years, there has been an increasing interest in other analytical approaches such as artificial intelligence (AI) including machine learning (ML). Popular examples of such AI/ML methods are support vector machines, tree-based learning (e.g. random forests), and neural networks (including deep learning) (6). With increasing accessibility to software and high computational power (e.g., through cloud computing), developing prediction models in the healthcare domain with AI/ML methods has become overwhelming. The ease with which prediction models can now be developed has undoubtedly contributed to the vast numbers of prediction models we see in the biomedical literature. Also, changes in infrastructure for healthcare data, such as the increasing use of electronic health records, and the advancements in natural language processing to make use of the unstructured data from these records, resulted in the availability of large amounts of data further facilitating the development and training of prediction models, with both prevailing statistical and with AI/ML techniques.

Despite the abundance of prediction models available in the healthcare literature and guidelines, numerous reviews in the past two decades have shown that the majority of published prediction models, including AI/ML based prediction models, are of poor quality, the reported predictive performance is at high risk of bias (2-5, 7, 8), and fairness-related issues of prediction models may affect the predictive performance of models related to certain groups (9, 10). Prediction model studies including AI/ML based prediction model studies, also systematically suffer from overinterpretation (referred to as 'spin') regarding the applicability, validity and usefulness of the resulting models (11, 12). Further, poor science practices and inefficient translation of poorly performing models lead to research waste (13). This problem is compounded by the lack of scrutiny and oversight since prediction model use is largely unregulated and non-standardised. All this casts considerable doubt on the validity, accuracy and thus safety and applicability of prediction models in medical guidelines or healthcare practice, and potentially creates or further widens existing healthcare disparities (14, 15).

In response to these developments, the Cochrane Prognosis Methods group was established in 2008, with a focus on systematic reviews of prognosis studies (16-18). To facilitate the appraisal of prediction models included in systematic reviews, the Prediction model Risk Of Bias ASsessment Tool (PROBAST; www.probast.org) for the appraisal of model development and evaluation (validation) studies in the healthcare domain, was published in 2019 (19, 20). Risk of bias refers to the potential for a systematic error (bias) in the estimators of the model's predictive performance for the target population(s) of interest. Bias may act in either direction, such that there may be overestimation or underestimation of the true model performance. Applicability refers to whether the prediction model study in question is relevant for the assessor's review question or to the assessor's intended use of the prediction model including target population(s) and setting(s). Several tools are currently available for assessing methodological quality and applicability of diagnostic and prognostic tests and models, which were recently summarized accompanied by a decision tree to determine which quality assessment tool to use for which context, purpose and situation (21).

PROBAST thus assesses the risk of bias and applicability with 20 signalling questions across four domains: Participants, Predictors, Outcome, and Analysis (19,20). PROBAST enables a focused and transparent approach to assessing risk of bias and applicability of studies that develop, update or evaluate (validate) the performance of a prediction model. PROBAST was accompanied with a detailed Explanation and Elaboration document providing the rationale behind each domain and signalling question, examples of how to use the tool, and a discussion of issues causing risk of bias and

applicability concerns in prediction model studies (19). Additional guidance is available in other methodological papers (18, 22-25).

Advances in AI/ML methods and feedback from PROBAST users necessitated an update to allow additional data and prediction model quality appraisal, and to better address the methodological considerations for a broader set of modelling approaches given the uptake in applying AI/ML for prediction modelling. For example, inherently different approaches to handling predictors in tree-based learning (e.g., random forest) and neural networks, or the often wrongly used methods to address so-called class imbalances in a data set, were not addressed in the original PROBAST (hereafter referred to as PROBAST-2019). Moreover, in recent years, there have been important methodological advancements, including sample size guidance for both developing (26-30) and evaluating prediction models (31-33) using either regression based or AI/ML techniques. Finally, notably with the introduction of AI/ML-based prediction models there has been an important recognition of the concerns of prediction model (un)fairness, discrimination (6, 10, 11, 12, 34, 35), and reproducibility (36). An update of PROBAST-2019 was considered needed to reflect these latest developments, and capture the potential consequences for the quality, risks of bias and applicability assessment of prediction models in the healthcare domain, regardless the data analytical method that was used for the prediction modelling (i.e. either prevailing statistical or AI/ML techniques).

In this paper we describe the process of updating the PROBAST tool (referred to as PROBAST+AI), present the PROBAST+AI tool, and or ‘probably yes’ and teachers, in the critical appraisal of one or more prediction models without conducting an explicit systematic review (Table 1). A glossary of terms (Box 1) defines key concepts used within the field of prediction modelling.

Box 1: Glossary of terms (in alphabetical order)

<i>Algorithmic bias</i>
When the predictions or classifications by the algorithm (model) benefit or disadvantage certain groups of individuals more than others, without a justified reason for such unequal impacts.
<i>Apparent performance</i>
A type of a model performance evaluation. In apparent performance, model (prediction or classification) performance is estimated using exactly the same data as used for model development.
<i>Applicability</i>
Whether the study in question is relevant (applicable) for the assessor's review question or to the assessor's intended use of a model including target population(s) and setting(s).

Box 1: Glossary of terms (in alphabetical order) (*continued*).*Artificial intelligence*

There are many definitions of AI, some of which are very extensive and complex, see for instance the European AI act (37). In the context of prediction in healthcare, the term AI is commonly used for statistical learning approaches that do not fall under the family of generalized linear models (e.g. logistic regression) or survival modeling (e.g. Cox regression). For instance, analytical models are commonly referred to AI if they are based on support vector machines, tree-based learning (e.g. random forests) and neural networks (e.g. deep learning). In this context, machine learning (ML) is often used synonymously. However, any strict distinction made between statistical versus ML/AI quickly become a false dichotomy (38).

Calibration

The agreement between the model's estimated probabilities and the observed outcome probabilities. Calibration is typically assessed graphically using a plot of the observed outcome values on the y-axis and the estimated outcome values on the x-axis, with a calibration curve in the individual data.

Data leakage

When data from the model development phase is somehow inadvertently included during the model evaluation (testing) phase, typically leading to overly optimistic performance estimates or inaccurate predictions or classifications of the model. This occurs because the evaluated model has learned from information in the "leaked" data.

Development or training data

The data used to build or fit, referred to as develop or train, the prediction model.

Discrimination

How well the estimated outcome values from the model differ between those with observed outcome values. Discrimination is typically quantified by the c-index for time-to-event outcomes, and the c-statistic (sometimes referred to as the area under the receiver operating characteristic curve) for binary outcomes.

Evaluation or test data

The data used to estimate the (prediction or classification) performance of a prediction model, sometimes referred to as test data or validation data. Evaluation data should ideally be different from the data used to train the model, do model selection, or tune hyperparameters, such that there is no overlap in participants between the training and evaluation data (see also data leakage).

External validation

A type of model performance evaluation. In external validation, model performance is estimated using participant data that was not used for development (including internal validation) of the model.

Fairness

Property of prediction models that do not disadvantage particular groups of people based on characteristics such as age, sex/gender, race/ethnicity, or socioeconomic status.

Feature

Measurable property that is used as input for a prediction model, in this paper consistently referred to as predictor.

Hyperparameters

Values that control the model development or learning process.

Box 1: Glossary of terms (in alphabetical order) (*continued*).

Hyperparameter tuning

Finding the optimal settings of (hyper)parameters in the prediction model building strategy.

Imprecision

When a model's performance estimate is based on a small evaluation sample, leading to wide confidence intervals of the performance estimates.

Internal validation or evaluation

A type of model performance evaluation. The process of assessing a prediction model's performance using some form of splitting, resampling or cross-validation technique on the development data set.

Machine learning

A subfield of artificial intelligence that focuses on developing models that are capable of learning and making predictions or decisions from data, without being explicitly programmed. In the context of prediction models in health, often used synonymously with AI.

Model evaluation

Evaluating the predictive performance of a model by estimating its, for example, overall predictive accuracy (e.g., Brier score), model discrimination (e.g., *c*-statistic), model calibration (e.g., calibration plot, calibration slope), and clinical usefulness (e.g., decision curve analysis). Evaluation types can include the assessment of the model's apparent performance, internal validation performance, and external validation performance.

Outcome

The diagnostic or prognostic health state, value, or their probabilities that is being predicted. In machine learning, this is often referred to as the target value or response variable.

Predictor

A characteristic that can be measured or attributed at an individual level (such as age, sex, systolic blood pressure, disease stage, radiomics features), or group level (e.g., country). A predictor is also often referred to as a feature, input, independent variable, or covariate.

Risk of bias

The potential for a systematic error (bias) in the estimators of the model's predictive performance for the target population(s) of interest. Bias may act in either direction, such that there may be overestimation or underestimation of the true model performance.

Data pre-processing

Typical preparatory step for predictors before data analysis. For example, transforming a continuous predictor or outcome, (re-)categorising a predictor or outcome, or collapsing rare predictor or outcome categories.

Box 1: Glossary of terms (in alphabetical order) (*continued*).*Validation data*

This can have different meanings. Typically, in the medical literature it refers to data that is not used for model development (or training) but is only used to evaluate (validate) a model's (predictive or classification) performance, often referred to as external validation. The differences between internal and external validation are explained above and in the main text. In the computer science literature, it typically refers to 'held back data' that is used after the model development phase for (hyper)parameter tuning of the model, that will then go forward for a model's predictive performance evaluation. To avoid any ambiguity and harmonise terminology, we here refer with validation data to any data used to evaluate a model's performance.

These definitions relate to the specific context of PROBAST+AI and the use of these terms in the PROBAST+AI tool. They are not necessarily applicable to other areas of research. Developed based on Collins et al. TRIPOD+AI statement (60).

DEVELOPMENT OF PROBAST+AI

A working group with extensive experience in prediction model research (using AI/ML/statistical methods), systematic reviews and applying PROBAST-2019 was formed comprising KGMM, JAAD, TK, CAN, PD, LH, JBR, RDR, GSC and MvS to oversee the PROBAST+AI development process. The protocol for the update of PROBAST was published in the BMJ (39) and made available on the Open Science Framework (OSF) (40). In the summer of 2019, we announced on the PROBAST website (www.probast.org) the conduct of a large international project comprising a series of systematic reviews on the methodological (including risk of bias) and reporting quality of prediction model, including AI/ML based prediction model, studies published in the field of cancer as well as in the generic healthcare literature. Furthermore, after previous research on the inter-rater agreement with PROBAST-2019 (41, 42), we conducted a comprehensive study on the application of the tool (paper submitted). All these findings informed the update of PROBAST-2019 (6, 11, 12, 19, 20, 21, 43, 44). These reviews consistently showed that the vast majority of studies on AI-based prediction models were poorly reported and poorly conducted with the development and evaluation of the included models rated at high risk of bias (6, 11, 12, 43, 44). Many factors contributed to this, including (too) small sample sizes, poor handling of missing data, failure to deal with model overfitting, and lack of adequate assessment of predictive performance. The conclusions from these reviews are supported by many other prediction model reviews that have appeared in the past years (31, 45-58).

Ethical approval

This project was qualified as non-medical research involving human subjects (non-WMO) according to the Dutch Central Committee on Research Involving Human Subjects (CCMO) and formal ethical approval was waived. The modified Delphi survey procedure was approved by the data management board of the Julius Center of the University Medical Center Utrecht (Netherlands). The modified Delphi procedure was subject to regular monitoring. Delphi survey participants provided electronic informed consent prior to completing the survey. All participant data was pseudonymized and stored securely on a server of the University Medical Center Utrecht (59).

Patient involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for design or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to the patient community.

Candidate domain and signalling question list generation

We used PROBAST-2019 as starting point (19, 20). PROBAST-2019 included 20 signalling questions across four domains (Participants, Predictors, Outcome, and Analysis). We anticipated that these domains and signalling questions would already largely apply to AI-based prediction model studies as well as regression-based approaches.

In the autumn of 2020, we conducted a survey among over 50 authors who participated in the earlier mentioned living review on prediction models for COVID-19 (5). Each of these authors have applied PROBAST-2019 on multiple occasions, including on papers with AI-based prediction models. The survey asked about their experiences using PROBAST-2019; its applicability to AI-based prediction model studies; suggestions for improvement or changes in wording of existing domains and signalling questions; and whether domains or signalling questions needed to be deleted or added. All suggestions and recommendations were harmonised into an initial list of 26 candidate signalling questions distributed over four risk of bias domains, then labelled as: Participants and Data Sources; Predictors; Outcomes; Analysis. The domains Participants and Data Sources, Predictors, and Outcomes each included an applicability subdomain. This initial list served as the basis for the first online Delphi round.

Delphi participant recruitment

Participants were identified by the PROBAST+AI working group, from authors of relevant publications, via social media (e.g., X), and based on personal recommendations by Delphi participants (snowballing recruitment). Participants were recruited covering all key stakeholder groups listed in Table 1, from a range of settings (e.g., universities, primary care, hospitals, biomedical journals, non-profit and for-profit organisations), and having geographical and other aspects of diversity in mind.

Delphi process

The Delphi surveys were designed and shared electronically using the REDCap online platform (www.projectredcap.org), and later the Castor online platform (www.castoredc.com). Due to a system change within the University Medical Center, we were required to switch data capturing programs after Round 1.

The first round of the Delphi involved rating each of the initial 26 candidate items. For each signalling question, participants were asked whether they ‘strongly disagree’, ‘disagree’, ‘neither agree nor disagree’, ‘agree’, and ‘strongly agree’ to their inclusion in PROBAST+AI. Participants were also invited to comment on any domain, on the signalling questions, and to suggest new items. For an item to be included a level of agreement (response to ‘strongly agree’ and ‘agree’) of 80 percent or higher had to be achieved. For items that did not surpass the threshold of 80 percent but were identified as essential, we checked whether a reformulation of the signalling question, i.e., based on received comments and expert opinion, led to a higher result in the next round. The narrative responses were analysed by CAN, JAAD, and TK, and agreed upon by KGMM, MvS, JBR, LH, GSC, RDR, and PD. After each round, the aggregated quantitative agreements results were presented to the participants of the next round. Responses were anonymous.

Round 1

The first round of the Delphi was open between July 12th, 2021, and September 12th, 2021. A total of 201 individuals received an invitation to participate, of which 95 completed the survey (Supplementary Figure 1). Panellists were based in various countries and represented six continents.

Round 2

The second round of the Delphi was open between January 20th, 2023, and March 10th, 2023. All participants who completed the first Delphi round were invited to the second

Delphi round. Individuals who did not respond to the first round of the Delphi were reinvited. Additional participants were invited that were identified or recommended after the first round, to improve diversity of the participants either in their expertise (i.e., experts in AI, algorithmic bias and fairness) or geographical location. These additional participants were either identified by their participation in the TRIPOD+AI development (60) or they contacted us after some PROBAST+AI co-authors advertised the survey via X (formerly Twitter). A total of 294 individuals received an invitation to participate in round 2, of which 144 responded to the survey, including 12 who provided partial responses (Supplementary Figure 1, Supplementary Table 1).

In round 2, survey participants were given a summary of and link to the aggregated ratings from round one of the surveys (available at DOI 10.17605/OSF.IO/W3CFE). The second Delphi round included four domains, similar to PROBAST-2019 and the first survey round, although slightly reworded: Participants and Data Sources; Predictors; Outcome; Analysis. Major changes to the previous round were the more explicit distinction between model development and model performance evaluation, and the more distinguished focus on quality assessment (for model development) and risk of bias assessment (for evaluation of model performance) (Box 1). The number and type of signalling questions differed between model development (21 items) and model evaluation (22 items) (Supplementary Table 2).

Round 3

The third and final round of the Delphi was open between May 11th, 2023, and July 7th, 2023. All participants who completed the first or second Delphi round were invited to the third Delphi round. Individuals who did not respond to the previous rounds were reinvited, as well as participants identified or recommended after these rounds. All invitees received the aggregated responses of round 1 and 2. A total of 299 people received an invitation to participate in round 3, of which 131 responded to the survey (Supplementary Figure 1, Supplementary Table 1).

The list of signalling questions for Delphi round 3 for the model development phase included the same four domains as round 2, with 18 items for model development and 19 items for model evaluation (Supplementary Table 2).

Consensus meeting

A hybrid consensus meeting was held on October 17th, 2023, and was chaired by KGMM, LH, JAAD, PD, RDR, and MvS. Twenty-nine participants were invited to attend the consensus meeting of which 26 attended (Supplementary Figure 1). Participants

were identified to ensure a balanced representation of the key stakeholder groups and geographic and other aspects of diversity. In preparation of the final consensus meeting, KGMM, JAAD, TK, CAN, PD, LH, JBR, RDR, GSC and MvS developed a pre-final PROBAST+AI tool based on the input of the Delphi rounds in several hybrid meetings. Ten days prior to the consensus meeting, participants were emailed a pre-meeting document containing a brief overview of PROBAST+AI, format of and instructions for the consensus meeting, summary of the aggregated responses from the last round of the Delphi survey, and the draft PROBAST+AI tool. The tool shared with the consensus meeting participants included four domains and 16 signalling questions for quality assessment of model development, and the same four domains and 18 signalling questions for risk of bias assessment of the model evaluation. Given the high endorsement achieved for many items in round 3 of the Delphi survey, only a subset of 10 signalling questions were selected for plenary discussion and voting during the consensus meeting. The 10 items had either undergone rewording following round 3 of the Delphi survey or were new items introduced after round 3.

PROBAST+AI

After the consensus meeting, the working group developed the final PROBAST+AI tool (Supplementary Tables 3 and 4). The most noteworthy change in PROBAST+AI, compared to the PROBAST-2019 tool (20), is the more explicit distinction between signalling questions to assess the methodological quality of the process of model development versus assessment of the risk of bias in the evaluation of the model performance. A detailed rationale for this distinction can be found in the next section. Box 2 summarises noteworthy changes and additions to PROBAST-2019, such as this more explicit distinction between quality of model development and risk of bias in the model performance evaluation, as well as more explicit attention for algorithmic bias and fairness throughout the tool. Supplementary table 5 provides a more detailed comparison between PROBAST+AI and PROBAST-2019.

In summary (Table 2), PROBAST+AI includes four domains, 34 signalling questions (model development: $n=16$; model evaluation $n=18$) and 6 applicability items (model development: $n=3$; model evaluation: $n=3$). The first three domains share the same signalling questions for assessing either the quality of model development or risk of bias in model evaluation. All domains focus on quality and risk of bias and applicability concerns. Domain 1 (Participants and Data Sources) covers issues related to the participants and data sources used for model development or evaluation. Domain

2 (Predictors) covers the definition or measurement of predictors included in the prediction model development or evaluation, while Domain 3 (Outcome) covers the definition and measurement of the outcome predicted. Domain 4 (Analysis) deals with data analysis methods. It assesses aspects related to the (choice of) analysis method and whether key statistical considerations (e.g., handling of missing data) were correctly addressed. Domain 4 has five signalling questions to support the quality assessment for the model development and seven to support the risk of bias assessment for model evaluation. Detailed information on all items is available in the Explanation and Elaboration Light (PROBAST+AI E&E Light, Supplementary Table 4). The differences between PROBAST-2019 and PROBAST+AI can be seen side by side in Supplementary Table 5.

Box 2: Noteworthy changes and additions in PROBAST+AI

- Updated tool for assessing quality, risk of bias, and applicability, that covers prediction model studies regardless of the modelling approach applied, e.g., regression or AI/ML methods.
 - Model development and evaluation of model performance are more explicitly addressed as separate phases, with updated signalling questions.
 - PROBAST+AI distinguishes three types of model performance evaluation: apparent performance, internal validation, and external validation (Box 1).
 - Particular emphasis on ‘fairness’ and ‘algorithmic bias’, to assess whether specific methods were used to ensure fairness and address algorithmic bias (Box 1). Aspects of fairness and algorithmic bias are embedded throughout the signalling questions of the four domains of PROBAST+AI.
 - Harmonisation of nomenclature between different fields of expertise (e.g., statistics, AI/ML, epidemiology) (Box 1).
 - PROBAST+AI is not only useful for assessing prediction model studies when one aims to conduct a systematic review of prediction models, but also when one aims to appraise (the applicability, quality and risk of bias) of one or more prediction models, e.g. when developing medical or healthcare, guidelines, policy or recommendations.
-

Assessment of quality (model development) versus risk of bias (model evaluation)

In the original PROBAST tool (19, 20), signalling questions for model development and model evaluation could both be used to assess risk of bias that is, systematic error in the estimate of the model’s true predictive performance. With PROBAST+AI, we clarify that assessment of model development addresses quality, whereas assessments of model evaluation addresses bias. The former assesses the quality of how a prediction model is developed (model development), whereas the latter addresses the risk of bias in the predictive or classification performance of a developed model (model evaluation).

Model development is the actual process of constructing, producing or manufacturing a prediction model, from data collection and study design to fitting the model on data and producing or fitting the final model or algorithm. Every model is developed only once; it can be compared with the manufacturing or production of a medical test, device or drug. When methodological weaknesses or shortcomings in the design, conduct, and analysis of a model development, production or manufacturing process are present, this may lead to a prediction model with less reliable or accurate predictions and weak predictive performance when evaluated or applied in (data of) other individuals than used for the model development (19, 20). With PROBAST+AI, we introduce the concept of methodological quality (or simply quality) of the model development or production process: concerns about a lower quality of the model development process, as indicated by the signalling questions of the first part of PROBAST+AI, should thus be seen as a ‘red flag’ or raise concern about a poorly developed (manufactured) model.

Model evaluation is the process of estimating the model’s predictive performance, for example, in terms of calibration, discrimination, and net benefit. While a prediction model is developed or manufactured only once, it can and ideally should be more than once evaluated on its predictive performance, in participant data not used for model development. In other words, a particular model has only one model development process or study, but it can have multiple external evaluations or evaluation studies. This can also be compared to a manufactured medical test, device or drug that is manufactured only once, but can be evaluated on its accuracy or effectiveness multiple times. When methodological weaknesses or shortcomings in the design, conduct, and analysis of a model evaluation are present, reported model performance estimates may systematically differ from the true model performance (19, 20). Estimating model performance can be done in various ways (Box 1) (19, 24, 60): (i) using exactly the same participant data as that used for model development (i.e., apparent performance); (ii) using some form of splitting, resampling or cross-validation technique on the data of the development set (i.e., internal validation); or (iii) using different participant data than the development data set (i.e., external validation). Thus, the second part of PROBAST+AI assesses the risk of bias in the quantification or evaluation of the performance estimates of a prediction model (for each of the apparent, internal and external validation components, as relevant) by assessing the study’s design, conduct, and analysis using a series of signalling questions.

In the context of evaluating the performance of a prediction model, bias thus refers to systematic error in the estimates of the model’s true predictive performance

(19, 20). Bias may act in either direction, such that there may be systematic over- or underestimation of the true model performance. In the context of developing or manufacturing a prediction model, we cannot speak of the ‘true performance’ of that model, and it is thus more appropriate to speak of quality than of bias. A poorly developed prediction model (i.e., the development study was judged as having ‘low quality’) may likely have weak predictive performance (e.g., small sample sizes used to develop models tend to result in lower discrimination performance when the model is evaluated or applied in new individuals’ data). Models of lower quality may also be more susceptible to bias concerns in the performance of the model when evaluated or applied in new participant data, but that needs to be examined using the ‘evaluation’ component of PROBAST+AI. We note that bias must not be confused with imprecision, which arises when a model’s performance estimate is based on a small evaluation sample, leading to wide confidence intervals of the performance estimates (19).

Who should use PROBAST+AI and why

PROBAST+AI includes a quality appraisal of the model development process, provides a formal tool to assess risk of bias of a model’s performance evaluation (validation) study, and enables an assessment of the applicability of a prediction model to the intended purposes of the user of PROBAST+AI. PROBAST+AI is not only useful for (systematic) reviewers of prediction model (development or evaluation) studies, but for anyone who wants to appraise the applicability, quality and risk of bias of one or more prediction models (Table 1). PROBAST+AI is thus also useful for researchers or (medical technology or device) manufacturers aiming to develop or evaluate a prediction model with or without accompanying software; healthcare professionals determining whether or not to implement a prediction model in their daily medical practice; health policy regulators and guideline organisations appraising prediction models for their clinical guidance (such as the World Health Organisation (WHO), the Food and Drug Administration (FDA), the National Institute for Health and Care Excellence (NICE)); journal editors, reviewers and ethical review boards aiming to critically appraise prediction model studies, or anyone else who wants to judge the applicability and quality of a prediction model for their specific context, situation or purposes. Table 1 details the different purposes, reasons and for whom PROBAST+AI is useful.

Table 1 Purposes of PROBAST+AI: Users/stakeholders, actions and potential benefits.

User/stakeholder	Proposed action	Potential benefits
Academic institutions	• Promote or require adherence to PROBAST +AI by investigators developing, evaluating, assessing, or implementing prediction models. • Provide training to early-career researchers on the importance of methodological quality assessment, including requiring doctoral students to adhere to the quality criteria underlying PROBAST+AI.	• Enhances transparency in the design, analysis and reporting of prediction model research • Improves quality, accountability, reproducibility, replicability, fairness, and usefulness of produced research • Avoid research waste
Researchers	• Adhere to quality criteria underlying PROBAST+AI when developing or evaluating prediction models	• Improves methodological quality • Increases knowledge of the minimal quality criteria required and expected when performing a prediction model study • Improves accountability, reproducibility, replicability, fairness and usefulness of produced research • Avoid research waste
Systematic reviewers and meta-researchers	• Use PROBAST+AI to assess quality, risk of bias, and applicability of prediction models	• Can be used to assess study quality (e.g., design, methods) and applicability of prediction models • Increases trust in research findings • Improves quality, accountability, reproducibility, replicability, fairness, and usefulness of published research

Table 1 Purposes of PROBAST+AI: Users/stakeholders, actions and potential benefits. (continued).

User/stakeholder	Proposed action	Potential benefits
<i>Journal editors</i>	• Recommend or mandate authors to use PROBAST+AI, and submit a completed methodological quality assessment as part of a systematic review/ meta-analysis • Recommend peer reviewers to use PROBAST+AI to assess the methodological quality of prediction models for studies developing/ validating a model	• Improves understanding of journal requirements and expectations for prediction model publications • Increases efficiency of peer review resulting from improved author understanding of journal requirements for prediction model publications. • Improves quality, accountability, reproducibility, replicability, fairness, and usefulness of published research
<i>Peer reviewers</i>	• Use PROBAST+AI to assess the methodological quality of prediction models	• Improves efficiency of peer reviews • Facilitates and directs specific feedback to authors on where important details are missing
<i>Commercial manufacturers of prediction models</i>	• Adhere to quality criteria underlying PROBAST+AI when developing or evaluating a particular prediction model.	• Improves methodological quality • Increases awareness of the minimal quality criteria required and expected when developing a prediction model • Improves accountability, reproducibility, replicability, fairness, and usefulness of produced models
<i>Funders</i>	• Recommend or mandate use of quality criteria established in PROBAST+AI by investigators when reviewing a grant for prediction model research	• Increases usefulness and fairness of research findings • Reduces avoidable research waste due to inadequate methodological quality • Ensures that funded research can be used by others
<i>Policy-makers</i>	• Use or promote PROBAST+AI to ensure research is methodologically sound	• Ensures decisions to implement a prediction model are based on adequate methodology • Adds integrity for evidence-based policy recommendations

Table 1 Purposes of PROBAST+AI: Users/stakeholders, actions and potential benefits. (*continued*).

User/stakeholder	Proposed action	Potential benefits
<i>Regulators</i>	Clinical reviewers use PROBAST+AI to assess adequate methodological quality for ‘software as medical device’ regulatory submissions where the operating principle of the product is a prediction model	- Aligns reported intended use with regulatory intended purpose - Aligns medical device regulatory review with pivotal investigational reporting - Encourages manufacturers to publish clinical investigation reports by encouraging one common standard
<i>Healthcare professionals</i>	Verify whether a prediction model meets methodological standards and whether the model is applicable before purchasing or using a model to support clinical use	- Improves understanding of the target population of a model and the clinical decision it is intended for - Improves understanding of model predictions and awareness of limitations - Improves trust in research findings
<i>Institutional Ethical review boards</i>	Verify whether a proposal for a prediction model (study) meets the required methodological standards	Improves quality of prediction model development and evaluation studies
<i>Patients, public, study participants</i>	Understand and advocate use of PROBAST+AI by authors, peer reviewers, journals and funders.	- Improves trust in research findings - Improves understanding of prediction model research - Promotes health equity considerations in research

Developed based on PROBAST-2019 (19, 20) and on Collins et al. TRIPOD+AI statement (60).

Steps of using PROBAST+AI (Supplementary Table 3)

The PROBAST+AI tool can be used regardless of the modelling approach, prevailing statistical or AI/ML techniques, used for the model development. PROBAST+AI therefore outpaces PROBAST-2019 (19, 20). The PROBAST-2019 Explanation and Elaboration document (19) remains the comprehensive background document of PROBAST+AI and serves as an important pedagogical document to provide rationale and examples for the majority of the PROBAST+AI items. In Supplementary Table 4, we provide a bullet-point structure for each signalling question including a brief Explanation and Elaboration (Explanation and Elaboration Light) to facilitate implementation of PROBAST+AI. Differences in item scoring between traditional regression-based models and models

based on AI/ML techniques are, where needed, also described in the Explanation & Elaboration Light.

Using PROBAST+AI involves the same four steps as in PROBAST-2019 (19, 20) and explained in Supplementary Table 3:

1. Specify the intended purpose of the prediction model assessment or prediction model systematic review;
2. Classify the type of prediction model study;
3. Assess the quality (model development), risk of bias (model performance evaluation) and applicability (to the intended purpose) of the prediction model, for the separate domains;
4. Assess the overall quality, risk of bias and applicability of the prediction model.

Step 1: Specify the intended purpose of the prediction model assessment or of the prediction model systematic review

Conform PROBAST-2019 (19, 20), when using PROBAST+AI we advise to specify the purpose of the assessed prediction model(s). For this we recommend to define the PICOTS (Population, Index model(s), Comparator model(s), Outcome(s), Timing, Setting and intended use of the prediction model) criteria as provided by the guidance of the Cochrane Prognosis Methods group (<https://methods.cochrane.org/prognosis/>) and described in the CHARMS checklist (Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (61)). Defining the PICOTS directly indicates the aim of the assessment or review of the prediction models.

Step 2: Classify the type of prediction model study

As explained, prediction model studies can include model development, model evaluation or a combination of both (19, 20, 60). PROBAST+AI includes different signalling questions depending on the type of prediction model study. Therefore, we recommend assessors and reviewers state whether the study addresses model development only, model evaluation only, or both model development and model evaluation. Furthermore, as described in Box 1, model evaluation distinguishes between estimation of the model's apparent performance, internal validation performance and external validation performance. When a publication focuses on updating a previously developed model, e.g., by adding one or more new predictors, the model development part of PROBAST+AI should be used. When a publication focuses on only evaluating the

performance of an existing model in other (external) participant data, only the model evaluation part should be used.

Step 3: Assess the quality, risk of bias, and applicability of the prediction model for each domain

Step 3 aims to identify areas where concerns regarding quality and risk of bias may be introduced in the prediction model study, or where concerns regarding applicability may exist (19, 20). For each domain the quality (for model development) and risk of bias (for model evaluation) assessment comprises four sections (see Table 2, Supplementary Table 3 and 4 with detailed guidance for use): 1. General information from the study or model to support answering the signalling questions of that domain; 2. Answering the signalling questions; 3. A judgement of concerns regarding quality or risk of bias per domain; and 4. Rationale for the overall quality or risk of bias judgement of the prediction model (study).

Alike with PROBAST-2019 (19, 20), in the 'rationale for judgement' box, assessors can record any additional information used to answer the signalling questions. Signalling questions are answered as 'yes', 'probably yes', 'no', 'probably no', 'no information' or, where appropriate, 'not applicable'. Quality concern is judged as 'low', 'high', or 'unclear', and risk of bias is judged as 'low', 'high', or 'unclear'. All signalling questions are phrased such that 'yes' or 'probably yes' indicate low concern for quality (= high quality) or low risk of bias. Any signalling question answered as 'no' or 'probably no' flags the potential for quality concerns or bias. Subsequently, assessors need to use their judgement to determine based on these answers whether the *entire domain* should be rated as 'low', 'high', or 'unclear' quality concern or risk of bias. If a signalling question is answered with 'no' that does not automatically result in a high concern for quality or risk of bias rating of the entire domain. The 'no information' category should be used only when reported information is insufficient to permit a judgement. The 'not applicable' category may be available for items which are not applicable for certain types of models or situations. When the rationale for the overall domain judgement is recorded, the domain's quality or risk of bias rating will be more transparent. This can also facilitate discussion among different reviewers or assessors who complete assessments independently.

The first three domains, conform PROBAST-2019 (19, 20), also include assessment of concerns regarding the applicability of the prediction model (low, high, unclear) to the prediction model review question or to the assessors' intended use of the assessed model(s). Applicability is defined as any concern that either the included (data of the) participants and setting (Domain 1); or the definition, pre-processing, assessment

or timing of assessment of the predictors (Domain 2); or the outcome definition, assessment or timing of assessment (Domain 3) do not match the prediction model review question or the assessor's intended use of the model. Accordingly, applicability refers to both applicability of a study to the question of the reviewer (for example when one conducts a systematic review of prediction model studies) or when one assesses whether a particular developed or evaluated model is indeed useful for the intended use or purpose of the assessor. For example, a model can have a low concern for quality if data and participant selection were appropriate for the modeller's intended purpose, but a high concern for applicability if this does not match how the reviewer intends to use that model.

Table 2 PROBAST+AI: Summary of Step 3—Assessment of quality, risk of bias and concerns regarding applicability.

Model development			
Participants and data sources	Predictors	Outcome	Analysis
<i>Signalling questions</i>			
1.1 Were appropriate data sources used?	2.1 Were predictors defined and assessed in a similar way for all participants?	3.1 Were outcomes defined and assessed appropriately?	4.1 Was there evidence that the sample size was reasonable?
1.2 Was an appropriate study design used?	2.2 Was any pre-processing of predictors similar for all participants?	3.2 Were outcomes defined and assessed in a similar way for all participants?	4.2 Were continuous and categorical predictors handled appropriately?
1.3 Did the in- and exclusions of study participants result in a representative dataset?	2.3 Were predictor assessments made without knowledge of outcome data?	3.3 Were outcome assessments made without use or knowledge of predictor data?	4.3 Were participants with missing or censored data handled appropriately in the analysis?
	2.4 Were the predictors included in the model available at the time the model was intended to be used?	3.4 Was the time interval between predictor assessment and outcome assessment appropriate?	4.4 If methods to address class imbalance were used, was the model or the model predictions recalibrated?
			4.5 Were methods used to address potential model overfitting?
<i>Quality</i> Concern regarding quality of selection of participants and data sources	Concern regarding the quality of the predictors or their assessment	Concern regarding quality of the outcome or its determination	Concern regarding quality of the analysis
<i>Applicability</i> Concern that the (data of the) included participants do not match the review question or the assessor's intended use of the prediction model	Concern that the definition, pre-processing, assessment, or timing of assessment of the predictors in the model do not match the review question or the assessor's intended use	Concern that the outcome, its definition, assessment, or timing of assessment do not match the review question or the assessor's intended use	

Table 2 PROBAST+AI: Summary of Step 3—Assessment of quality, risk of bias and concerns regarding applicability. *(continued)*.

Model development			
Participants and data sources	Predictors	Outcome	Analysis
Signalling questions			
1.1 Were appropriate data sources used?	2.1 Were predictors defined and assessed in a similar way for all participants?	3.1 Were outcomes defined and assessed appropriately?	4.1 Was model evaluation based on only apparent performance avoided?
1.2 Was an appropriate study design used?	2.2 Was any pre-processing of predictors similar for all participants?	3.2 Were outcomes defined and assessed in a similar way for all participants?	4.2 Was there evidence that the sample size was reasonable?
1.3 Did the in- and exclusions of study participants result in a representative dataset?	2.3 Were predictor assessments made without knowledge of outcome data?	3.3 Were outcome assessments made without use or knowledge of predictor data?	4.3 Were participants with missing or censored data handled appropriately in the analysis?
	2.4 Were the predictors included in the model available at the time the model was intended to be used?	3.4 Was the time interval between predictor assessment and outcome assessment appropriate?	4.4 If methods to address class imbalance were used, was the evaluation done in a dataset without imbalance correction?
			4.5 If data splitting was done to create training and test datasets, was there evidence that data leakage was avoided?
			4.6 If resampling methods were used to evaluate model performance, were all model development steps replicated in the resampling process?

Table 2 PROBAST+AI: Summary of Step 3—Assessment of quality, risk of bias and concerns regarding applicability. *(continued)*.

Model development			
Participants and data sources	Predictors	Outcome	Analysis
			4.7 Was the predictive performance of the model evaluated appropriately, e.g., calibration, discrimination, and net benefit?
<i>Risk of bias</i> Risk of bias introduced by the selection of participants and data sources	Risk of bias introduced by the predictors or their assessment	Risk of bias introduced by the outcome or its determination	Risk of bias introduced by the analysis
<i>Applicability</i> Concern that the (data of the) included participants do not match the review question or the assessor's intended use of the prediction model	Concern that the definition, pre-processing, assessment, or timing of assessment of the predictors in the model do not match the review question or the assessor's intended use	Concern that the outcome, its definition, assessment, or timing of assessment do not match the review question or the assessor's intended use	

Developed based on PROBAST-2019 (19, 20). For further details, please see the PROBAST+AI Explanation and Elaboration Light (Supplementary Table 4), and PROBAST-2019 Explanation and Elaboration (19). Signalling questions are answered as yes, probably yes, probably no, no, no information, or not applicable. Quality concerns, risk of bias and applicability concerns are rated as low, high, or unclear.

Step 4: Overall quality, risk of bias, and applicability judgement

The final step of the PROBAST+AI tool is similar to PROBAST-2019 (19, 20), in which the four domain ratings are combined into an overall judgement on the quality (for model development), risk of bias (for model evaluation) and applicability of the prediction model for the prediction model review question or the assessor's intended use of the model. This overall judgement is scored as either low, high, or unclear concern of the quality; low, high, or unclear risk of bias; and low, high, or unclear concern of the applicability. Supplementary Table 3 (step 4) provides guidance on how to make

such overall judgement on quality, risk of bias and applicability, as well as the original PROBAST-2019 guidance (19, 20).

For example, a high overall concern of the applicability (both for the model development or model evaluation part) indicates a limited or poor applicability of the scored model for the review question or the assessor's intended purpose of the model, whereas a low overall concern of the applicability indicates a good applicability of the scored model (19, 20). Similarly, for the quality judgement of the model development, an overall high concern indicates a low quality of the model development (production) process, whereas an overall low concern indicates a high quality of the model development process. And similarly for the risk of bias judgment of the model evaluation, a low risk of bias indicates that the reported model performance estimates are valid (unbiased), whereas a high risk of bias indicates the performance estimates may systematically differ from the true model performance. For all three (i.e., judgement of applicability, quality or risk of bias), an unclear overall judgement indicates that there was insufficient reported information to make an adequate judgment.

These overall judgements may sometimes involve changing or reclassifying from overall high concern to low concern (for the quality of the model development process) or from high risk of bias to low risk of bias (for the model performance estimates) (19, 20). For example, for the model evaluation part, if a model performance was evaluated without any external performance evaluation, one might have scored domain 4 as high risk of bias. However, if the model was (typically in that same study) developed (i.e., fitted) on a very large data set and evaluated with some form of internal model performance validation, one may change this high risk of bias to an overall low risk of bias rating, provided that the other three domains of course had low concern regarding quality and risk of bias.

Multiple PROBAST+AI assessments and extending answers from model development to model evaluation

When a study reports the development and evaluation of more than one prediction model, all domains should be completed for each distinct prediction model (19, 20). The same publication may even address model development, its evaluation with apparent performance estimates, with its performance estimates after internal validation, and its performance with some form of external validation. Also, the same report may describe the development or evaluation of multiple models. We recommend that for each model a separate PROBAST+AI assessment is done, which is conform PROBAST-2019.

Note that for studies only developing models, usually both parts of PROBAST+AI still have to be completed. In these instances, the responses to the signalling questions (notably in the domains Participants and Data Sources, Predictors, and Outcome) will be the same for the development, the apparent performance evaluation and the interval validation (based on resampling techniques): accordingly, the answers across the three can easily be copied and pasted.

However, if in study in which a prediction model was developed and also external data sources were used for evaluation of its performance or accuracy, the responses for the first three domains may differ.

Finally, a model evaluation study only may describe the evaluation of one or more prediction models that have been developed in other previously published development studies (19, 20, 60). In these instances, only the second part of PROBAST+AI needs to be filled in, though for each evaluated (validated) model separately, where often the responses to many signalling questions will be the same for each assessment and can be copied.

The signalling questions of PROBAST+AI have just as PROBAST-2019 (19, 20) a natural order, as is roughly encountered when reviewing a prediction model report or study, although this may depend on journal formatting policies. The items and issues addressed by the domains and signalling questions of PROBAST+AI are the minimal and most essential issues to be assessed. Users may always assess additional aspects of a study to obtain a better view of the quality, risk of bias, or applicability of a prediction model.

DISCUSSION

PROBAST+AI has been developed through a comprehensive, phased and international multi-stakeholder consensus process. It provides explicit criteria for assessing the methodological quality, risk of bias, and applicability of studies or reports describing the development or evaluation of prediction models using any data analytical (i.e. prevailing statistical or AI/ML) method. Noticeable changes in PROBAST+AI (Box 2) are a clear distinction between the concepts of quality in the model development process and risks of bias in the estimates of the model's performance, and a more explicit emphasis on fairness and algorithmic bias throughout the tool, as described in the PROBAST+AI Explanation and Elaboration Light (Supplementary Tables 3 and 4). PROBAST+AI is a direct extension and update of PROBAST-2019 (19, 20 + Supplementary Table 5), but

because of these specific changes and additions, PROBAST+AI outpaces and replaces PROBAST-2019.

We illustrated that methodological quality of a model development or manufacturing process is different from a risk of bias assessment in the evaluation or quantification of the model's performance. The assessor may have high confidence in the performance estimates (low risk of bias) from a well conducted evaluation of an initially poorly developed model (low quality). But, it is also possible that an assessor is sceptical (high risk of bias) about the performance estimates in a poor evaluation study of a previously well-developed model (high quality). We have emphasized that a single study may include a model development and several types of performance evaluations of that same model, i.e., using either the same participant data for development and performance evaluation or using different (external) data for the performance evaluation (19, 20, 60). PROBAST+AI is to be used to assess both the methodological quality of a model development process and to assess the risks of bias in the evaluated predictive performance estimates.

We further stress that with improved data infrastructures and thereby increasing availability of data that were not collected primarily for research purposes (e.g. data from administrative patient registries, electronic health records or from other so-called real world contexts), it has become more important that not only the methods used to develop or evaluate a prediction model should be assessed, but also the quality of the source of data and the inclusiveness or fairness of the relevant individuals in the data set. Fairness in prediction model research is particularly important in healthcare, also or perhaps certainly when AI/ML methods are used to develop or evaluate the models (60, 62). Fairness means that prediction models are designed and used avoiding adverse discrimination against any particular group of individuals and not perpetuating any disparities in healthcare provision and patient or citizen outcomes (62). One important aspect of fairness is ensuring that the data used to develop or evaluate prediction models is diverse and representative. The STANDING Together initiative has developed standards for data diversity, inclusivity, and generalisability (35). This means that the used data sources should include information from individuals representing relevant axes of diversity for the use case, such as those of different ages, sex/genders, race/ethnicities, as well as individuals with different health conditions/comorbidities and potentially from different geographic locations that are representative of the target population in whom the prediction model is intended to be used. If the data used to develop the model is not diverse and representative, the resulting model may not be effective or fair, and thus not applicable, for the individuals the model is intended

to be used (by the assessor of the model). Furthermore, if the data used to evaluate a model is not representative of the targeted context of the assessor of the model, the estimates of predictive performance in particular subgroups may be misleading. PROBAST+AI has therefore stressed more explicitly aspects on fairness of the model development and evaluation throughout the four domains, to ensure due consideration in the appraisal of prediction models and prediction model studies. While algorithmic bias and fairness assessments for the model development and model evaluation procedure are crucial, they should always be approached with caution. Algorithmic bias and fairness-related issues may not be identified by exploratory data analysis alone. The ultimate assessment of algorithmic bias and fairness occurs when the model is deployed in daily healthcare practice (10).

CONCLUSION

We anticipate that PROBAST+AI will help all stakeholders (e.g., prediction model developers and companies, researchers, editors, reviewers, health care professionals, patients, ethical review boards, guideline developers and health policy organisations), encountering prediction models in the healthcare sector understand and appraise the quality, risk of bias, and applicability of healthcare prediction models and their studies. Using PROBAST+AI also to guide the design of a prediction model development and/or evaluation (validation) study will help to reduce research waste while improving the accuracy, effectiveness, generalisability, and appropriate use and fairness of prediction models in any healthcare setting or domain where prediction or classification plays a role, and regardless of the used data analytical modelling (i.e. prevailing statistical or AI/ML) technique.

Acknowledgements

PROBAST+AI Authors: Karel Moons (UMC Utrecht, The Netherlands), Maarten van Smeden (UMC Utrecht, The Netherlands), Richard Riley (University of Birmingham, UK), Gary Collins (University of Oxford, UK), Paula Dhiman (University of Oxford, UK), Johannes Reitsma (UMC Utrecht, The Netherlands), Johanna Damen (UMC Utrecht, The Netherlands), Tabea Kaul (UMC Utrecht, The Netherlands), Lotty Hooft (Cochrane Netherlands, The Netherlands), Constanza Andaur Navarro (UMC Utrecht, The Netherlands), Bada Yang (UMC Utrecht, The Netherlands), Andrew Beam (Harvard School of Public Health, US), Ben Van Calster (KU Leuven, Belgium), Leo Celi (Massachusetts Institute of Technology, US), Spiros Denaxas (University College

London, UK), Alastair Denniston (University of Birmingham, UK), Marzyeh Ghassemi (Massachusetts Institute of Technology, US), Georg Heinze (Medical University of Vienna, Austria), André Pascal Kengne (University of Cape Town, South Africa), Xiaoxuan Liu (University of Birmingham, UK), Patricia Logullo (University of Oxford, UK), Lena Maier-Hein (German Cancer Research Centre, Germany), Melissa McCradden (The Hospital for Sick Children, Canada), Nan Liu (Duke-NUS Medical School, Singapore), Lauren Oaken-Rayner (University of Adelaide, Australia), Karandeep Singh (University of Michigan, US), Daniel Ting (Stanford University, US), Laure Wynants (KU Leuven, Belgium).

Delphi Survey participants: We would like to thank the PROBAST+AI Delphi panel members for their time and valuable contribution in helping to develop PROBAST+AI Statement. We also acknowledge all contributors to the original PROBAST guidance. The full list of Delphi participants are as follows: Jose A Calvache, Elie Akl, Elena Albu, Lucy Archer, Sarah Barman, Valentina Bellini, Laura Bonnett, Patrick Bossuyt, Anne-Laure Boulesteix, Randy Boyes, Peter- Bram 't Hoen, Danilo Bzdok, Jennifer Camaradou, Guido Camps, Jonathan Chen, Evangelia Christodolou, Jeremie Cohen, Darren Dahly, Maarten De Vos, Thomas Debray, Jon Deeks, Andre Dekker, Jac Dinnes, Edgar Efrén Lozada Hernández, Joie Ensor, Ari Ercole, Andre Esteva, Ji Eun Park, Lavinia Ferrante di Ruffano, Alan Fraser, Shan Gao, Geert-Jan Geersing, Bart Geerts, Robert Golub, Benjamin Gravesteijn, Olivier Groot, Saskia Haitjema, Michael Harhay, Frank Harrell, Ulrike Held, Tina Hernandez-Boussard, Alejandro Hernández-Arango, Pauline Heus, Bethany Hillier, Michael Hoffman, Jeroen Hoogland, Mohammed Hudda, Merel Huisman, Ivana Isgum, Jan Jaap Baalbergen, Patricia Jaspers, David Jenkins, Kevin Jenniskens, Charles Kahn, Vineet Kamal, Michael Kammer, Evangelos Kanoulas, Ilse Kant, Teus Kappen, Christopher Kelly, Nina Kreuzberger, Jethro Kwong, Joanna Lane, Linda Lapp, Artuur Leeuwenberg, Tim Leiner, Brooke Levis, Qui Li, Christopher Lovejoy, Kim Luijken, Pat Lyons, Stephen M Borstelmann, Jie Ma, Dennis Makau, Sue Mallett, Konstantinos Margetis, Iain Marshall, Glen Martin, Bilal Mateen, Michael Matheny, Matthew McDermott, David McLernon, Jamie Miles, Antonio Moura, Leila Mureebe, Myura Nagendran, Charlie Nederpelt, Daan Nieboer, Wiro Niessen, Steven Nijman, Quentin Noirhomme, Daniel Oberski, Johan Ordish, Almilaji Orouba, Ravi Parikh, Seong Park, Andre Pascal Kengne, Niels Peek, Bas Penning de Vries, Daniel Pinto dos Santos, Robert Platt, Frank Rademakers, Erik Ranschaert, Kelly Reeve, Samuel Relton, Dimitris Rizopoulos, Sherri Rose, Laura Rosella, Jan Roth, Alicja Rudnicka, Rupa Sakar, Pui San Tan, Katie Scandrett, Michael Schlussel, Ewoud Schuit, Martijn Schut, Mark Sendak, Jamie Sergeant, Chunhu Shi, Nicole Skoetz, Kym Snell, Adrian Soto-Mota, Viknesh Sounderajah, Matthew Sperrin, Benjamin Spivak,

Ewout Steyerberg, Tom Stocker, Matthew Strother, Herdiantri Sufriyana, Xin Sun, Tom Syer, Toshihiko Takada, Halil Tanboga, Cristian Tebé, Paul Tiffin, Jim Tol, Eric Topol, Darren Treanor, Ioanna Tzoulaki, Wouter Veldhuis, Kamal Vineet, Christine Wallish, Junfeng Wang, Peter Watkinson, Wim Weber, Gary Weissman, Penny Whiting, Rebecca Whittle, Jack Wilkinson, Tyler Williamson, Robert Wolff, Aaron Y Lee, Christopher Yau, Valentijn de Jong, Annemarie van 't Veen, Wouter van Amsterdam, Bas van Bussel, Peter van der Heijden, Iwan van der Horst, Florian van Royen, Kim van der Braak, Marie Westwood.

Authors' contributions

KGMM, JBR, RDR, GSC and MvS conceived this paper. KGMM, JAAD, TK, CAN, PD, LH, JBR, RDR, GSC and MvS designed the PROBAST+AI tool (signalling questions for the quality, evaluation and applicability components) and subsequently designed the surveys carried out to inform the guideline content. TK, CAN, and JAAD analysed the survey results and free-text comments from the survey. TK designed the materials for the consensus meeting with input from KGMM. TK took consolidated notes from the consensus meeting. KGMM and MvS led drafting of the manuscript with initial edits from JAAD, TK, CAN, PD, LH, JBR, RDR, and GSC. All authors were involved in revising the article critically for important intellectual content. All authors approved the final version of the article. KGMM is the guarantor of this work. The corresponding author, KGMM, attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding

RDR, GSC and PD are supported by an EPSRC grant for 'Artificial intelligence innovation to accelerate health research' (number: EP/Y018516/1). RDR and GSC are supported by an NIHR-MRC Better Methods Better Research grant (MR/V038168/1); and RDR is supported by the National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. RDR and GSC are National Institute for Health and Care Research (NIHR) Senior Investigators. GSC and PL are supported by Cancer Research UK (programme grant: C49297/A27294). PD is supported by Cancer Research UK (project grant: PRCPJT-Nov21\100021). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Competing interests

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

KGMM is director of Health Innovation Netherlands (HI-NL), editor in chief of BMC Diagnostic & Prognostic Research, and principal investigator and author of the Guidance for high quality AI in healthcare (<https://guideline-ai-healthcare.com>).

GSC is the Director of the UK EQUATOR Centre, editor in chief of BMC Diagnostic & Prognostic Research, and a statistics editor for the BMJ. PL is a meta-researcher with the UK EQUATOR Centre.

RR is a statistics editor for the BMJ and receives royalties for two textbooks: 'Prognosis Research in Healthcare' and 'Individual Participant Data Meta-Analysis'.

Ethical approval

This project qualified as non-medical research involving human subjects (non-WMO) according to the Dutch Central Committee on Research Involving Human Subjects (CCMO) and formal ethical approval was waived. The modified Delphi survey procedure was approved by the data management board of the Julius Centre of the University Medical Centre Utrecht (Netherlands). The modified Delphi process was subject to regular monitoring. Delphi survey participants provided electronic informed consent before completing the survey. All participant data were pseudonymised and stored securely on a server of the University Medical Centre Utrecht.

Registration and protocol

The protocol of this study was published in BMJ Open (*BMJ Open* 2021;11:e048008) and is available on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/4BJWT>).

Data availability

Aggregated Delphi survey responses are available on the Open Science Framework "PROBAST+AI" repository (<https://osf.io/w3cfe/>).

Supplementary data

Supplementary data to this article can be found online.



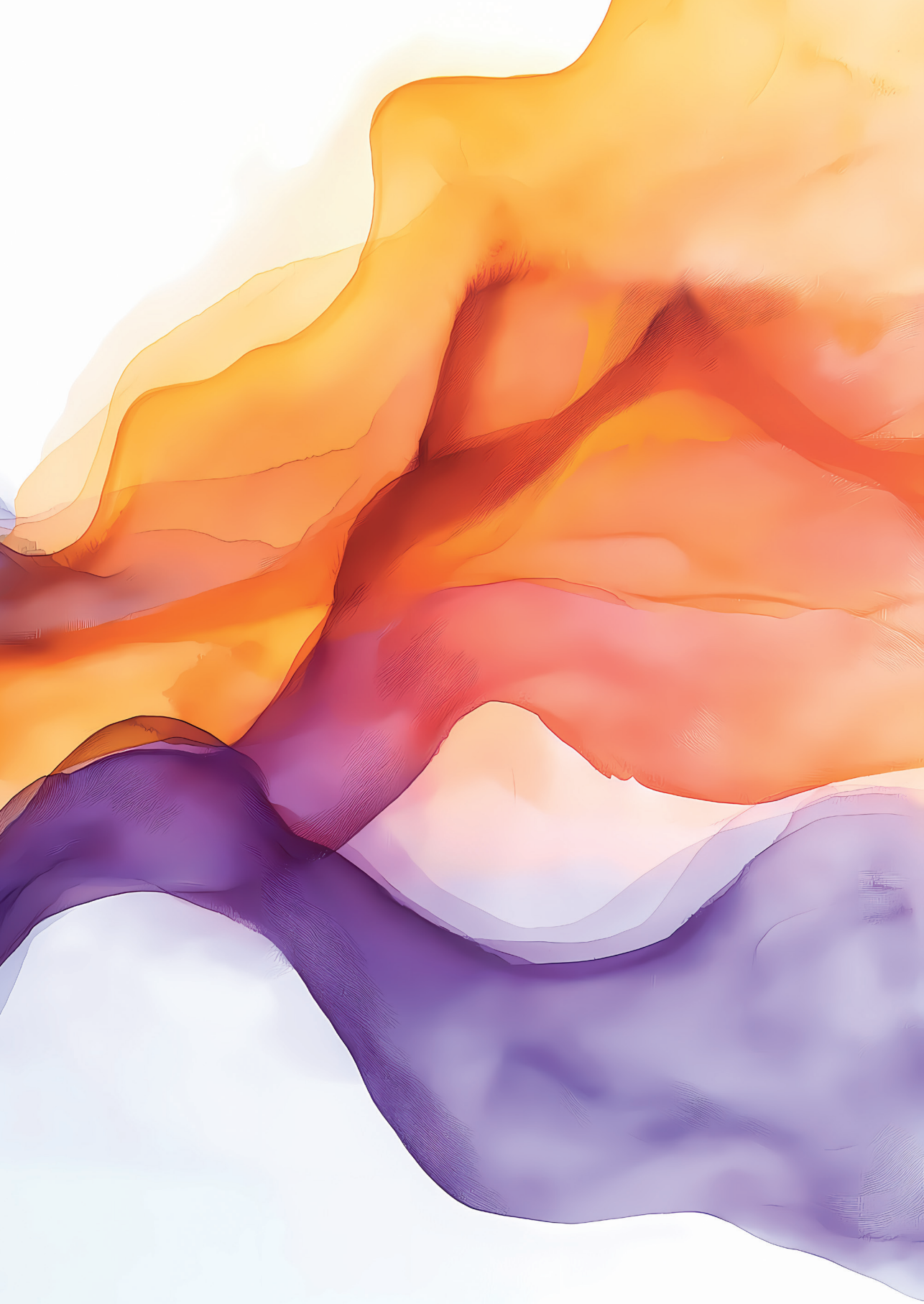
REFERENCES

- (1) Hippisley-Cox J, Coupland CAC, Bafadhel M, Russell REK, Sheikh A, Brindle P, et al. Development and validation of a new algorithm for improved cardiovascular risk prediction. *Nature Medicine*. 2024;30:1440-7.
- (2) Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416.
- (3) 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.
- (4) de Munter L, Polinder S, Lansink KW, Cnossen MC, Steyerberg EW, de Jongh MA. Mortality prediction models in the general trauma population: A systematic review. *Injury*. 2017;48:221-9.
- (5) 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.
- (6) Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J Clin Epidemiol*. 2023;154:8-22.
- (7) Bouwmeester W, Zuthoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLOS Medicine*. 2012;9:e1001221.
- (8) White N, Parsons R, Collins G, Barnett A. Evidence of questionable research practices in clinical prediction models. *BMC Medicine*. 2023;21:339.
- (9) Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*. 2021;27:2176-82.
- (10) Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*. 2024.
- (11) Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *Journal of Clinical Epidemiology*. 2023;157:120-33.
- (12) Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Systematic review finds “spin” practices and poor reporting standards in studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*. 2023;158:99-110.
- (13) London AJ, Kimmelman J. Against pandemic research exceptionalism. *Science*. 2020;368:476-7.
- (14) van Royen FS, Moons KGM, Geersing G-J, van Smeden M. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *European Respiratory Journal*. 2022;60:2200250.
- (15) Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. *J Clin Epidemiol*. 2021;138:219-26.
- (16) Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. *J Clin Epidemiol*. 2007;60:863-5; author reply 5-6.
- (17) Moons KG, Hooft L, Williams K, Hayden JA, Damen JA, Riley RD. Implementing systematic reviews of prognosis studies in Cochrane. *Cochrane Database Syst Rev*. 2018;10:Ed000129.

- (18) Damen JAA, Moons KGM, van Smeden M, Hooft L. How to conduct a systematic review and meta-analysis of prognostic model studies. *Clin Microbiol Infect.* 2023;29:434-40.
- (19) 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-w33.
- (20) 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.
- (21) Kaul T, Kellerhuis BE, Damen JAA, Schuit E, Jenniskens K, van Smeden M, et al. Methodological quality assessment tools for diagnosis and prognosis research: overview and guidance. *J Clin Epidemiol.* 2024;177:111609.
- (22) Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *Bmj.* 2017;356:i6460.
- (23) Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. *Diagnostic and Prognostic Research.* 2019;3:13.
- (24) Collins GS, Dhiman P, Ma J, Schluskel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ.* 2024;384:e074819.
- (25) Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ.* 2024;384:e074820.
- (26) Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Jr., Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med.* 2019;38:1262-75.
- (27) Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med.* 2019;38:1276-96.
- (28) Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *Bmj.* 2020;368:m441.
- (29) van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology.* 2016;16:163.
- (30) van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res.* 2019;28:2455-74.
- (31) 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.
- (32) van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology.* 2014;14:137.
- (33) Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine.* 2019;17:230.
- (34) Gichoya J, McCoy L, Celi L, Ghassemi M. Equity in essence: A call for operationalising fairness in machine learning for healthcare. *BMJ Health & Care Informatics.* 2021;28:e100289.
- (35) Ganapathi S, Palmer J, Alderman JE, Calvert M, Espinoza C, Gath J, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med.* 2022;28:2232-3.
- (36) McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med.* 2021;13.
- (37) Artificial Intelligence Act (Regulation (EU) 2024/1689), Official Journal version of 13 June 2024. Interinstitutional File: 2021/0106(COD). <https://eur-lex.europa.eu/eli/reg/2024/1689>

- (38) Finlayson SG, Beam AL, van Smeden M. Machine Learning and Statistics in Clinical Research Articles-Moving Past the False Dichotomy. *JAMA Pediatr.* 2023;177:448-50.
- (39) Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open.* 2021;11:e048008.
- (40) Andaur Navarro CLD, Johanna AA ; Hooft, Lotty; Reitsma, Johannes B; van Smeden, Maarten; Riley, Richard D; Collins, Gary S; Moons, Karel GM; . Protocol for development of a risk of bias assessment tool for diagnostic and prognostic prediction studies based on artificial intelligence: PROBAST-AI 2022; cited 2024 August].
- (41) Langenhuijsen LFS, Janse RJ, Venema E, Kent DM, van Diepen M, Dekker FW, et al. Systematic metareview of prediction studies demonstrates stable trends in bias and low PROBAST inter-rater agreement. *J Clin Epidemiol.* 2023;159:159-73.
- (42) Kaiser I, Pfahllberg AB, Mathes S, Uter W, Diehl K, Steeb T, 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.
- (43) Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol.* 2021;138:60-72.
- (44) Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Medical Research Methodology.* 2022;22:12.
- (45) Rech MM, de Macedo Filho L, White AJ, Perez-Vega C, Samson SL, Chaichana KL, et al. Machine Learning Models to Forecast Outcomes of Pituitary Surgery: A Systematic Review in Quality of Reporting and Current Evidence. *Brain Sci.* 2023;13.
- (46) Munguía-Realpozo P, Etchegaray-Morales I, Mendoza-Pinto C, Méndez-Martínez S, Osorio-Peña Á D, Ayón-Aguilar J, et al. Current state and completeness of reporting clinical prediction models using machine learning in systemic lupus erythematosus: A systematic review. *Autoimmun Rev.* 2023;22:103294.
- (47) Kee OT, Harun H, Mustafa N, Abdul Murad NA, Chin SF, Jaafar R, et al. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. *Cardiovascular Diabetology.* 2023;22:13.
- (48) Song Z, Yang Z, Hou M, Shi X. Machine learning in predicting cardiac surgery-associated acute kidney injury: A systemic review and meta-analysis. *Front Cardiovasc Med.* 2022;9:951881.
- (49) Yang Q, Fan X, Cao X, Hao W, Lu J, Wei J, et al. Reporting and risk of bias of prediction models based on machine learning methods in preterm birth: A systematic review. *Acta Obstet Gynecol Scand.* 2023;102:7-14.
- (50) 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.
- (51) Lans A, Kanbier LN, Bernstein DN, Groot OQ, Ogink PT, Tobert DG, et al. Social determinants of health in prognostic machine learning models for orthopaedic outcomes: A systematic review. *J Eval Clin Pract.* 2023;29:292-9.
- (52) Li B, Feridooni T, Cuen-Ojeda C, Kishibe T, de Mestral C, Mamdani M, et al. Machine learning in vascular surgery: a systematic review and critical appraisal. *npj Digital Medicine.* 2022;5:7.
- (53) Groot OQ, Bindels BJJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop.* 2021;92:385-93.

- (54) Yusuf M, Atal I, Li J, Smith P, Ravaud P, Fergie M, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open*. 2020;10:e034568.
- (55) Araújo ALD, Moraes MC, Pérez-de-Oliveira ME, Silva VMD, Saldivia-Siracusa C, Pedroso CM, et al. Machine learning for the prediction of toxicities from head and neck cancer treatment: A systematic review with meta-analysis. *Oral Oncol*. 2023;140:106386.
- (56) Sheehy J, Rutledge H, Acharya UR, Loh HW, Gururajan R, Tao X, et al. Gynecological cancer prognosis using machine learning techniques: A systematic review of the last three decades (1990-2022). *Artif Intell Med*. 2023;139:102536.
- (57) Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLoS One*. 2020;15:e0234722.
- (58) Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagnostic and Prognostic Research*. 2020;4:16.
- (59) Andaur Navarro CL, Kaul T. Data Management Plan: PROBAST+AI Delphi Survey 2022 [cited 25 August 2024]. Available from: https://dmponline.dcc.ac.uk/plans/78317/export.pdf?export%5Bquestion_headings%5D=true.
- (60) Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
- (61) Moons KG, 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:e1001744.
- (62) Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health*. 2021;3:e260-e5.





Adherence to the TRIPOD+AI Guideline: An Updated Reporting Assessment Tool

Tabea Kaul*, Emilie de Kanter*, Pauline Heus, Johanna AA Damen R Harmen Kuijten,
Tom M de Groot Johannes B Reitsma Gary S Collins Lotty Hooft, Karel GM Moons

* contributed equally and shared first author

Manuscript in preparation

ABSTRACT

Background: Incomplete reporting of research limits its usefulness and contributes to research waste. Numerous reporting guidelines have been developed to support complete reporting of healthcare research. Reporting guidelines also facilitate editors, reviewers, readers, and guideline developers to determine whether the study design, methods, findings, and conclusions are completely and transparently reported. Completeness of reporting can be measured by evaluating the adherence to reporting guidelines. However, assessing adherence to a reporting guideline often lacks uniformity. In 2019, we developed a reporting adherence tool for the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement. With recent advances in regression and artificial intelligence (AI)/machine learning (ML) based methods, TRIPOD+AI (www.tripod-statment.org) was developed to replace the TRIPOD statement. The aim of this study was to develop an updated adherence tool for TRIPOD+AI.

Methods: Based on the TRIPOD+AI full reporting guideline including the accompanying Explanation and Elaboration light and TRIPOD+AI for Abstracts, we updated and expanded the original TRIPOD adherence tool and refined the adherence elements and their scoring rules through discussions within the author team.

Results: The updated tool comprises the necessary adherence elements and includes several automated scoring rules. We developed separate TRIPOD+AI adherence tools for model development, model evaluation, and for studies describing both in one paper.

Conclusion: A uniform approach to assessing reporting adherence of TRIPOD+AI allows for comparisons across various fields, reporting to be monitored over time, and incentivizes primary study authors to comply.

BACKGROUND

Incomplete reporting of research limits its usefulness and contributes to research waste (1-3). Reporting guidelines facilitate better research reporting by supporting researchers in the manuscript writing stage with a list of items that are essential for complete reporting of their research. They also facilitate editors, reviewers, readers, and guideline developers to determine whether the study design, methods, findings, and conclusions are completely and transparently reported. The Enhancing the QUALity and Transparency Of health Research (EQUATOR) Network is an international initiative that includes a database of reporting guidelines in medical research (4). Widely used examples include the Consolidated Standards of Reporting Trials (CONSORT) Statement (5), Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) (6), Standards for Reporting Diagnostic Accuracy (STARD) statement (7), and the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (8, 9). Following recent advances in artificial intelligence (AI) and machine learning (ML), various reporting guidelines have been revised or updated to include reporting items relevant to AI components (e.g., interventions with AI components, prediction models based on AI/ML methods) (10-13).

In 2024, TRIPOD+AI (www.tripod-statement.org) updated and replaced the TRIPOD statement (8, 9, 12). The revised statement extends TRIPOD with items relevant to both regression- and AI/ML based prediction model studies, such as patient and public involvement, open science practices (e.g., details on study registration), and a particular emphasis on fairness (i.e., the property of prediction models not to discriminate against individuals or groups based on sensitive attributes, such as age, race/ethnicity) (12). TRIPOD+AI comprises 52 items (including main and subitems) considered essential for complete reporting of model development studies (i.e., development of a new prediction model) and model evaluation studies (i.e., evaluating the predictive performance of an existing model in new data) (Box 1a&b) (12). The number of applicable TRIPOD+AI items varies per study type (i.e., model development and model evaluation).

Previous reports have identified poor adherence to reporting guidelines (14-19). For instance, only half of all essential TRIPOD items were adhered to across various clinical fields (14). To incentivize complete reporting, to allow for adherence comparisons between different fields, and to guide reviewers, editors, readers and guideline developers in determining whether a primary prediction model study is completely and transparently reported, adherence assessments play a vital role (20). However, adherence assessments have been shown to lack uniformity (21). To enable

a streamlined approach to TRIPOD adherence assessments, Heus et. al developed an adherence assessment tool including pre-defined scoring rules (20). The tool provides a quantitative adherence score and allows for comparison across any medical field and enables monitoring of completeness of reporting over time.

The TRIPOD adherence tool was published in 2019. In light of the recent transition from TRIPOD to TRIPOD+AI, an update of the adherence tool also became necessary. In this paper, we provide an update of the TRIPOD adherence tool to the TRIPOD+AI statement, to allow for continued systematic adherence measurement across various fields where TRIPOD+AI can be applied.

METHODS

Figure 1 shows the flow of the TRIPOD+AI adherence tool updating process. First, we compared the TRIPOD statement and the adherence tool to the new TRIPOD+AI statement (including the Explanation and Elaboration light and TRIPOD+AI for Abstracts) (Box 1a&b) (8, 9, 12, 20). RHK and TDG prepared a first draft of the TRIPOD+AI adherence tool, which was updated by EdK, TK, JAAD, and PH. This initial draft was then discussed among the broader author team regarding content, wording, and scoring rules in several hybrid roundtable discussions. The author team includes developers of the original TRIPOD adherence tool (PH, JAAD, JBR, GSC, KGMM, LH) and of the TRIPOD+AI statement (GSC, JBR, KGMM).

RESULTS

The TRIPOD adherence tool comprised 32 items (main and subitems) and 97 adherence elements (20). The updated TRIPOD+AI adherence tool consists of 52 items (main and subitems, labelled from item 1 to item 27c) and 131 adherence elements (Supplementary file 1). The TRIPOD+AI adherence elements cover individual reporting topics in a more granular manner to enable adherence scoring (e.g., TRIPOD+AI item: “4b. Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up”, translates into the adherence elements: “4bi: The starting date of accrual is reported” and “4bii: The end date of accrual is reported” [Supplementary file 1]). In accordance with the TRIPOD+AI statement, the number of items varies depending on the type of prediction model study (i.e., development, external evaluation, or the combination of both development and external evaluation).

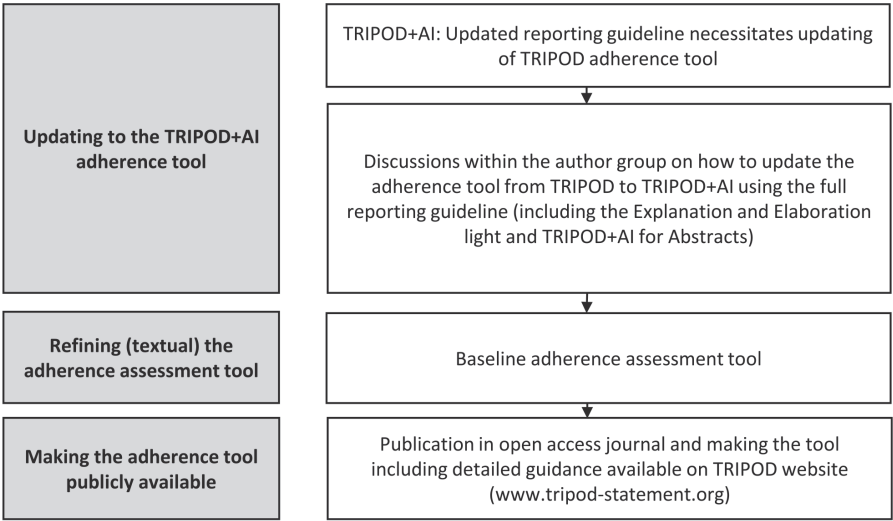


Figure 1 Process of developing the TRIPOD+AI adherence tool

Box 1a. The TRIPOD+AI statement

Title
1. Title (D;E) Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted
Abstract
2. Abstract (D;E): See Box 1b
Introduction
3. Background a. Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models (D;E) b. Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public) (D;E) c. Describe any known health inequalities between sociodemographic groups(D;E)
4. Objectives Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both) (D;E)

Box 1a. The TRIPOD+AI statement (*continued*).

Methods
5. Data
a. Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data (D;E)
b. Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up (D;E)
6. Participants
a. Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres (D;E)
b. Describe the eligibility criteria for study participants (D;E).
c. Give details of any treatments received, and how they were handled during model development or evaluation, if relevant (D;E)
7. Data preparation
• Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups (D;E)
8. Outcome
a. Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups (D;E)
b. If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors (D;E)
c. Report any actions to blind assessment of the outcome to be predicted (D;E)
9. Predictors
a. Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building (D;E)
b. Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors) (D;E)
c. If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors (D;E)
10. Sample size
• Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation (D;E)
11. Missing data
• Describe how missing data were handled. Provide reasons for omitting any data (D;E)
12. Analytical methods
a. Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements (D)

Box 1a. The TRIPOD+AI statement (*continued*).

b. Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardization) (D)

c. Specify the type of model, rationale, all model building steps, including any hyperparameter tuning, and method for internal validation (D)

d. Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations (D;E)

e. Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple model (D;E)

f. Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings (E)

g. For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application, programming interface) (E)

13. Class imbalance

- If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions (D;E)

14. Fairness

- Describe any approaches that were used to address model fairness and their rationale (D;E)

15. Model output

- Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified (D)

16. Training versus evaluation

- Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors (D;E)

17. Ethical approval

- Name the institutional research board or ethics committee that approved the study and describe the participant informed consent or the ethics committee waiver of informed consent, (D;E)

Open science**18. Funding**

- a. Give the source of funding and the role of the funders for the present study (D;E)

18. Conflicts of interest

- b. Declare any conflicts of interest and financial disclosures for all authors (D;E)

18. Protocol

- c. Indicate where the study protocol can be accessed or state that a protocol was not prepared (D;E)

18. Registration

- d. Provide registration information for the study, including register name and registration number, or state that the study was not registered (D;E)

18. Data sharing

Box 1a. The TRIPOD+AI statement (*continued*).

e. Provide details of the availability of the study data (D;E)
18. Code sharing
e- Provide details of the availability of the analytical code (D;E)
Patient and public involvement
19. Patient and public involvement
• Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement (D;E)
Result
20. Participants
a. Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful (D;E) b. Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups (D;E) c. For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome) (E)
21. Model development
• Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation) (D;E)
22. Model specification
• Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third party evaluation and implementation, including any restrictions to access or reuse (e.g., freely available, proprietary) (D)
23. Model performance
a. Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation (D;E) b. If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD-Cluster for additional details (D;E)
24. Model updating
• Report the results from any model updating, including the updated model and subsequent performance (E)
Discussion
25. Interpretation
• Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies (D;E)
26. Limitations

Box 1a. The TRIPOD+AI statement (*continued*).

- Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability (D;E)

27. Usability of the model in the context of current care

a. Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model (D)

b. Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users (D)

c. Discuss any next steps for future research, with a specific view to applicability and generalisability of the model (D;E)

D; E: item relevant to both development and external evaluation; D: item only relevant to development; E: item only relevant to external evaluation

Box 1b. The TRIPOD+AI Statement for Abstracts**Abstract** (D;E)

Background

- Provide a brief explanation of the healthcare context and rationale for developing or evaluating the performance of all models

Objectives

- Specify the study objectives, including whether the study describes model development, evaluation, or both

Methods

a. Describe the sources of data

b. Describe the eligibility criteria and setting where the data were collected

c. Specify the outcome to be predicted by the model, including time horizon of predictions in case of prognostic models

d. Specify the type of model, a summary of the model-building steps, and the method for internal validation†

e. Specify the measures used to assess model performance (e.g., discrimination, calibration, clinical utility)

Results

a. Report the number of participants and outcome events

b. Summarise the predictors in the final model†

c. Report model performance estimates (with confidence intervals)

Discussion

- Give an overall interpretation of the main results

Registration

- Give the registration number and name of the registry or repository

D; E: item relevant to both development and external evaluation; D: item only relevant to development; E: item only relevant to external evaluation

Using the TRIPOD+AI adherence assessment tool

Besides wording and content changes related to the replacement of TRIPOD by TRIPOD+AI, the updated adherence tool does not contain significant differences for scoring (20). *Supplementary file 1* provides an overview of various TRIPOD+AI scoring scenarios, while *Supplementary file 2* provides a fillable adherence tool in Excel format that includes score calculations.

Guidance for use

Similar to review practices, we recommend TRIPOD+AI adherence assessments to be performed by at least two users of the same studies independently, including consensus meetings, before the final consolidated adherence score (see below) is calculated. Further, users should decide before starting the assessment which elements are applicable to their assessment context and operationalize what they regard as adhered to for their specific situation. The guidance documents of TRIPOD+AI (including the Explanation & Elaboration light) can help users make an informed choice or clarify aspects where necessary (12).

Scoring adherence per TRIPOD+AI item

First, users may judge adherence by assessing a prediction model study against each and all relevant elements in the adherence tool (n=131). Adherence to each element is answered as “yes” (Y), indicating that the element was adhered to, “no” (N), indicating lack of adherence, as well as “not applicable” (NA), indicating that the element does not apply in the specific situations (e.g., item 5a: follow-up is not applicable to diagnostic studies). If one paper includes both model development and evaluation, the items should be scored separately, although some elements need to be adhered to only once (denoted by ‘OR’ in the scoring rules; e.g., adherence element 3ai: “The healthcare context [including whether diagnostic or prognostic] is explained.”). In the previous TRIPOD adherence form, some elements were also considered adhered to if they were (only) described in a cited reference rather than in the prediction model paper itself. These were indicated with the response “referenced”. This has changed in the TRIPOD+AI adherence form. We now advise users to assess the adherence of the reporting in the cited reference and score “yes” (instead of “referenced”) if the information is indeed provided in the reference. If the manuscript states that specific information is described in a cited reference, however, the information is not available, the element should be scored as “no”.

Once adherence to each element is assessed, adherence to each TRIPOD+AI item can be determined. An item is considered to be adhered to if the response to all adherence elements is either “yes” or “not applicable”. If a manuscript describes both model development and evaluation, these need to be combined first before adherence to the item can be determined. In some situations, information needs to be described only once (either for model development or for evaluation) to adhere to the TRIPOD+AI item (e.g., Title item 1ii). However, in most situations, information needs to be described for both development and evaluation in order to adhere to the TRIPOD+AI item. The various scoring scenarios are outlined in the adherence form (Supplementary file 1).

Overall adherence to TRIPOD+AI

Overall adherence should be calculated by dividing the sum of adhered TRIPOD+AI items by the total number of applicable TRIPOD+AI items, thus the total (i.e., denominator) can vary. The total number of applicable TRIPOD+AI items for model development is 51, for external evaluation 46, and for development and external evaluation is 52 (Supplementary files 1&2). A total of 32 adherence elements may be scored as “not applicable” in the model development context, while 29 adherence elements can be scored as “not applicable” in the model evaluation context (Supplementary files 1&2). When assessing the adherence of multiple prediction model studies simultaneously, an adherence score can be calculated by dividing the total number of studies that adhered to a specific TRIPOD+AI item by the number of studies to which the specific TRIPOD+AI item was applicable (Supplementary file 2).

CONCLUDING REMARKS

Adhering to reporting standards is a key step in ensuring that research findings can be subsequently used. Our TRIPOD+AI adherence tool (www.tripod-statement.org) streamlines assessment efforts and thus allows for comparison of adherence practices across prediction model studies including either regression or AI (including machine learning techniques) in any biomedical field, domain or setting. The TRIPOD+AI adherence tool is not only helpful for reviewers, editors, readers and guideline developers, it may also encourage primary study authors to report prediction model studies adequately. The assessment tool also facilitates the monitoring of reporting over time to identify trends (i.e., improvements) in reporting, and identifies specific areas of poor reporting that could benefit from targeted education on the importance of and how to better report. With recent developments in AI and its integration in study review

or evaluation processes, promoting adequate reporting becomes even more important (22). Although methods have previously been described how to synthesize evidence from incomplete reports for, e.g., systematic reviews of therapeutic or preventive intervention research (23), it is still lacking for systematic reviews of prediction model research. Therefore, transparent and complete reporting of prediction model studies remains key to reducing research waste and enabling subsequent use of research findings.

Acknowledgements

None.

Authors' contributions

I contributed to the design of the adherence tool and finalised it together with EdK after several rounds of discussions with all co-authors. Subsequently, EdK and I wrote the first draft of the paper and circulated it among all co-authors. After receiving the initial round of feedback, I integrated the co-authors' suggestions into the manuscript, which is now included in this thesis chapter. As a result, EdK and I are recognized as shared first authors.

Funding

None.

Competing interests

PH, JAAD, JBR, GSC, KGMM, and LH are the developers of the original TRIPOD adherence tool; KGMM, GSC, and JBR are part of the TRIPOD+AI group. All other authors have nothing to declare.

Ethical approval

Not required for this work.

Registration and protocol

This study was not registered and does not include a publishable protocol.

Data availability

The full TRIPOD+AI statement is published in the BMJ (BMJ 2024;385:e078378). The TRIPOD+AI adherence tool can be found in the supplementary data.

Supplementary data

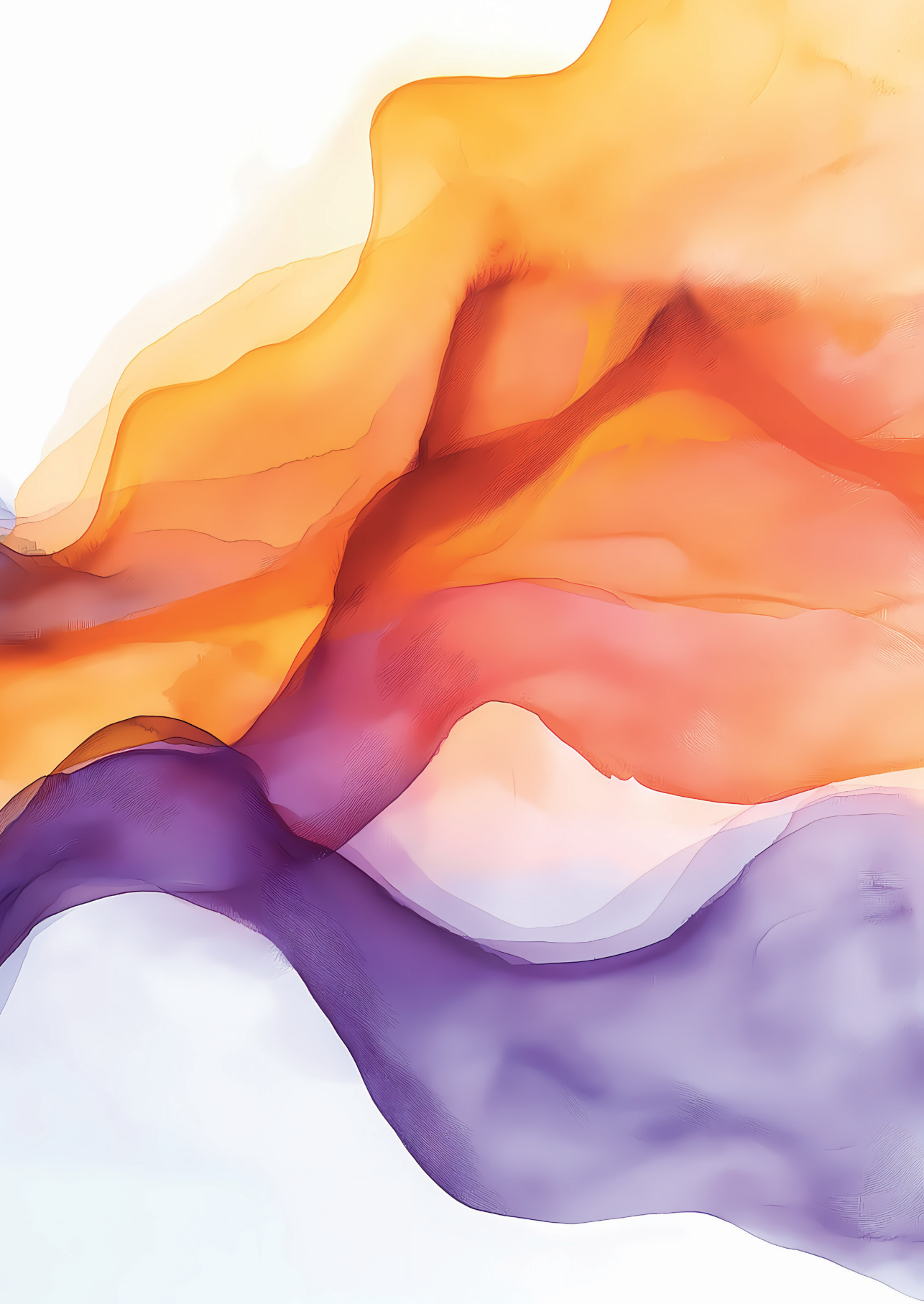
Supplementary data to this article can be found online.



REFERENCES

- (1) Liberati A. An unfinished trip through uncertainties. *BMJ*. 2004;328(7438):531.
- (2) Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *The Lancet*. 2009;374(9683):86-9.
- (3) Okumura Y. Reducing Research Waste Through Good Reporting Practices. *Journal of Epidemiology*. 2016;26(8):397-8.
- (4) EQUATOR network. Enhancing the quality and transparency of health research 2025 [cited 2025 20 February]. Available from: <https://www.equator-network.org/about-us/>.
- (5) Schulz KF, Altman DG, Moher D, the CG. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*. 2010;8(1):18.
- (6) 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. *BMJ*. 2021;372:n71.
- (7) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Bmj*. 2015;351:h5527.
- (8) Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*. 2015;13(1):1.
- (9) 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(1):W1-73.
- (10) Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11(6):e047709.
- (11) Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020;370:m3164.
- (12) Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
- (13) Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
- (14) Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Medicine*. 2018;16(1):120.
- (15) Feller D, Wingbermuhle R, Berg B, Vigdal ØN, Innocenti T, Grotle M, et al. Improvements Are Needed in the Adherence to the TRIPOD Statement for Clinical Prediction Models for Patients With Spinal Pain or Osteoarthritis: A Metaresearch Study. *The Journal of Pain*. 2024;25(11):104624.
- (16) Liu LQ, Morris PJ, Pengel LH. Compliance to the CONSORT statement of randomized controlled trials in solid organ transplantation: a 3-year overview. *Transpl Int*. 2013;26(3):300-6.
- (17) Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev*. 2012;11(11):Mr000030.
- (18) Hogan KO, Fraga GR. Compliance With Standards for STARD 2015 Reporting Recommendations in Pathology. *Am J Clin Pathol*. 2020;154(6):828-36.

- (19) Dal Santo T, Rice DB, Amiri LSN, Tasleem A, Li K, Boruff JT, et al. Research waste from poor reporting of core methods and results and redundancy in studies of reporting guideline adherence: a meta-research review. medRxiv. 2022:2022.12.19.22283669.
- (20) Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open. 2019;9(4):e025611.
- (21) Korevaar DA, van Enst WA, Spijker R, Bossuyt PMM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evidence Based Medicine. 2014;19(2):47-54.
- (22) Affengruber L, van der Maten MM, Spiero I, Nussbaumer-Streit B, Mahmić-Kaknjo M, Ellen ME, et al. An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review. BMC Medical Research Methodology. 2024;24(1):210.
- (23) Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020;368:l6890.



9

Conclusions and General Discussion

This thesis aimed to facilitate the development of more robust and reliable evidence synthesis methods. To promote the generation of high-quality evidence and refine evidence synthesis (systematic review) methods, we carefully reviewed methodological guidance and its application.

LESSONS LEARNT

1. Clinical trials registry searches may be sufficient in identifying randomised controlled trials (RCTs) for evidence synthesis, however, non-registered RCTs may be missed (Chapter 2).
2. The reporting of trial results on clinical trials registries has improved over time, but accessibility and usability remains challenging. With targeted efforts, clinical trials registries have the potential to become indispensable resources for transparent and efficient evidence dissemination (Chapter 3).
3. Prediction models should be validated in the specific setting they are intended for before integrating them into the clinical care pathway (Chapter 4).
4. A multitude of methodological quality and risk of bias assessment tools exists for diagnosis and prognosis research. Structural similarities and item overlap can be observed, but there are also clear differences between the tools. Five key considerations can help users select the most appropriate tool by ensuring alignment with their specific research question and aim of their assessment (Chapter 5).
5. The inter-rater reliability of the Prediction model Risk Of Bias ASsessment Tool (PROBAST) for individual signalling questions and risk of bias domains was moderate to substantial, but high for the overall study judgement. Scoring should be operationalised a priori, and well-structured consensus meetings should be held to ensure valid and less subjective assessments (Chapter 6).
6. Evidence synthesis of clinical prediction models encounters new challenges with recent developments in regression-, artificial intelligence (AI)-, and machine learning (ML)- based methods. Therefore, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) can be replaced by the updated version, PROBAST+AI, to aid key stakeholders (e.g., systematic reviewers, guideline developers, physicians, patients) encountering prediction models in the healthcare sector regardless the data analytical technique (regression or AI/ML) with which the models were developed (Chapter 7).

7. Incomplete reporting hampers evidence synthesis. Adherence assessments of reporting guidelines may facilitate adequate reporting, e.g., by creating a reward for primary study authors. A reporting adherence tool for TRIPOD+AI was developed for that aim (Chapter 8).

FUTURE PERSPECTIVES FOR EVIDENCE SYNTHESIS METHODS

Reporting, reporting, reporting

Without access to medical research findings, we cannot use them to improve healthcare. Billions of US dollars are wasted every year due to avoidable research waste, straining humans and economies (1, 2). With estimates going as high as 85% being avoidable (2). Interestingly, local reports suggest that research funding and burden of disease only correlate moderately (3, 4). Whilst these figures may appear disheartening, there is room for optimism: the international community demonstrates a consistent dedication to eradicating sources of research waste and pursues the objective of producing high-quality evidence (2, 5-8). In this thesis, we contribute to this commitment by providing improvements to various aspects of evidence synthesis to increase research transparency and eliminate sources of resource waste, with the ultimate aim of improving healthcare.

In chapters 2, 3, and 8, we highlight the importance of adequate reporting practices. Currently, legislation and agreements on registration of research is focused on randomised (intervention) trials (Chapters 2 and 3) (7-11). Nevertheless, this registration is frequently still regarded as mere formality for manuscript submission to medical journals (Chapter 3), which (understandably) significantly reduces the motivation for primary study authors to comprehensively document trial information on registries. Furthermore, previous research suggests that trial information differences between registry record entries and manuscripts have no impact on acceptance for publication in a medical journal (12). This creates a dead end: the trial is prospectively registered (and the box for submission can be checked), but trial information in the registry record is not or only partially updated.

A reward system can be helpful in incentivizing transparent reporting practices. Such a system could benefit primary study authors and principal investigators in several ways. For instance, publication lists (e.g., those needed when applying for grants) may be extended to include reporting on registries (i.e., whether the study was reported on a registry, and if so, whether results were reported in the record). The time between

study completion and reporting of results on registries is shorter compared to waiting for the publication processes of medical journals (13) (Chapter 3). Therefore, extending publication lists to include reporting on registries could reward the investigators quicker.

Adding reporting adherence scores to medical journal publications can help assess and promote research transparency. For submission, most journals already require authors to follow reporting guidelines (e.g., CONSORT or TRIPOD+AI), and authors occasionally need to provide a document stating on which manuscript page or section specific aspects are described (14-18). This could be further extended to include a score on reporting adherence. Reporting adherence tools, such as the one we developed in Chapter 8 for TRIPOD+AI, could facilitate this and may even be extended to checking reporting adherence on trials registries. However, implementation of such a reporting adherence tool would require extensive validation and flexibility for various study designs (19). AI-based screening tools have shown promising results in supporting review processes (e.g., StatReviewer) and could support reporting adherence checks by adding a publicly available adherence score to publications and even to registry records (20-23).

Registration is a common practice for clinical trials, but not for, e.g., clinical prediction model studies (9, 24-29). While agreements for clinical trial registration exist, clinical prediction model studies are not (yet) included in such agreements (26, 28). Model registration would involve two key stages: first, prior to development and evaluation, to register the planned study; and second, after completion, to update the registry record with the finalised model, to enable further assessments (26).

Previous reports have shown an abundance of clinical prediction models developed for the same population and topic, raising concerns of a duplication of effort and thus research waste (30, 31). Rather than developing yet another model, more evaluation and impact studies of existing models are needed to enable faster integration into the clinical care pathways and daily practices. Similar to intervention research, registries could support the identification of previously developed or evaluated prediction models, to assess the risk of missing any relevant studies due to the lack of a published manuscript (i.e., publication bias) (32-34). In Chapter 4, we developed prognostic models only after an evaluation study by Damen et. al showed insufficient performance of COVID-19 prognostic models in low- and lower middle-income countries and thus necessitated the model developments (35).

Many clinical prediction model studies suffer from poor methodological quality and questionable research practices (e.g., p-hacking) have been reported (36-38). Registries can act as central repositories for essential documentation of clinical prediction model

studies (e.g., on data source, pre-processing, fairness aspects). Such information may be essential for the interpretation of results and assessment of methodological quality but less suitable for inclusion in journal publications. Standard methodological and reporting formats, guided by items included in PROBAST+AI (www.probast.org; Chapter 7) and TRIPOD+AI (www.tripod-statement.org) (17) can be used to build these repositories. This will increase their value for further assessment (external evaluation and updating), reduce duplication of effort, avoid research waste, and help identify issues with study validity (2, 26, 29).

Clinical prediction model registration will greatly improve research transparency and potentially reduce research waste. The first step towards improving registration is by creating awareness about the possibility to register prediction models. For example, the recently published updated reporting guideline TRIPOD+AI includes one item motivating researchers to report a registration number, creating awareness about the possibility to register such a study (17). A second step could include expanding existing agreements and policies (e.g., ICMJE requirement (28)) – currently focused on clinical trials – to also cover clinical prediction model studies. The proposed actions and identified implementer groups to enhance research transparency of clinical trials, as described by Chapter 3 of this thesis, can further inspire such actions for clinical prediction model studies, and could even be extended to other types of medical research designs.

Primary study quality and bias: could one size fit all?

We evaluated different aspects linked to primary study quality and biases in this thesis. In Chapter 5, we found that there are many tools for assessing methodological quality (including bias assessments) in diagnosis and prognosis research and many of the included items overlapped. This raises the question whether tool development processes could be streamlined or integrated for greater efficiency and how to select the most appropriate tool for different types of diagnosis or prognosis studies. Nonetheless, quality assessment of rare or complex study designs may still be limited by tool inflexibility and adequate tailoring can be difficult. Also, more general assessment tool challenges have been described, such as in their design (e.g., ceiling effects in scoring domains), usability (e.g., tool complexity, time invest), or assessment guidance (e.g., lack of clarity) (39).

Considering all of this, a solution could involve using a flexible tool that incorporates items of selected methodological quality assessment tools. In Chapter 5, we grouped and compared items of several tools in a matrix and developed guiding questions to help users specify their assessment needs and select an appropriate tool for their

specific question. These serve as a foundational starting point to address design issues and maximize tool effectiveness and use. The selected tools can then even be tailored – if needed – using specific items from other tools. In the future, an ‘overarching’ tool, informed by experts, could further support (less experienced) reviewers in tailoring tools based on their context and research aim. The implementation of such a ‘combined quality assessment tool’ would necessitate a substantial collaborative effort and would require extensive piloting and quality checks to ensure its reliability.

Applying methodological quality assessment tools poses additional challenges. In Chapter 6, we engage in a discussion of the practical application and use of the risk of bias tool PROBAST. We found that PROBAST ratings between two raters changed after between-rater-discussion, even occasionally in case of initial agreement between them. This raises questions on the operationalization of quality assessment tools and highlights the need for adequate assessment instructions to ensure less variability and subjectivity in quality assessments using PROBAST. While our analysis relied on human reviewers, recent advancements in AI show promising results in their application for evidence synthesis (40). AI algorithms require clear operational instructions for assessment and may, hopefully, be less vulnerable to variability and subjectivity. Questions on how AI algorithms can be safely integrated into the current evidence synthesis pipeline and to what extent the human component can be automated (i.e., human in the loop) will guide evidence synthesis methods research in the coming years (41).

AI algorithms have already transformed various sectors and will certainly continue to do so in the future. Although we are only at the beginning, challenges of AI use in healthcare have already presented themselves. These include equity-related issues (e.g., algorithmic bias/fairness) and sustainability concerns (42-46).

With Chapter 7 of this thesis (PROBAST+AI), we offer reviewers and primary study authors a set of signalling questions highlighting aspects related to the prediction model’s (regardless AI or regression based) fairness (i.e., that it does not disadvantage groups of people based on attributes such as age, sex, race/ethnicity, or socioeconomic status) and help assess algorithmic bias (i.e., results of the model [algorithm] systematically disadvantage a particular group of people, based on attributes such as age, sex, race/ethnicity, or socioeconomic status), which may affect the predictive performance for specific sociodemographic subgroups (42, 44, 47). Unfortunately, current solutions to tackle algorithmic bias, e.g., by limiting the effect of decisions on disparities or the effects of sensitive attributes, have been shown to even worsen the disparities (48-50). The European Union’s AI Act represents a key legal framework

that addresses the risks related to AI use and emphasizes the importance of fair, non-discriminatory use (51). This act serves as a crucial first step, and with a collaborative effort from all stakeholders—such as WHO’s Global Initiative on AI for Health (52)—will it help ensure that health equity remains a research priority in the years to come.

Sustainability is a critical consideration in AI use in healthcare. While systems are running, energy and water consumption (i.e., to cool servers, generate electricity, and supply-chain water for server manufacturing) are massive (45, 46, 53). For example, Li et al. estimated that training the Chat GPT-3 language model consumed approximately 700,000 litres of water on-site (53). The Green AI movement fosters sustainable practices in model design, training, and application (45, 46, 54). A recent systematic review reports that a Green AI approach can save at least 50% of energy (55). In the healthcare setting, recommendations for sustainable AI are various and include - amongst others - designing energy-efficient AI models, incorporating green computing infrastructure (i.e., providing adequate hardware), knowledge sharing between institutions, and regulatory compliance and reporting (i.e., impact assessments) (46).

Current initiatives to enhance research transparency

As the World Health Assembly stated in 2022, we have a commitment to produce high-quality evidence in medical research (8). Currently, many initiatives are working on increasing the value of the medical research data we have and produce. For example, the European Health Data Space is an initiative by the European Union to foster safe and secure data exchange to simplify the use and reuse of health data to benefit patients, researchers, innovators, and regulators (56); the European Open Science cloud, which aims to provide a safe environment and promotes FAIR (Findable, Accessibility, Interoperability, and Reusability) data management (57), facilitates sharing on a broader scale. Unfortunately, global data sharing requirements remain limited and are challenged by data privacy (58, 59).

CONCLUSION

Every patient and citizen deserves healthcare based on the best available evidence. We examined various evidence synthesis methods, including registration, reporting, and quality assessment practices, used to inform evidence-based medicine and provide key solutions to address current challenges. With collaborative efforts of the international research community, we aim to implement the proposed changes, leading to improved methods for evidence synthesis in medicine. To end: “We need less research, better research, and research done for the right reasons” (Doug Altman) (60).

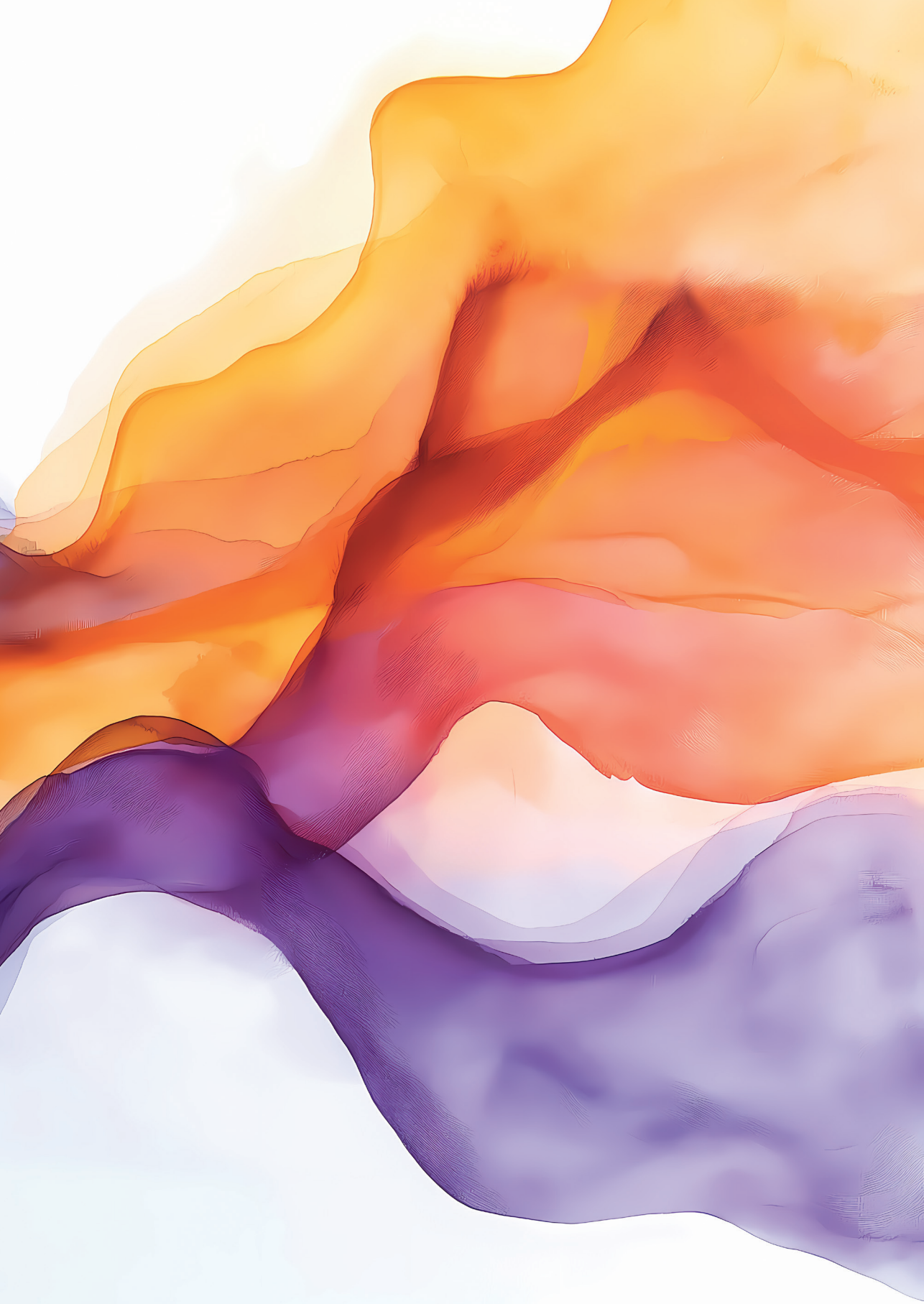
REFERENCES

- (1) Liberati A. An unfinished trip through uncertainties. *BMJ*. 2004;328(7438):531.
- (2) Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *The Lancet*. 2009;374(9683):86-9.
- (3) Gillum LA, Gouveia C, Dorsey ER, Pletcher M, Mathers CD, McCulloch CE, et al. NIH Disease Funding Levels and Burden of Disease. *PLOS ONE*. 2011;6(2):e16837.
- (4) Nozaki I. WHO's budgetary allocation and disease burden. *The Lancet*. 2013;382(9896):937-8.
- (5) Cochrane. About us 2025 [Available from: <https://www.cochrane.org/about-us>].
- (6) WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. Sect. 35.-36. (2013).
- (7) International Clinical Trial Registry Plattform. Joint statement on public disclosure of results from clinical trials, : World Health Organization; 2017 [cited 2025 20 February]. Available from: <https://www.who.int/news/item/18-05-2017-joint-statement-on-registration>.
- (8) Strengthening clinical trials to provide high-quality evidence on health interventions and to improve research quality and coordination, WHA 75.8. Sect. 75.8 (2022).
- (9) Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov — The Final Rule. *New England Journal of Medicine*. 2016;375(20):1998-2004.
- (10) European Parliament and the Council of the European Union. Regulation (EU) No 536/2014 of the European Parliament and of the council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC [Regulation]. *Official Journal of the European Union*; 2014 [updated May 27,2014; cited 2024 April 5]. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0536>.
- (11) U.S. Government. Food and Drug Administration Amendments Act of 2007: US Government Printing Office; 2007 [Available from: <https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf>].
- (12) van Lent M, IntHout J, Out HJ. Differences between information in registries and articles did not influence publication acceptance. *Journal of Clinical Epidemiology*. 2015;68(9):1059-67.
- (13) Showell M, Buckman S, Berber S, Ata Allah N, Patterson B, Cole S, et al. Publication bias in trials registered in the Australian New Zealand Clinical Trials Registry: Is it a problem? A cross-sectional study. *PLOS ONE*. 2023;18(1):e0279926.
- (14) The BMJ. Article types and preparation 2025 [cited 2025 20 January]. Available from: <https://www.bmj.com/about-bmj/resources-authors/article-types>.
- (15) *Journal of Clinical Epidemiology*. Guide for Authors 2025 [cited 2025 20 January]. Available from: <https://www.jclinepi.com/content/authorinfo>.
- (16) International Committee of Medical Journal Editors. Preparing a Manuscript for Submission to a Medical Journal 2025 [cited 2025 17 January]. Available from: <https://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#two>.
- (17) Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
- (18) Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and pharmacotherapeutics*. 2010;1(2):100-7.
- (19) Puljak L. Reporting checklists are used as measurement tools for assessing quality, even though they have not been validated for such use. *Trials*. 2019;20(1):676.
- (20) Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J Med Internet Res*. 2024;26:e48996.

- (21) Ge L, Agrawal R, Singer M, Kannapiran P, De Castro Molina JA, Teow KL, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. *Systematic Reviews*. 2024;13(1):269.
- (22) Shanahan D. A peerless review? Automating methodological and statistical review. *Research in Progress Blog*. 2016.
- (23) Heaven D. AI peer reviewers unleashed to ease publishing grind. *Nature*. 2018;563(7733):609-10.
- (24) Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: is it time? *Cmaj*. 2010;182(15):1638-42.
- (25) Zarin DA, Tse T. Moving toward transparency of clinical trials. *American Association for the Advancement of Science*; 2008. p. 1340-2.
- (26) Hofman JM, Chatzimparmpas A, Sharma A, Watts DJ, Hullman J. Pre-registration for predictive modeling. *arXiv preprint arXiv:231118807*. 2023.
- (27) White N, Parsons R, Borg D, Collins G, Barnett A. Planned but ever published? A retrospective analysis of clinical prediction model studies registered on clinicaltrials.gov since 2000. *Journal of Clinical Epidemiology*. 2024;173:111433.
- (28) Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*. 2004;351(12):1250-1.
- (29) Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KGM, et al. Improving the Transparency of Prognosis Research: The Role of Reporting, Data Sharing, Registration, and Protocols. *PLOS Medicine*. 2014;11(7):e1001671.
- (30) Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416.
- (31) Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JAA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *Journal of Clinical Epidemiology*. 2021;138:60-72.
- (32) Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. *BMJ*. 2017;356:j448.
- (33) Lefebvre C GJ, Briscoe S, Featherstone R, Littlewood A, Marshall C, Metzendorf M-I, Noel-Storr A, Paynter R, Rader T, Thomas J, Wieland LS. Chapter 4: Searching for and selecting studies. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. editor. *Cochrane Handbook for Systematic Reviews of Interventions version 63*. Cochrane: Cochrane; 2022.
- (34) Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. *Journal of Clinical Epidemiology*. 2011;64(12):1277-82.
- (35) Damen JAA, Arshi B, van Smeden M, Bertagnolio S, Diaz JV, Silva R, et al. Validation of prognostic models predicting mortality or ICU admission in patients with COVID-19 in low- and middle-income countries: a global individual participant data meta-analysis. *Diagn Progn Res*. 2024;8(1):17.
- (36) 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.
- (37) White N, Parsons R, Collins G, Barnett A. Evidence of questionable research practices in clinical prediction models. *BMC Medicine*. 2023;21(1):339.
- (38) Collins GS, Whittle R, Bullock GS, Logullo P, Dhiman P, de Beyer JA, et al. Open science practices need substantial improvement in prognostic model studies in oncology using machine learning. *Journal of Clinical Epidemiology*. 2024;165:111199.

- (39) Tomlinson E, Cooper C, Davenport C, Rutjes AWS, Leeflang M, Mallett S, et al. Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies. *Journal of Clinical Epidemiology*. 2024;171:111370.
- (40) Affengruber L, van der Maten MM, Spiero I, Nussbaumer-Streit B, Mahmić-Kaknjo M, Ellen ME, et al. An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review. *BMC Medical Research Methodology*. 2024;24(1):210.
- (41) Jackson JM, Pinto MD. Human Near the Loop: Implications for Artificial Intelligence in Healthcare. *Clinical Nursing Research*. 2024;33(2-3):135-7.
- (42) Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*. 2024;30(10):2838-48.
- (43) Ganapathi S, Palmer J, Alderman JE, Calvert M, Espinoza C, Gath J, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med*. 2022;28(11):2232-3.
- (44) Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform*. 2021;28(1).
- (45) Bolón-Canedo V, Morán-Fernández L, Cancela B, Alonso-Betanzos A. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*. 2024;599:128096.
- (46) Ueda D, Walston SL, Fujita S, Fushimi Y, Tsuboyama T, Kamagata K, et al. Climate change and artificial intelligence in healthcare: Review and recommendations towards a sustainable future. *Diagnostic and Interventional Imaging*. 2024;105(11):453-9.
- (47) Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*. 2021;27(12):2176-82.
- (48) Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*. 2023;24(1):14730-846.
- (49) McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*. 2020;2(5):e221-e3.
- (50) Goodman SN, Goel S, Cullen MR. Machine Learning, Health Disparities, and Causal Reasoning. *Annals of Internal Medicine*. 2018;169(12):883-4.
- (51) Eupean Parliament. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008,(EU) No 167/2013,(EU) No 168/2013,(EU) 2018/858,(EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU,(EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Off. J Eur Union. 2024;50:202.
- (52) World Health Organization. Global Initiative on AI for Health 2023 [cited 2025 03 February]. Available from: <https://www.who.int/initiatives/global-initiative-on-ai-for-health>.
- (53) Li P, Yang J, Islam MA, Ren S. Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. *arXiv preprint arXiv:230403271*. 2023.
- (54) Schwartz R, Dodge J, Smith NA, Etzioni O. Green ai. *Communications of the ACM*. 2020;63(12):54-63.
- (55) Verdecchia R, Sallou J, Cruz L. A systematic review of Green AI. *WIREs Data Mining and Knowledge Discovery*. 2023;13(4):e1507.
- (56) European Union. European Health Data Space, 2024 [cited 2025 23 January]. Available from: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en.
- (57) European Union. European Open Science Cloud (EOSC) 2022 [cited 2025 23 January]. Available from: https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/european-open-science-cloud-eosc_en.
- (58) Tan AC, Webster AC, Libesman S, Yang Z, Chand RR, Liu W, et al. Data sharing policies across health research globally: Cross-sectional meta-research study. *Research Synthesis Methods*. 2024;15(6):1060-71.

- (59) Bentzen HB, Castro R, Fears R, Griffin G, ter Meulen V, Ursin G. Remove obstacles to sharing health data with researchers outside of the European Union. *Nature Medicine*. 2021;27(8):1329-33.
- (60) Altman DG. The scandal of poor medical research. *BMJ*. 1994;308(6924):283-4.



Appendices

English Summary

Nederlandse Samenvatting

Deutsche Zusammenfassung

List of Manuscripts

Author Affiliations

PhD Portfolio

Curriculum Vitae

Acknowledgements

ENGLISH SUMMARY

By providing a synthesis of the available scientific evidence, systematic reviews inform evidence-based medicine. However, they are dependent on the quality and completeness of the underlying body of evidence. Challenges in evidence synthesis include the lack of evidence, biased evidence, and lack of established evidence synthesis guidance. For prediction research these challenges seem even more prominent. This thesis critically examined various aspects of evidence synthesis in evidence-based medicine with the aim to provide more reliable synthesis methods. We specifically focused on aspects related to methodological guidance and its application.

In **Chapter 2**, we evaluated two recommended search approaches for identifying randomised controlled trials (RCTs) for systematic reviews of interventions. The aim of this study was to determine whether searching clinical trials registries compared to medical journal databases yielded a higher number of completed RCTs and to quantify the number of eligible ongoing RCTs (relevant for updating systematic reviews). We searched various clinical trials registries and medical journal databases for RCTs based on the eligibility criteria of a systematic review from the cardiovascular field as a case study. We compared the search results to determine the extent to which the two sources identified the same RCTs. The clinical trials registry approach identified a larger number of completed RCTs. However, both search approaches missed some RCTs. Therefore, if the goal is to identify all eligible RCTs, both sources should continue to be searched. Searching in clinical trials registries alone may be possible once the barriers to searching in clinical trials registries are properly tackled, particularly through ensuring reporting of study results in registries and implementing simplified retrieval mechanisms of results and publications linked to clinical trial records.

Recent initiatives and mandates have established results reporting on clinical trials registries to improve transparency and reduce publication bias. However, local reports suggest that results reporting on registries is often inadequate, limiting their usefulness for evidence synthesis. In **Chapter 3**, we reviewed results reporting practices across clinical trials registries globally, identified barriers and facilitators to reporting and using results from trials registries, and developed recommendations for improvement. Only a minority of RCT registry records had results available. On a subset of registry records, we identified satisfactory (re-)usability for various results data available on the registries. Results reporting barriers include time, effort, and fear of interference with a publishable manuscript, while barriers to use results include mistrust in non-peer-reviewed data and difficulty locating them. Improving reporting and usability

requires better systems, infrastructures, policies, education, training, and funding. We developed expert-informed recommendations and identified key stakeholders to implement the proposed actions. With adequate support, registries have the potential to become essential tools for improving research transparency and efficiency in evidence dissemination.

The COVID-19 pandemic especially burdened low- and lower middle-income countries (LMICs). Clinical prediction models can support clinical decision-making by calculating an individual risk estimate for a COVID-19 patient to fall critically ill. However, none of the currently available prognostic models perform well in the LMIC setting. In **Chapter 4**, using data of the WHO Global Clinical Platform database for COVID-19, we developed two prognostic models for COVID-19-confirmed patients admitted to the hospital in LMICs at the first point of contact to assess 1) the risk of dying; and 2) the risk of requiring intensive care unit (ICU) admission. The final models included various demographic-, symptom-, and comorbidity-related predictors. We identified heterogeneity across countries in model performance measures. Both models performed well but showed some signs of miscalibration. While the two developed models provide ideal starting points for further validation in specific LMIC settings, we highly advise they should be further externally validated before integrating them into clinical care pathways.

Multiple tools exist for assessing the methodological quality of diagnosis and prognosis research. In **Chapter 5**, we provided an overview of existing methodological quality assessment (QA) tools for diagnosis and prognosis studies. We highlighted the overlap and differences among these tools and provided guidance for choosing a QA tool. We identified 14 QA tools and proposed a set of five guiding questions. Our guidance may help researchers, systematic reviewers, health policy makers, and guideline developers in specifying their purpose, and to select the most appropriate tool for their assessment.

Since 2019, the Prediction model Risk Of Bias ASsessment Tool (PROBAST: www.probast.org) has supported methodological quality assessments of prediction model studies. Bias refers to a systematic error in the estimators of the model's predictive performance for the target population of interest and may act in either direction such that there may be over- or under estimation of the true model performance. Using PROBAST, most prediction model studies are rated with a "High" risk of bias. Inter-rater reliability between assessors of the same study has been described as low. In **Chapter 6**, based on a case study, we assessed the inter-rater reliability of PROBAST ratings between assessors of the same study to understand reasons for discrepancies,

determined which items contributed most to domain-level risk of bias ratings, and explored the impact of consensus meetings between assessors of the same study. We found high inter-rater reliability at the overall study level, with some variation on item- and domain-level. To reduce variability, we advise PROBAST assessors to standardize item- and domain-level judgements and hold well-structured consensus meetings.

Advancements in prediction modelling methodologies, including artificial intelligence/machine learning based methods, and findings of Chapter 6 made an update of PROBAST necessary. In **Chapter 7**, we developed PROBAST+AI, an updated tool for assessing quality, risk of bias, and applicability in all types of studies developing or evaluating (validating) multivariable diagnostic and prognostic prediction models using any data analytical (statistical or artificial intelligence /machine learning) technique. PROBAST+AI replaces the original PROBAST and provides a more focused tool that allows all key stakeholders to examine the quality, risk of bias and applicability of any type of prediction model study in the healthcare sector, and embeds signalling questions addressing model fairness and algorithmic bias.

Incomplete research reporting limits its usefulness and contributes to research waste. Various reporting guidelines have been developed to aid researchers in reporting their study comprehensively. The completeness of reporting can be assessed by measuring adherence to such guidelines. However, adherence measurements lack uniformity. In 2019, a reporting adherence tool was developed to measure the completeness of reporting of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement. With recent advances in regression and artificial intelligence/machine learning based methods, TRIPOD+AI (www.tripod-statement.org) has been developed to replace the TRIPOD statement. In **Chapter 8**, we updated the TRIPOD adherence tool to TRIPOD+AI. The updated tool comprises the necessary adherence elements and includes several automated scoring rules. A uniform approach to assessing reporting adherence of TRIPOD+AI allows for comparisons across various medical fields, enables the monitoring of reporting over time, and incentivizes primary study authors to comply.

In conclusion, we have thoroughly evaluated various evidence synthesis methods, including registration, reporting, and quality assessment practices, used to inform evidence-based medicine. We provided key solutions to address current challenges. With collaborative efforts of the international research community, we aim to implement the proposed changes, leading to improved methods for evidence synthesis in medicine.

NEDERLANDSE SAMENVATTING

Systematische reviews leveren een synthese van het beschikbare wetenschappelijke bewijs en informeren daarmee evidence-based medicine. Ze zijn echter afhankelijk van de kwaliteit en volledigheid van het onderliggende bewijsmateriaal. Uitdagingen in de synthese van bewijs zijn onder andere het gebrek aan bewijs, bewijs van slechte kwaliteit en onvolledige methodologische richtlijnen voor de synthese van bewijsmateriaal. Voor diagnostisch en prognostisch onderzoek lijken deze uitdagingen nog prominenter. In dit proefschrift zijn verschillende aspecten van de synthese van bewijs in evidence-based medicine kritisch onderzocht met als doel betrouwbare synthesesmethoden te ontwikkelen. We hebben ons specifiek gericht op aspecten met betrekking tot methodologische richtlijnen en hun toepassing.

In **hoofdstuk 2** evalueerden we twee aanbevolen zoekstrategieën voor het identificeren van gerandomiseerde gecontroleerde studies (RCT's) voor systematische reviews van interventies. Het doel van dit onderzoek was om te bepalen of het doorzoeken van klinische onderzoeksregisters in vergelijking met databases van medische tijdschriften een hoger aantal afgeronde RCT's opleverde, en om het aantal relevante lopende RCT's te kwantificeren (relevant voor het bijwerken van systematische reviews). We doorzochten verschillende klinische onderzoeksregisters en medische tijdschriftendatabases voor RCT's op basis van de geschiktheidscriteria van een systematische review uit de cardiovasculaire sector als case study. We vergeleken de zoekresultaten om te bepalen in hoeverre de twee bronnen dezelfde RCT's identificeerden. De benadering op basis van de klinische onderzoeksregisters identificeerde een groter aantal voltooide RCT's. Beide zoekmethoden misten echter enkele RCT's. Als het doel is om alle in aanmerking komende RCT's te identificeren, moeten daarom beide bronnen worden doorzocht. Zoeken in alleen klinische onderzoeksregisters kan mogelijk zijn als de belemmeringen voor het zoeken in deze registers goed worden aangepakt, met name door ervoor te zorgen dat studieresultaten in registers worden gerapporteerd en door het vereenvoudigen van de identificatie van resultaten en publicaties die aan klinische onderzoeksregisters zijn gekoppeld.

Recente initiatieven en mandaten hebben rapportage van studieresultaten binnen klinische onderzoeksregisters mogelijk gemaakt. Dit verbetert de transparantie en verkleint de kans op publicatiebias. Lokale rapporten suggereren echter dat de rapportage van resultaten in registers vaak ontoereikend is, waardoor de bruikbaarheid van de resultaten voor de synthese van bewijsmateriaal wordt beperkt. In **hoofdstuk 3** hebben we de rapportage van resultaten van klinische trials in onderzoeksregisters

wereldwijd bekeken. We hebben barrières en faciliterende factoren voor het rapporteren en gebruiken van resultaten uit trials geïdentificeerd, en aanbevelingen voor verbetering ontwikkeld. Slechts een minderheid van de registraties van RCT's had resultaten beschikbaar. Voor een subset van registraties hebben we vastgesteld dat de (her)bruikbaarheid van verschillende in de registers beschikbare gegevens over resultaten voldoende is. Belemmeringen bij het rapporteren van resultaten zijn onder andere tijd, moeite en angst voor interferentie met een publiceerbaar manuscript, terwijl belemmeringen bij het gebruik van resultaten onder andere zijn: wantrouwen in niet-peer-reviewed gegevens en moeite om resultaten te vinden. Het verbeteren van de rapportage en de bruikbaarheid van studieresultaten vereist betere systemen, infrastructuur, beleid, onderwijs, training en financiering. We hebben aanbevelingen ontwikkeld en de belangrijkste belanghebbenden geïdentificeerd om de voorgestelde acties te implementeren. Met de juiste ondersteuning hebben registers het potentieel om essentiële hulpmiddelen te worden voor het verbeteren van onderzoekstransparantie en efficiëntie in het toegankelijk maken van bewijs.

De COVID-19-pandemie heeft vooral lage- en lagere middeninkomenslanden (LMIC's) zwaar getroffen. Klinische voorspelmodellen kunnen de klinische besluitvorming ondersteunen door een individuele risicoschatting te berekenen voor het ziekteverloop van een COVID-19-patiënt. Geen van de momenteel beschikbare prognosemodellen presteert echter goed in de LMIC-setting. In **hoofdstuk 4** ontwikkelden we met behulp van gegevens van de WHO Global Clinical Platform database voor COVID-19 twee prognostische modellen voor volwassen COVID-19-bevestigde patiënten bij het eerste contactmoment in het ziekenhuis in LMIC's om 1) het risico op overlijden en 2) het risico op opname op de intensive care (IC) te beoordelen. De uiteindelijke modellen bevatten verschillende demografische, symptoom- en comorbiditeitsgerelateerde predictoren. Er was heterogeniteit tussen landen in de voorspellende kwaliteiten van de modellen. Beide modellen presteerden goed, maar vertoonden tekenen van miscalibratie. Hoewel de twee ontwikkelde modellen ideale uitgangspunten bieden voor verdere validatie in specifieke LMICs, raden we ten eerste aan om ze eerst extern te valideren voordat ze worden geïntegreerd in klinische zorgpaden.

Er bestaan meerdere instrumenten om de methodologische kwaliteit van diagnose- en prognoseonderzoek te beoordelen. In **hoofdstuk 5** hebben we een overzicht gegeven van bestaande methodologische kwaliteitsbeoordelingsinstrumenten voor diagnose- en prognoseonderzoek. We benadrukten de overlap en verschillen tussen deze hulpmiddelen en gaven richtlijnen voor het kiezen van een kwaliteitsbeoordelingsinstrument. We identificeerden 14

kwaliteitsbeoordelingsinstrumenten en stelden een set van vijf leidende vragen voor. Onze richtlijnen kunnen onderzoekers, systematische reviewers, beleidsmakers op het gebied van gezondheid, en richtlijnontwikkelaars helpen bij het specificeren van hun doel en bij het kiezen van het meest geschikte kwaliteitsbeoordelingsinstrument voor hun beoordeling.

Sinds 2019 ondersteunt de Prediction Model Risk Of Bias Assessment Tool (PROBAST: www.probast.org) methodologische kwaliteitsbeoordelingen van voorspelmodelstudies. Bias verwijst naar een systematische fout in de schatters van de voorspellende prestaties van het model voor de doelpopulatie en kan in beide richtingen werken, zodat er sprake kan zijn van over- of onderschatting van de werkelijke modelprestaties. Met behulp van PROBAST worden de meeste onderzoeken naar voorspellingsmodellen beoordeeld met een “hoog” risico op bias. De interbeoordelaarsbetrouwbaarheid tussen beoordelaars van hetzelfde onderzoek is als laag omschreven. In **hoofdstuk 6** hebben we op basis van een case study de interbeoordelaarsbetrouwbaarheid van PROBAST-beoordelingen tussen beoordelaars van dezelfde studie beoordeeld om de redenen voor discrepanties te begrijpen. We hebben ook bepaald welke items het meest hebben bijgedragen aan de beoordelingen van het risico op bias op domeinniveau en we hebben de invloed van consensusbijeenkomsten tussen beoordelaars van dezelfde studie onderzocht. We vonden een hoge interbeoordelaarsbetrouwbaarheid op studieniveau, met enige variatie op item- en domeinniveau. Om de variabiliteit te verminderen, adviseren wij PROBAST-beoordelaars om beoordelingen op item- en domeinniveau te standaardiseren en goed gestructureerde consensusbijeenkomsten te houden.

De vooruitgang in voorspelmethodologieën, waaronder kunstmatige intelligentie (artificial intelligence, ‘AI’)/machine learning, en de bevindingen van hoofdstuk 6 maakten een update van PROBAST noodzakelijk. In **hoofdstuk 7** ontwikkelden we PROBAST+AI, een bijgewerkt hulpmiddel voor het beoordelen van kwaliteit, risico op bias en toepasbaarheid in alle soorten onderzoeken die multivariabele diagnostische en prognostische voorspelmodellen ontwikkelen of evalueren (valideren) ongeacht welke data-analytische techniek (statistische of AI/machine learning) gebruikt werd. PROBAST+AI vervangt de oorspronkelijke PROBAST en biedt een meer gericht hulpmiddel waarmee alle belangrijke belanghebbenden de kwaliteit, het risico van bias en de toepasbaarheid van elk type voorspelmodelstudie in de gezondheidszorg kunnen onderzoeken. Daarnaast bevat PROBAST+AI-items die betrekking hebben op fairness en algoritmische bias.

Onvolledige onderzoeksrapportage beperkt de bruikbaarheid van de resultaten en draagt bij aan onderzoeksverspilling. Er zijn verschillende rapportagerichtlijnen ontwikkeld om onderzoekers te helpen bij het volledig rapporteren van hun onderzoek. Het evalueren van de naleving van dergelijke richtlijnen kwantificeert de volledigheid van de rapportage. Echter, instrumenten om de kwaliteit van rapportage te scoren missen uniformiteit. In 2019 werd een rapportagetool ontwikkeld om de volledigheid van de rapportage van de TRIPOD-statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) te meten. Met de recente vooruitgang in statistische en AI/machine learning-gebaseerde methoden werd TRIPOD+AI (www.tripod-statement.org) ontwikkeld om de TRIPOD-statement te vervangen. In **hoofdstuk 8** hebben we het instrument om de kwaliteit van rapportage te scoren bijgewerkt voor TRIPOD+AI. De bijgewerkte tool bevat de noodzakelijke elementen kwaliteit van rapportage te bepalen, en bevat verschillende geautomatiseerde scoringsregels. Een uniforme aanpak voor het beoordelen van de naleving van het TRIPOD+AI-statement maakt vergelijkingen tussen verschillende gebieden mogelijk, maakt het monitoren van de rapportage in de loop van de tijd mogelijk en stimuleert auteurs van primaire onderzoeken om zich aan de rapportage regels te houden.

Concluderend hebben we verschillende methoden voor de synthese van bewijsmateriaal grondig geëvalueerd, inclusief registratie, rapportage en kwaliteitsbeoordelingspraktijken, die worden gebruikt om evidence-based geneeskunde te onderbouwen. We hebben belangrijke oplossingen aangedragen om de huidige uitdagingen aan te pakken. Met de gezamenlijke inspanningen van de internationale onderzoeksgemeenschap willen we de voorgestelde veranderingen implementeren, wat zal leiden tot verbeterde methoden voor de synthese van bewijsmateriaal in de geneeskunde.

DEUTSCHE ZUSAMMENFASSUNG

Systematische Reviews leisten einen zentralen Beitrag zur evidenzbasierten Medizin, indem sie die verfügbare wissenschaftliche Evidenz zusammenfassen. Ihre Aussagekraft hängt jedoch von der Qualität und Vollständigkeit der zugrunde liegenden Evidenz ab. Zu den Herausforderungen bei der Evidenzsynthese zählen fehlende Evidenz, verzerrte Evidenz sowie das Fehlen etablierter methodischer Leitlinien. Besonders im Bereich der Prognoseforschung scheinen diese Herausforderungen noch ausgeprägter zu sein. Die vorliegende Dissertation setzt sich kritisch mit verschiedenen Aspekten der Evidenzsynthese im Rahmen systematischer Reviews in der evidenzbasierten Medizin auseinander – mit dem Ziel, verlässlichere Methoden zur Evidenzsynthese zu entwickeln. Ein besonderer Fokus liegt dabei auf methodischen Leitlinien und deren praktischer Anwendung.

In **Kapitel 2** haben wir zwei empfohlene Suchstrategien zur Identifizierung randomisierter kontrollierter Studien (RCTs) für systematische Reviews im Bereich von Interventionen evaluiert. Ziel der Studie war es, festzustellen, ob die Recherche in klinischen Studienregistern im Vergleich zu Datenbanken medizinischer Fachzeitschriften eine größere Anzahl abgeschlossener RCTs liefert und wie viele geeignete laufende RCTs – relevant für die Aktualisierung systematischer Reviews – auf diese Weise identifiziert werden können. Dazu wurde in verschiedenen klinischen Studienregistern sowie in medizinischen Fachliteraturdatenbanken gezielt nach RCTs gesucht, basierend auf den Einschlusskriterien eines systematischen Reviews aus dem Bereich der Herz-Kreislauf-Erkrankungen, der als Fallbeispiel diente. Die Suchergebnisse beider Quellen wurden miteinander verglichen, um zu ermitteln, inwieweit sie dieselben Studien identifizierten. Die Analyse zeigte, dass über die Studienregister eine größere Anzahl abgeschlossener RCTs gefunden wurde. Allerdings verfehlten beide Ansätze jeweils einige relevante Studien. Daher sollten – zumindest derzeit – beide Quellen parallel durchsucht werden, um eine möglichst vollständige Erfassung aller relevanten RCTs sicherzustellen. Eine ausschließliche Recherche in klinischen Studienregistern könnte dann in Betracht gezogen werden, wenn bestehende Suchbarrieren behoben sind. Dies setzt insbesondere eine konsequente Ergebnisberichterstattung in den Registern sowie die Einführung vereinfachter Abrufmechanismen für Studienergebnisse und zugehöriger Publikationen voraus.

Initiativen und regulatorische Vorgaben haben die Ergebnisberichterstattung klinischer Studien über Studienregister eingeführt, um Transparenz zu fördern und Publikationsverzerrungen zu reduzieren. Regionale Forschungsberichte weisen jedoch

darauf hin, dass die Ergebnisberichterstattung in vielen Studienregistern weiterhin unzureichend ist, was deren Mehrwert für evidenzbasierte Synthesen erheblich einschränkt. In **Kapitel 3** wurde die Ergebnisberichterstattung randomisierter kontrollierter Studien in sechs verschiedenen klinischen Studienregistern systematisch untersucht. Dabei wurden die Nutzbarkeit der Einträge analysiert, Gründe für fehlende Ergebnisberichte identifiziert und praxisorientierte Empfehlungen zur Verbesserung der aktuellen Situation entwickelt. Unsere Analyse zeigt, dass nur ein geringer Anteil der untersuchten Einträge Ergebnisdaten enthielt. Bei einer Stichprobe konnten wir jedoch für bestimmte Ergebnisformate eine zufriedenstellende (Wieder-)Verwendbarkeit feststellen. Die ausbleibende Ergebnisberichterstattung lässt sich auf mehrere Faktoren zurückführen, darunter Zeitmangel, administrativer Aufwand sowie die Befürchtung, dass eine frühzeitige Veröffentlichung von Ergebnissen in Studienregistern die spätere Publikation in Fachzeitschriften beeinträchtigen könnte. Zusätzlich wird die Nutzung von Studienergebnisse aus klinischen Studienregistern durch deren begrenzten Auffindbarkeit sowie durch ein verbreitetes Misstrauen gegenüber nicht peer-reviewten Daten eingeschränkt. Zur Verbesserung der Ergebnisberichterstattung und der damit verbundenen Nutzbarkeit klinischer Studienregister sind gezielte Maßnahmen erforderlich – insbesondere der Ausbau technischer Infrastrukturen, verbesserte Anreizsysteme, standardisierte Berichtsformate, umfassende Schulungsangebote sowie ausreichende finanzielle und personelle Ressourcen. Basierend auf den Ergebnissen unserer Untersuchung sowie dem Input von Fachexpert:innen haben wir konkrete Handlungsempfehlungen formuliert und zentrale Akteursgruppen identifiziert, die für die Implementierung entsprechender Maßnahmen maßgeblich sind. Mit entsprechender Unterstützung besitzen klinische Studienregister das Potenzial, sich zu unverzichtbaren Instrumenten zur Förderung von Forschungstransparenz und zur Effizienzsteigerung in der Verbreitung wissenschaftlicher Evidenz zu entwickeln.

Die COVID-19-Pandemie hat insbesondere Länder mit niedrigem und mittlerem Einkommen (LMICs) stark belastet. Klinische Prognosemodelle können die medizinische Entscheidungsfindung unterstützen, indem sie eine individualisierte Risikoeinschätzung für einen kritischen Krankheitsverlauf bei COVID-19-Patient:innen ermöglichen. Bislang sind jedoch keine der derzeit verfügbaren Modelle in LMICs zuverlässig einsetzbar. In **Kapitel 4** entwickelten wir zwei Prognosemodelle unter Verwendung von Daten der WHO Global Clinical Platform für COVID-19. Ziel war es, bei erwachsenen COVID-19-Patient:innen, die in LMICs bei der Erstvorstellung in ein Krankenhaus aufgenommen wurden, 1) das Mortalitätsrisiko und 2) das Risiko für den Bedarf einer intensivmedizinischen Behandlung abzuschätzen. Die endgültigen Modelle

enthielten verschiedene demografische, symptom- und komorbiditätsbezogene Prädiktoren. Es zeigte sich eine länderübergreifende Heterogenität in der Modellgüte. Beide Modelle zeigten insgesamt gute prädiktive Leistungen, wiesen jedoch Hinweise auf eine unzureichende Kalibrierung auf. Obwohl die entwickelten Modelle eine vielversprechende Grundlage für weiterführende Validierungsstudien in spezifischen LMIC-Kontexten darstellen, sprechen wir uns ausdrücklich für eine externe Validierung aus, bevor eine Integration in klinische Versorgungsabläufe erfolgen kann.

Zur Bewertung der methodischen Qualität von Diagnose- und Prognosestudien steht eine Vielzahl von Instrumenten zur Verfügung. In **Kapitel 5** haben wir einen systematischen Überblick über bestehende Bewertungsinstrumente gegeben, deren Gemeinsamkeiten und Unterschiede herausgearbeitet und eine praxisorientierte Anleitung zur Auswahl eines geeigneten Instruments erarbeitet. Insgesamt identifizierten wir 14 Instrumente zur Qualitätsbewertung und leiteten daraus fünf zentrale Leitfragen ab, die als Orientierungshilfe bei der Auswahl dienen können. Unser Leitfaden richtet sich an Forschende, systematische Reviewer:innen, Leitlinienentwickler:innen und Entscheidungsträger:innen im Gesundheitswesen. Er unterstützt sie dabei, ihre jeweilige Forschungsfrage präzise zu formulieren und das methodisch am besten geeignete Bewertungsinstrument für ihre spezifische Fragestellung auszuwählen.

Seit 2019 steht mit dem *Prediction model Risk Of Bias ASsessment Tool* (PROBAST; www.probast.org) ein etabliertes Instrument zur methodischen Qualitätsbewertung von Studien zu Vorhersagemodellen zur Verfügung. *Bias* bezeichnet hierbei einen systematischen Fehler, der zur Verzerrung in der Schätzung der Modellgüte in Bezug auf die Zielpopulation führt. Solche Verzerrungen können sowohl zu einer Überschätzung als auch zu einer Unterschätzung der tatsächlichen Modellgüte führen. In der Praxis werden die meisten Vorhersagemodellstudien mit PROBAST als mit einem „hohen“ Verzerrungsrisiko behaftet bewertet. Gleichzeitig wurde die Inter-Rater-Reliabilität – also die Übereinstimmung zwischen verschiedenen Bewerter:innen bei der Einschätzung derselben Studie – als gering beschrieben. In **Kapitel 6** haben wir im Rahmen einer Fallstudie die Inter-Rater-Reliabilität der PROBAST-Bewertungen analysiert, um Ursachen für Abweichungen zwischen Bewerter:innen besser zu verstehen. Dabei untersuchten wir, welche Items den größten Einfluss auf die Bewertung des Verzerrungsrisikos auf Domänenebene hatten und welche Rolle strukturierte Konsenssitzen zwischen den Bewerter:innen spielten. Unsere Ergebnisse zeigten eine insgesamt hohe Inter-Rater-Reliabilität auf Studienebene, jedoch Unterschiede auf Item- und Domänenebene. Zur Reduktion von Bewertungsvariabilität empfehlen wir eine stärkere Standardisierung der

Bewertung auf Item- und Domänenebene sowie den gezielten Einsatz gut strukturierter Konsenssitzungen.

Die jüngsten Fortschritte in der Entwicklung von Vorhersagemodellen – insbesondere durch den Einsatz von künstlicher Intelligenz und maschinellem Lernen – sowie die Erkenntnisse aus Kapitel 6 machten eine Aktualisierung von PROBABT erforderlich. In **Kapitel 7** entwickelten wir PROBABT+AI, ein erweitertes Bewertungsinstrument zur Beurteilung der Qualität, des Verzerrungsrisikos und der Anwendbarkeit von Studien, die multivariable diagnostische oder prognostische Vorhersagemodelle mithilfe beliebiger Datenanalysemethoden – sei es durch klassische statistische Verfahren oder durch Ansätze der künstlichen Intelligenz bzw. des maschinellen Lernens – entwickeln oder validieren. PROBABT+AI ersetzt das ursprüngliche PROBABT und bietet ein zielgerichteteres, zeitgemäßes Instrument, das es allen relevanten Interessengruppen – darunter Forschenden, systematische Reviewer:innen, Leitlinienentwickler:innen und Entscheidungsträger:innen im Gesundheitswesen – ermöglicht, die methodische Qualität, das Verzerrungsrisiko sowie die Anwendbarkeit jeglicher Art von Vorhersagemodellstudie systematisch zu bewerten. Darüber hinaus integriert PROBABT+AI erstmals auch Aspekte wie Fairness und Algorithmic Bias.

Unvollständige wissenschaftliche Berichterstattung mindert deren Nutzen und trägt zur Verschwendung von Forschungsressourcen bei. Um Forschende bei der umfassenden und transparenten Darstellung ihrer Studien zu unterstützen, wurden verschiedene Leitlinien für die Berichterstattung entwickelt. Die Vollständigkeit der Berichterstattung kann durch die Messung der Adhärenz an solche Leitlinien bewertet werden. Allerdings mangelt es derzeit an einem einheitlichen Vorgehen zur Adhärenzmessung. Im Jahr 2019 wurde ein Instrument zur Bewertung der Berichterstattungsadhärenz in Bezug auf die TRIPOD-Richtlinie (*Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis*) entwickelt. Angesichts der jüngsten Fortschritte in regressionsbasierten Methoden sowie in Verfahren der künstlichen Intelligenz und des maschinellen Lernens wurde TRIPOD+AI (www.tripod-statement.org) als aktualisierte Version zur Ablösung der ursprünglichen TRIPOD-Richtlinie entwickelt. In **Kapitel 8** haben wir das TRIPOD-Adhärenz-Tool entsprechend an TRIPOD+AI angepasst und aktualisiert. Das überarbeitete Instrument umfasst alle erforderlichen Adhärenzelemente und enthält mehrere automatisierte Bewertungsregeln. Ein standardisierter Ansatz zur Bewertung der Adhärenz an TRIPOD+AI ermöglicht disziplinübergreifende Vergleiche, unterstützt die zeitliche Überwachung der Berichterstattungspraxis und bietet Autor:innen primärer Studien einen Anreiz zur Einhaltung hoher Berichtstandards.

Abschließend haben wir verschiedene Methoden der Evidenzsynthese – einschließlich Registrierung, Berichterstattung und Verfahren zur Qualitätsbewertung – umfassend evaluiert, wie sie zur Unterstützung der evidenzbasierten Medizin eingesetzt werden. Wir haben zentrale Lösungsansätze zur Bewältigung aktueller Herausforderungen aufgezeigt. Mit gemeinschaftlichen Anstrengungen der internationalen Forschungsgemeinschaft streben wir die Umsetzung der vorgeschlagenen Veränderungen an – mit dem Ziel, die Methoden der Evidenzsynthese in der Medizin nachhaltig zu verbessern.

LIST OF MANUSCRIPTS

Manuscripts published and included in this doctoral thesis

Kaul T, Colombijn JMT, Vernooij RWM, Spijker R, Idema DL, Huis In 't Veld LF, Damen JAA, Hooft L. Both clinical trial register and electronic bibliographic database searches were needed to identify randomized clinical trials for systematic reviews: an evaluation study. *J Clin Epidemiol*. 2024 May;169:111300. doi: 10.1016/j.jclinepi.2024.111300. Epub 2024 Feb 24. PMID: 38402998.

Kaul T & Kellerhuis BE, Damen JAA, Schuit E, Jenniskens K, van Smeden M, Reitsma JB, Hooft L, Moons KGM, Yang B. Methodological quality assessment tools for diagnosis and prognosis research: overview and guidance. *J Clin Epidemiol*. 2025 Jan;177:111609. doi: 10.1016/j.jclinepi.2024.111609. Epub 2024 Nov 12. PMID: 39536993.

Kaul T, Damen JAA, Wynants L, Van Calster B, van Smeden M, Hooft L, Moons KGM. Assessing the quality of prediction models in health care using the Prediction model Risk Of Bias ASsessment Tool (PROBAST): an evaluation of its use and practical application. *J Clin Epidemiol*. 2025 Feb 25;181:111732. doi: 10.1016/j.jclinepi.2025.111732. Epub ahead of print. PMID: 40010583.

Moons KGM, Damen JAA, **Kaul T**, Hooft L, Andaur Navarro C, Dhiman P, Beam AL, Van Calster B, Celi LA, Denaxas S, Denniston AK, Ghassemi M, Heinze G, Kengne AP, Maier-Hein L, Liu X, Logullo P, McCradden MD, Liu N, Oakden-Rayner L, Singh K, Ting DS, Wynants L, Yang B, Reitsma JB, Riley RD, Collins GS, van Smeden M. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025 Mar 24;388:e082505. doi: 10.1136/bmj-2024-082505. PMID: 40127903; PMCID: PMC11931409.

Manuscripts published but not included in this doctoral thesis

Colombijn JMT, Idema DL, van Beem S, Blokland AM, van der Braak K, Handoko ML, In 't Veld LFH, **Kaul T**, Kolagasigil-Akdemir N, Kusters MPT, Meijvis SCA, Oosting IJ, Spijker R, Bots ML, Hooft L, Verhaar MC, Vernooij RWM. Representation of Patients With Chronic Kidney Disease in Clinical Trials of Cardiovascular Disease Medications: A Systematic Review. *JAMA Netw Open*. 2024 Mar 4;7(3):e240427. doi: 10.1001/jamanetworkopen.2024.0427. PMID: 38451526; PMCID: PMC10921252.

Manuscripts in preparation

Kaul T, Damen JAA, Seidler AL, Willson M, Karam G, Idema D, Kusters M, Dowsett MA, Abutahoun TIH, Hooft L, Hunter KE. Improving the Reporting and Use of Trial Results in Clinical Trials Registries: Global Practices, Barriers, and Recommendations.

Kaul T, van den Brink J, de Jong VMT, Bertagnolio S, Diaz JV, Silva R, Thwin SS, Wynants L, van Smeden M, Moons KGM, Damen JAA. Development of Prognostic Models for In-Hospital Mortality and ICU Admission in Confirmed COVID-19 Patients: An IPD Meta-Analysis from LMICs.

Kaul T & de Kanter E, Heus P, Damen JAA, Kuijten RH, de Groot TM, Reitsma JB, Collins GS, Hooft L, Moons KGM. Adherence to TRIPOD+AI Guideline: An Updated Reporting Assessment Tool.

AUTHOR AFFILIATIONS

Tabea Kaul

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Karel GM Moons

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Lotty Hooft

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Johanna AA Damen

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Julia MT Colombijn

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht, Netherlands

Robin WM Vernooij

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht, Netherlands

Rene Spijker

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Amsterdam UMC, University of Amsterdam, Amsterdam Public Health, Medical Library, Amsterdam, Netherlands

Demy L Idema

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Linde F Huis in 't Veld

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Anna Lene Seidler

NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

Department of Child and Adolescent Psychiatry, University Medical Centre Rostock, Rostock, Germany

German Center for Child and Adolescent Health (DZKJ), partner site Greifswald/Rostock, Rostock, Germany

Melina Willson

NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

Ghassan Karam

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Mike Kusters

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Mary Ann Dowsett

NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

Tala Ibrahim Hasan Abutahoun

NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

Kylie E Hunter

NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia

Jasper van den Brink

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Valentijn MT de Jong

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Data Analytics and Methods Task Force, European Medicines Agency, Amsterdam, Netherlands

Silvia Bertagnolio

World Health Organization, Geneva, Switzerland

Janet V Diaz

World Health Organization, Geneva, Switzerland

Ronaldo Silva

World Health Organization, Geneva, Switzerland

Soe Soe Thwin

World Health Organization, Geneva, Switzerland

Laure Wynants

Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

Department of Development and Regeneration, KU Leuven, Leuven, Belgium

Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

Maarten van Smeden

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Bas E Kellerhuis

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Ewoud Schuit

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Kevin Jenniskens

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Johannes B Reitsma

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Bada Yang

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Ben Van Calster

Department of Development and Regeneration, KU Leuven, Leuven, Belgium
Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

Constanza Andaur Navarro

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Paula Dhiman

Centre for Statistics in Medicine, UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, United Kingdom

Andrew L Beam

Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

Leo Anthony Celi

Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA
Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA

Spiros Denaxas

Institute of Health Informatics, University College London, London, United Kingdom
British Heart Foundation Data Science Centre, Health Data Research Centre UK, London, United Kingdom

Alastair K Denniston

College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom

Marzyeh Ghassemi

Department of Electrical Engineering and Computer Science; Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Georg Heinze

Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria

André Pascal Kengne

Department of Medicine, University of Cape Town, Cape Town, South Africa

Lena Maier-Hein

Division of Intelligent Medical Systems (IMSY), German Cancer Research Center (DKFZ), Heidelberg, Germany

National Centre for Tumour Diseases (NCT) Heidelberg, Heidelberg, Germany

Xiaoxuan Liu

College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom

University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom

School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom

National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, United Kingdom

Patricia Logullo

Centre for Statistics in Medicine, UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, United Kingdom

Melissa D McCradden

Department of Bioethics, The Hospital for Sick Children, Toronto, Ontario, Canada

Nan Liu

Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore

Lauren Oakden-Rayner

Australian Institute for Machine Learning, University of Adelaide, Adelaide, South Australia, Australia

Karandeep Singh

Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan, 48109, USA

Daniel S Ting

Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore
AI Office, Singapore Health Service; Duke-NUS Medical School, Singapore

Richard D Riley

School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom
National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, United Kingdom

Gary S Collins

Centre for Statistics in Medicine, UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, United Kingdom

Emilie de Kanter

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Pauline Heus

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

R Harmen Kuijten

Division of Imaging and Oncology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Tom M de Groot

Department of Orthopaedic Surgery, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

PHD PORTFOLIO

Name of PhD student: Tabea Kaul
 UMCU department: Julius Center – Epidemiology – Methodology
 PhD period: February 2022 – February 2025
 Promotoren: Prof.dr. Karel GM Moons
 Prof.dr. Lotty Hooft
 Co-promotor: Dr. Johanna AA Damen

1. PhD training

Courses at the MSc Epidemiology (Track: Clinical Epidemiology)		Year	ECTS
Mandatory Courses			
Introduction to Epidemiology	2022	1.5	
Introduction to Statistics	2022	1.5	
Study Design in Etiologic Research	2023	3.0	
Classical Methods in Data Analysis	2023	6.0	
Modern Methods in Data Analysis	2023	4.5	
Writing Research Proposals	2023	1.5	
Research Ethics an Introduction	2023	1.5	
Presenting Your Research Confidently	2024	0.5	
Systematic Reviews in Intervention Research	2024	1.5	
Clinical Epidemiology	2024	1.5	
Clinical Trials and Drug Risk Assessment	2024	1.5	
Electives			
Prognostic Research	2023	1.5	
Systematic Reviews in Diagnostic Studies	2024	1.5	
Systematic Reviews and Meta-Analysis of Prognosis Studies	2024	1.5	
Missing Data	2024	1.5	
Methodology in Health Economic Evaluation	2024	1.5	
Systematic Reviews and Meta-Analysis of Individual Participant Data	2024	1.5	
Big Data	2024	1.5	
Research project	2024	56	
Courses at the PhD Graduate School of Life Sciences			
Writing a Scientific Paper	2022	1.5	
The Art of Presenting Science	2023	1.4	
Introduction to R for Life Sciences	2024	3.0	
Responsible Conduct of Research – Year 1	2022	0.15	
Responsible Conduct of Research – Year 2	2023	0.15	

Responsible Conduct of Research – Year 3	2024	0.15
Responsible Conduct of Research – Year 4	2024	0.15
Conferences and Scientific Meetings		
Cochrane Colloquium – Two oral presentations	2023	1.0
WEON – poster presentation	2023	1.0
WEON – pre-conference workshop (3h)	2024	-
Global Evidence Summit – Oral presentation	2024	1.0
Teaching		
START block	2022	-
START block	2024	-
Supervising master's student research internship (1 student)	2024	-
Parameters of Esteem		
Strategic Network Development Grant for research visit	2024	-
Other Activities		
Peer reviewing for international journals (2 manuscripts)	2024	-
Visiting PhD researcher at NHMRC Clinical Trials Centre, University of Sydney, Sydney, Australia (February – April)	2024	-

CURRICULUM VITAE

Tabea Kaul was born on August 23rd, 1993, in Arnstadt, Germany. During her youth, Tabea lived in Germany, the United States, and Spain. In 2014, she began her career in healthcare with her paramedic training at Malteser and the Bavarian Red Cross, where she discovered her passion for public health. In 2016, she pursued a bachelor's degree in health sciences at the Technical University of Munich. During her studies, she interned at the Helmholtz Association in Brussels and participated in an exchange program at Keio University in Tokyo. Her bachelor's thesis was a systematic review



on HIV pre-exposure prophylaxis. Tabea then continued her academic career at the Vrije Universiteit Amsterdam, where she earned a master's in health sciences, focusing on infectious diseases. She completed her master's thesis on active tuberculosis case finding in Pakistan at the Royal Tropical Institute (KIT) in Amsterdam. Following her master's degree, Tabea worked at a COVID-19 vaccination center and a women's clinic. In 2022, Tabea started her PhD at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, supervised by prof. dr. Carl Moons, prof. dr. Lotty Hooft, and dr. Anneke Damen. During her PhD, she earned a postgraduate degree in clinical epidemiology, completed a two-month research visit at the University of Sydney under the supervision of dr. Kylie Hunter, supervised a student, and presented her work at both Dutch and international conferences, including a pre-conference workshop. Tabea enjoys topics related to evidence synthesis, with a particular interest in healthcare policy and infectious disease research.

The results of her PhD research, titled 'Evidence Synthesis in Evidence-Based Medicine: Guidance and Application' are presented in this thesis.

ACKNOWLEDGEMENTS

Completing this PhD has been an incredibly enriching and transformative experience, and I would like to express my heartfelt gratitude to all those who supported me along the way.

First and foremost, I am deeply thankful to my supervisors—Carl, Lotty, and Anneke—for your trust, guidance, and continuous encouragement throughout the past years. Your mentorship shaped not only the quality of the research projects in this thesis but also my growth as a researcher. I truly appreciated your openness, expertise, and the supportive atmosphere you created—both in the academic setting and beyond.

Carl, your enthusiasm, energy, and guidance were always motivating. Lotty, thank you for granting me the freedom to explore and for encouraging my academic independence. The opportunity to do a research exchange was crucial for my confidence and development. Anneke, your ongoing support, and thoughtful feedback were instrumental across all projects. Working together so closely has been a privilege, and I deeply value the trust and collaboration we shared.

To all co-authors and collaborators involved in the different projects— thank you for your valuable input, constructive discussions, and fresh perspectives. Your contributions greatly enriched the work presented in this thesis and challenged me to think more critically and creatively.

A special thank-you goes to Constanza. From the very beginning, you were there to guide and support me—both academically and personally. I am especially grateful that our collaboration continued even after you had moved on to a new position.

I also want to thank my paranympths, Amanda and Anum:

Amanda, thank you for bringing me back to the Netherlands and for being a constant support throughout this journey. From our master's in Amsterdam to coffee chats at work, weekend trips, and shared milestones – it meant a lot to go through this with you.

Anum, we started this journey together, and I am so grateful we get to finish it side by side. What we built went far beyond work — it turned into a friendship I deeply treasure. Thank you for being part of it all.

A huge thank you to the Epi-Methods team for creating a collaborative and motivating work environment. Special thanks to Kim, for always joining me on the dance floor at conferences, and to Lotta, for bringing such joy to R problem-solving sessions.

To the team at the University of Sydney—Kylie, Lene, Melina, and many others—thank you for the warm welcome and inspiring exchange. This experience gave me a renewed sense of confidence and belonging in the research community. I am especially grateful for the continued collaboration even across continents and time zones.

To my fellow PhD candidates in the van Geuns building, especially Pauline, Julia, Vivian, Jorna, Sophie, Paul, Lieke, and Niklas: thank you for the shared coffees, conversations, and support. It meant a lot to be part of a group that celebrated both successes and setbacks together.

To my friends in the Netherlands – Anaïs, Dina, Bea, Syrka, Claudia, and ‘the Villagers’ : thank you for being part of this chapter in my life. Whether it was long walks, runs, yoga sessions, weekend getaways, spontaneous nights out, or simply laughing over a glass of wine or coffee – your presence made this time lighter and more joyful. I will always carry these memories with me, wherever I go.

Ein besonderer Dank gilt meinen Freund*innen in Deutschland und Österreich:

Алёна, Kathi, Julia M., Julia N. & David K., David M., Michi, Anna und Marie – eure Freundschaft hat mir in intensiven Phasen Kraft gegeben. Ob in München, Regensburg, oder Wien: Danke für euer offenes Ohr, eure Geduld und all die Momente, in denen ich einfach ich selbst sein konnte. Ihr habt mir immer wieder gezeigt, was Heimat bedeutet.

Vor allem aber möchte ich meiner Familie von Herzen danken:

Mama und Andreas – danke für eure unermüdliche Unterstützung, eure Liebe, und die aufmunternden Worte, zu jeder Tag- und Nachtzeit.

Opa Dieter und Birgit – danke für euer anhaltendes Interesse und dafür, dass ihr mir stets einen willkommenen Anlass gegeben habt, meine Forschung in verständlicher Alltagssprache zu erklären.

Andi – ohne dich hätte ich diese letzten Jahre nicht so durchstehen können. Danke für deine Geduld, dein Verständnis und dafür, dass du in jeder Phase für mich da warst. Du warst mein Rückhalt, mein Ruhepol, mein Zuhause.

