

The importance of methodology and transparency in predictive AI

Gary S. Collins, PhD
Professor of Medical Statistics

University of Birmingham, UK

01-October-2026

Declaration of Helsinki

Methods →

Scientific Requirements and Research Protocols

21. Medical research involving human participants must have a scientifically sound and rigorous design and execution that are likely to produce reliable, valid, and valuable knowledge and avoid research waste. The research must conform to generally accepted scientific principles, be based on a thorough knowledge of the scientific literature, other relevant sources of information, and adequate laboratory and, as appropriate, animal experimentation.

The welfare of animals used for research must be respected.

Protocols →

22. The design and performance of all medical research involving human participants must be clearly described and justified in a research protocol.

The protocol should contain a statement of the ethical considerations involved and should indicate how the principles in this Declaration have been addressed. The protocol should include information regarding aims, methods, anticipated benefits and potential risks and burdens, qualifications of the researcher, sources of funding, any potential conflicts of interest, provisions to protect privacy and confidentiality, incentives for participants, provisions for treating and/or compensating participants who are harmed as a consequence of participation, and any other relevant aspects of the research.

In clinical trials, the protocol must also describe any post-trial provisions.

Funding
COIs
Expertise
Benefits →

Research Registration and Publication and Dissemination of Results

35. Medical research involving human participants must be registered in a publicly accessible database before recruitment of the first participant.

← Registration

36. Researchers, authors, sponsors, editors, and publishers all have ethical obligations with regard to the publication and dissemination of the results of research. Researchers have a duty to make publicly available the results of their research on human participants and are accountable for the timeliness, completeness, and accuracy of their reports. All parties should adhere to accepted guidelines for ethical reporting. Negative and inconclusive as well as positive results must be published or otherwise made publicly available. Sources of funding, institutional affiliations, and conflicts of interest must be declared in the publication. Reports of research not in accordance with the principles of this Declaration should not be accepted for publication.

← Reporting

Special Communication

October 19, 2024

World Medical Association Declaration of Helsinki Ethical Principles for Medical Research Involving Human Participants

World Medical Association

Article Information

JAMA. 2025;333(1):71-74. doi:10.1001/jama.2024.21972

Methodology over metrics

- The overall quality of medical research remains poor, despite longstanding criticisms.
- The scientific enterprise is business-like and consistently undervalues its own backbone, methodology.
- Despite great initiatives to improve research quality, progress is modest.
- Top-down action from journals, funding agencies, universities and governments is needed to break the cycle. These actions should give methodology a central place in funding acquisition as well as study design, conduct, and reporting.



Journal of Clinical Epidemiology 138 (2021) 219–226

Journal of
Clinical
Epidemiology

COMMENTARY

Methodology over metrics: current scientific standards are a disservice to patients and society

Ben Van Calster^{a,b,c,*}, Laure Wynants^{a,c,d}, Richard D Riley^e, Maarten van Smeden^f, Gary S Collins^{g,h,i}

Daly et al. *Trials* (2025) 26:110
<https://doi.org/10.1186/s13063-025-08747-4>

Trials

METHODOLOGY

Open Access

Tolerating bad health research (part 2): still as many bad trials, but more good ones too

Anna Daly^{1,2}, Shaun Treweek³, Genevieve Shiely Hayes^{1,2} and Frances Shiely^{1,2,4*}



Purpose of the research publication

- **Articles are written for multiple readerships:**
 - **Healthcare professionals**
 - To learn how to treat their patients better
 - **Researchers:**
 - To inform their own research
 - To help plan a similar study
 - To include the study in a systematic review/evidence synthesis
 - **Patients/consumers:**
 - To aid personal decision-making
 - **Policy makers/purchasers:**
 - To aid policy decision-making
- ... should present sufficiently detailed information to allow assessment of study **reliability and relevance** and **comparison across studies**

Obligation

- Scientific manuscripts should present sufficient information so that the reader can fully evaluate this new information and reach their own conclusions about the results
 - Often the only tangible evidence that the study was ever done
- We need research we can rely on
- Transparent reporting is an essential part of good research
 - open science, reproducibility and research(er) integrity

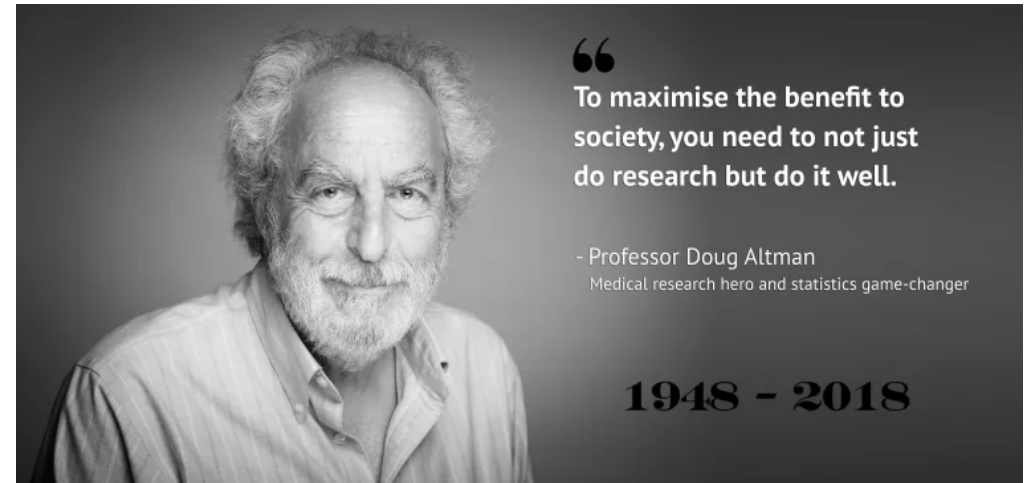
Transparency & Reporting



www.equator-network.org

"Readers should not have to infer what was probably done, they should be told explicitly"

Altman, BMJ 1996



Sauerbrei et al, Biom J 2021

Research waste from poor reporting

- “**inadequate reporting occurs in all types of studies**—animal and other preclinical studies, diagnostic studies, epidemiological studies, **clinical prediction research [predictive AI]**, surveys, and qualitative studies”
- “high amount of waste also warrants **future investment** in the **monitoring of and research into reporting of research**, and active implementation of the findings to **ensure that research** reports better **address the needs** of the range of **research users**”

Reporting guidelines

- They are a **minimum** set of essential items when reporting a study
 - **Reminders of scientific content** for authors
 - Recommendations and guidance, not requirements
 - Depends on journal enforcement
- Based on **evidence** and **international consensus**
 - Community driven typically involving a multidisciplinary group
- Often accompanied by a long **Explanation & Elaboration** (E&E) paper
 - **Rationale** on the importance of the items
 - **Examples** of good reporting
 - **Educational**
- The **EQUATOR Network** (an international initiative) brings all the guidelines together
 - Promotes **transparent and accurate reporting** of health research

www.equator-network.org



The EQUATOR (Enhancing the QUALity and Transparency Of health Research) Network is an international initiative that seeks to improve the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines.

It is the first coordinated attempt to tackle the problems of inadequate reporting systematically and on a global scale; it advances the work done by individual groups over the last 15 years.



Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	
Quality improvement studies	SQUIRE	Extensions
Economic evaluations	CHEERS	Extensions

[See all 659 reporting guidelines](#)

Reporting guidelines are endorsed by

- Journals & publishers
- Research organisations
- Editorial organisations
- Funders



ICMJE

laboration will not always be possible, practical, or desired, the efforts of those who generated the data must be recognized.

IV. MANUSCRIPT PREPARATION AND SUBMISSION

A. Preparing a Manuscript for Submission to a Medical Journal

1. General Principles

The text of articles reporting original research is usually divided into Introduction, Methods, Results, and Discussion sections. This so-called "IMRAD" structure is not an arbitrary publication format but a reflection of the process of scientific discovery. Articles often need sub-headings within these sections to further organize their content. Other types of articles, such as meta-analyses, may require different formats, while case reports, narrative reviews, and editorials may have less structured or unstructured formats.

www.icmje.org

the primary manuscript.

2. Reporting Guidelines

Reporting guidelines have been developed for different study designs; examples include CONSORT (www.consort-statement.org) for randomized trials, STROBE for observational studies (<http://strobe-statement.org/>), PRISMA for systematic reviews and meta-analyses (<http://prisma-statement.org/>), and STARD for studies of diagnostic accuracy (<http://www.equator-network.org/reporting-guidelines/stard/>). Journals are encouraged to ask authors to follow these guidelines because they help authors describe the study in enough detail for it to be evaluated by editors, reviewers, readers, and other researchers evaluating the medical literature. Authors are encouraged to refer to the SAGER guidelines for reporting of sex and gender information in study design, data analyses, results, and interpretation of findings: www.equator-network.org/reporting-guidelines/sager-guidelines/. Authors of review manuscripts are

15

Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals

encouraged to describe the methods used for locating, selecting, extracting, and synthesizing data; this is mandatory for systematic reviews. Good sources for reporting guidelines are the EQUATOR Network (www.equator-network.org/home/) and the NLM's Research Reporting Guidelines and Initiatives (www.nlm.nih.gov/services/research_report_guide.html).

figures and tables were actually included with the manuscript and, because tables and figures occupy space, to assess if the information provided by the figures and tables warrants the paper's length and if the manuscript fits within the journal's space limits.

Disclosure of relationships and activities. Disclosure information for each author needs to be part of the manuscript; each journal should develop standards with regard to the form the information should take and

Reporting of prediction models: 'pre-Machine learning' era (i.e., regression models)

Example: 228 articles [development of 408 prognostic models for patients with chronic obstructive pulmonary disease]

- **12% did not report the modelling method**
 - e.g., logistic/cox regression
- **64% did not describe how missing data were handled**
- **70% did not report the model**
 - e.g., full regression equation/code (no model → no prediction)
- **78% did not evaluate assess calibration**
 - e.g., no calibration plot
- **24% did not evaluate model discrimination (e.g., AUROC)**

RESEARCH

OPEN ACCESS

Check for updates

Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal

Vanesa Bellou,^{1,2} Lazaros Belbasis,¹ Athanasios K Konstantinidis,² Ioanna Tzoulaki,^{1,3,4} Evangelos Evangelou^{1,3}

¹Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece
²Department of Respiratory Medicine, University Hospital of Ioannina, University of Ioannina Medical School, Ioannina, Greece
³Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
⁴MRC-PHE Center for Environment, School of Public Health, Imperial College London, London, UK

Correspondence to: E Evangelou
evangelis@uoi.gr
(or greevangelou on Twitter; ORCID 0000-0002-5488-2999)

ABSTRACT
OBJECTIVE
To map and assess prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease (COPD).
DESIGN
Systematic review.
DATA SOURCES
PubMed until November 2018 and hand searched references from eligible articles.
ELIGIBILITY CRITERIA FOR STUDY SELECTION
Studies developing, validating, or updating a prediction model in COPD patients and focusing on any potential clinical outcome.
RESULTS
The systematic search yielded 228 eligible articles, describing the development of 408 prognostic models, the external validation of 28 models, and

examined the calibration of the developed model. For 286 (70%) models a model presentation was not available, and only 56 (14%) models were presented through the full equation. Model discrimination using the C statistic was available for 311 (76%) models. 38 models were externally validated, but in only 12 of these was the validation performed by a fully independent team. Only seven prognostic models with an overall low risk of bias according to PROBAST were identified. These models were ADO, B-AE-D, B-AE-D-C, extended ADO, updated ADO, updated BODE, and a model developed by Bertens et al. A meta-analysis of C statistics was performed for 12 prognostic models, and the summary estimates ranged from 0.611 to 0.769.
CONCLUSIONS
This study constitutes a detailed mapping and assessment of the prognostic models for outcome

TRIPOD Statement

- **Started in 2010, published in Jan 2015, in 11 journals**
- **Focus on models developed using regression methods**
 - Guidance is relevant for ML but not explicitly covered
- **Explanation document (73 pages) focusses solely on regression**
 - Touches on design, conduct, 'how to' (best practice)
 - Highlights good methodology
 - Highlights bad practice
- **Widely cited / included in journal author instructions**
 - Statement paper >12000 times; E&E paper >5000 times
- **Needs to be tailored to the AI/ML community (TRIPOD+AI)**
 - e.g., examples, terminology, model presentation & availability, fairness, open science, PPI
 - Harmonise the two fields (statistics/machine learning)

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement

Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Douglas G. Altman, DSc; and Karel G.M. Moons, PhD

Prediction models are developed to aid health care providers in estimating the probability or risk that a specific disease or condition is present (diagnostic models) or that a specific event will occur in the future (prognostic models), to inform their decision making. However, the overwhelming evidence shows that the quality of reporting of prediction model studies is poor. Only with full and clear reporting of information on all aspects of a prediction model can risk of bias and potential usefulness of prediction models be adequately assessed. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative developed a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. This article describes how the TRIPOD Statement was developed. An extensive list of items based on a review of the literature was created, which was reduced after a Web-based survey and revised during a 3-day meeting in June

2011 with methodologists, health care professionals, and journal editors. The list was refined during several meetings of the steering group and in e-mail discussions with the wider group of TRIPOD contributors. The resulting TRIPOD Statement is a checklist of 22 items, deemed essential for transparent reporting of a prediction model study. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. The TRIPOD Statement is best used in conjunction with the TRIPOD explanation and elaboration document. To aid the editorial process and readers of prediction model studies, it is recommended that authors include a completed checklist in their submission (also available at www.tripod-statement.org).

Ann Intern Med. 2015;162:55-63. doi:10.7326/M14-0697 www.annals.org
For author affiliations, see end of text.
For contributors to the TRIPOD Statement, see the Appendix (available at www.annals.org).

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accom-

panied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698 www.annals.org
For author affiliations, see end of text.
For members of the TRIPOD Group, see the Appendix.

Do we have a problem with the design, methods, reporting or spin in AI research?...YES



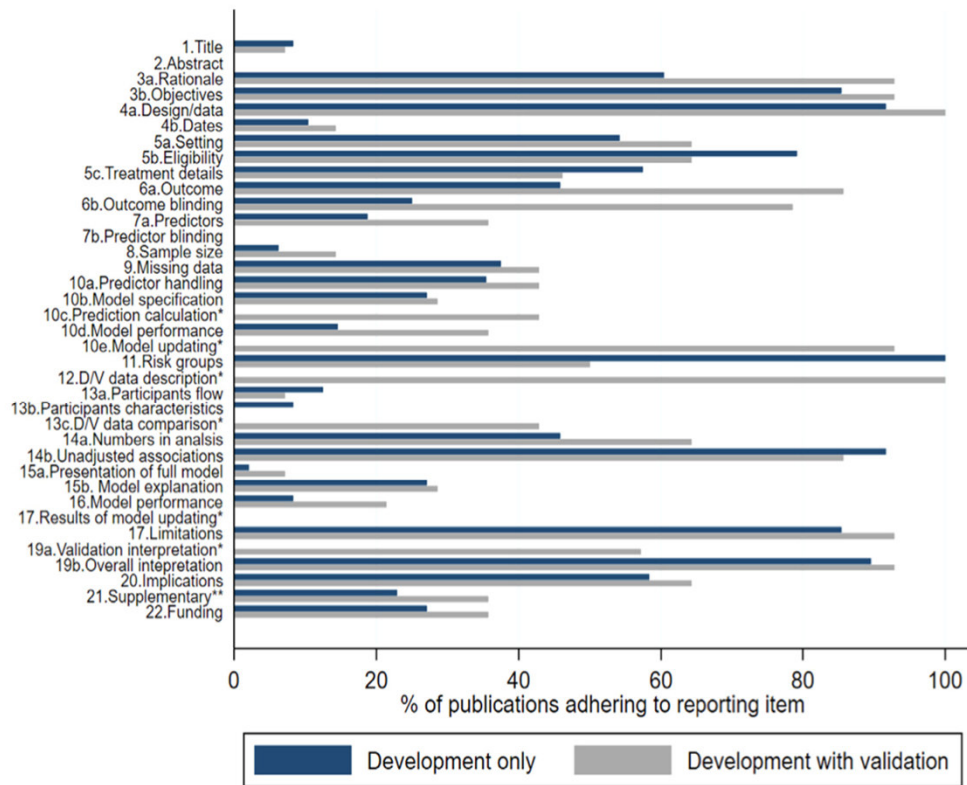
Oxford
(oncology)



Utrecht
(general medical journals)

Artificial intelligence -> Field of computer science that focuses on developing models and algorithms capable of performing tasks that typically require human intelligence.

Adherence to TRIPOD



**Oxford
(oncology)**

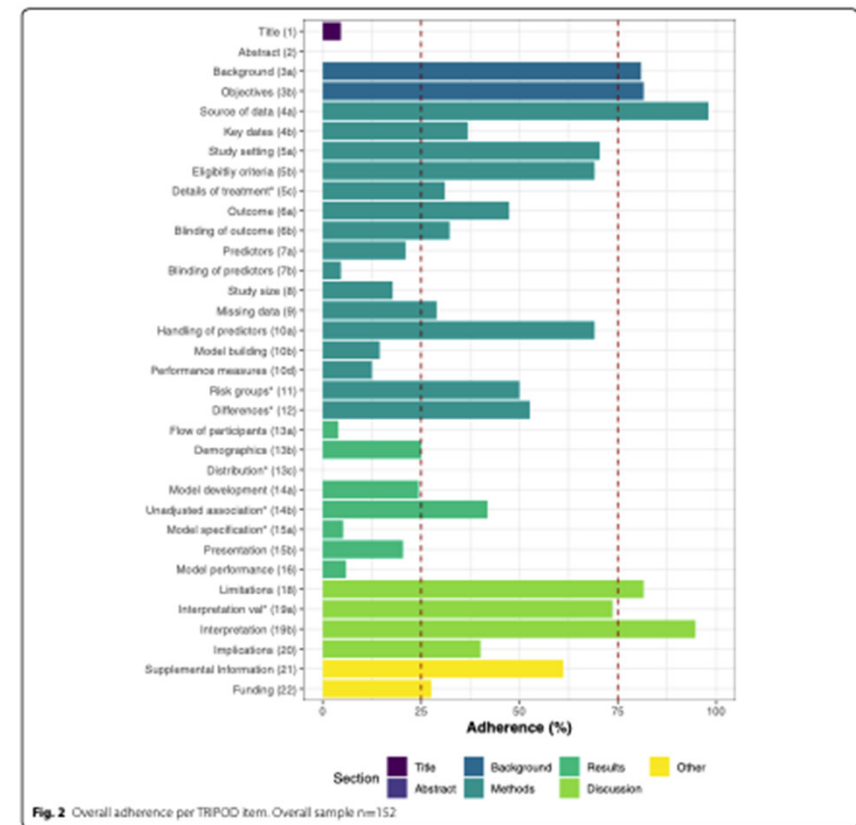


Fig. 2 Overall adherence per TRIPOD item. Overall sample n=152

**Utrecht
(general medical journals)**

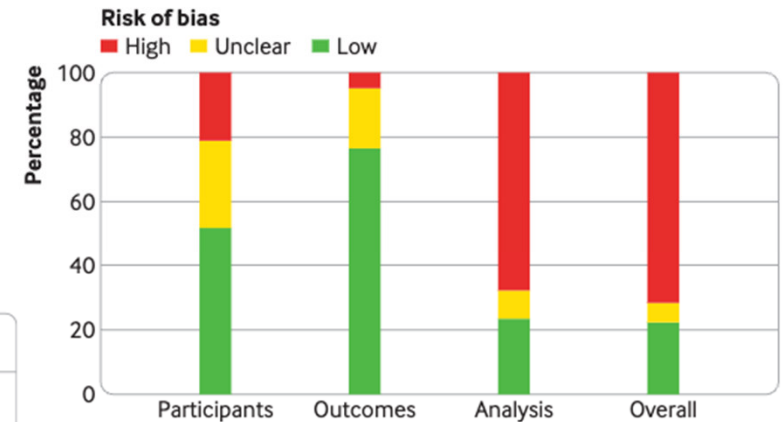
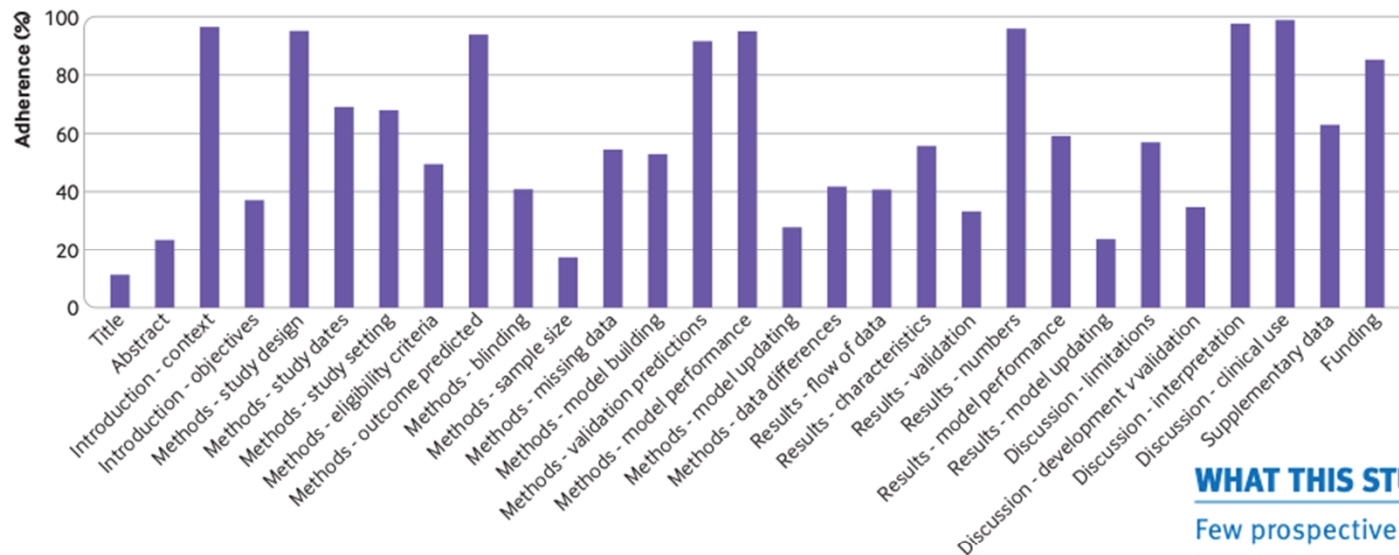
RESEARCH

OPEN ACCESS

Check for updates

Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies

Myura Nagendran,¹ Yang Chen,² Christopher A Lovejoy,³ Anthony C Gordon,^{1,4} Matthieu Komorowski,⁵ Hugh Harvey,⁶ Eric J Topol,⁷ John P A Ioannidis,⁸ Gary S Collins,^{9,10} Mahiben Maruthappu³



WHAT THIS STUDY ADDS

Few prospective deep learning studies and randomised trials exist in medical imaging

Most non-randomised trials are not prospective, are at high risk of bias, and deviate from existing reporting standards

Data and code availability are lacking in most studies, and human comparator groups are often small

Future studies should diminish risk of bias, enhance real world clinical relevance, improve reporting and transparency, and appropriately temper conclusions

COVID-19 prediction models

RESEARCH

OPEN ACCESS

Check for updates

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,^{1,2} Ben Van Calster,^{2,3} Gary S Collins,^{4,5} Richard D Riley,⁶ Georg Heinze,⁷ Ewoud Schuit,^{8,9} Elena Albu,² Banafsheh Arshi,¹ Vanesa Bellou,¹⁰ Marc M J Bonten,^{8,11} Darren L Dahly,^{12,13} Johanna A Damen,^{8,9} Thomas P A Debray,^{8,14} Valentijn M T de Jong,^{8,9} Maarten De Vos,^{2,15} Paula Dhiman,^{4,5} Joie Ensor,⁶ Shan Gao,² Maria C Haller,^{7,16} Michael O Harhay,^{17,18} Liesbet Henckaerts,^{19,20} Pauline Heus,^{8,9} Jeroen Hoogland,⁸ Mohammed Hudda,²¹ Kevin Jenniskens,^{8,9} Michael Kammer,^{7,22} Nina Kreuzberger,²³ Anna Lohmann,²⁴ Brooke Levis,⁶ Kim Luijken,²⁴ Jie Ma,⁵ Glen P Martin,²⁵ David J McLernon,²⁶ Constanza L Andaur Navarro,^{8,9} Johannes B Reitsma,^{8,9} Jamie C Sergeant,^{27,28} Chunhu Shi,²⁹ Nicole Skoetz,²² Luc J M Smits,¹ Kym I E Snell,⁶ Matthew Sperrin,³⁰ René Spijker,^{8,9,31} Ewout W Steyerberg,³ Toshihiko Takada,^{8,32} Ioanna Tzoulaki,^{10,33} Sander M J van Kuijk,³⁴ Bas C T van Bussel,^{1,35} Iwan C C van der Horst,³⁵ Kelly Reeve,³⁶ Florian S van Royen,⁸ Jan Y Verbakel,^{37,38} Christine Wallisch,^{7,39,40} Jack Wilkinson,²⁴ Robert Wolff,⁴¹ Lotty Hoof,^{8,9} Karel G M Moons,^{8,9} Maarten van Smeden⁸

Abstract

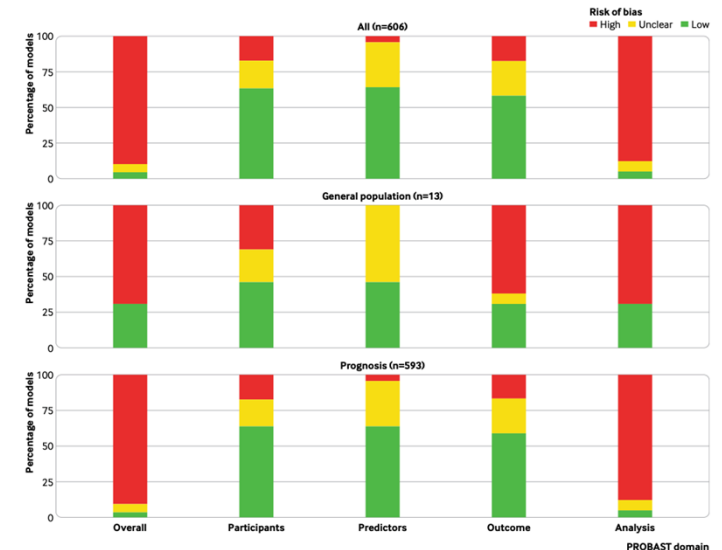
OBJECTIVE

To review and appraise the validity and usefulness of published and preprint reports of prediction models for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of covid-19 infection or being admitted to hospital.

DATA EXTRACTION

At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool).

RESULTS

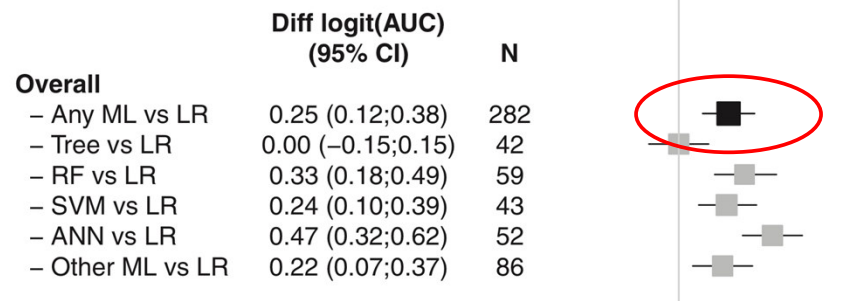


- **606 models** -> “29 had low risk of bias, 32 had unclear risk of bias, and **545 had high risk of bias**”
- “Most of the 606 models were **appraised to have high or uncertain risk of bias** owing to a combination of **poor reporting** and **poor methodological conduct**”

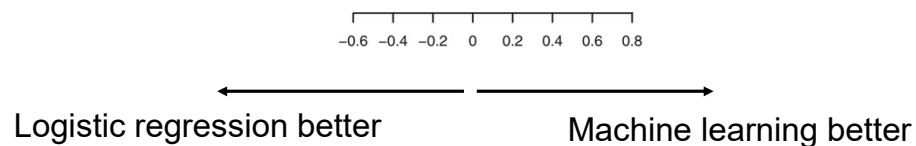
For numbered affiliations see end of the article

Correspondence to: L Wynants laure.wynants@maastrichtuniversity.nl (ORCID 0000-0002-3037-122X) Additional material is published online only. To view please visit the journal online.

Why methods & transparency matters: risk of bias ('off the shelf' ML)



- Complete and transparent reporting aids risk of bias assessment
 - Were the design/methods robust?
 - Need authors to transparently tell readers all the key details
- Impacts on how we interpret study findings and conclusions



Journal of Clinical Epidemiology 110 (2019) 12–22



REVIEW

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou^a, Jie Ma^b, Gary S. Collins^{a,c}, Ewout W. Steyerberg^d, Jan Y. Verbakel^{a,e,f}, Ben Van Calster^{a,d,g}

^aDepartment of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium
^bCentre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK

^cOxford University Hospitals NHS Foundation Trust, Oxford, UK

^dDepartment of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands

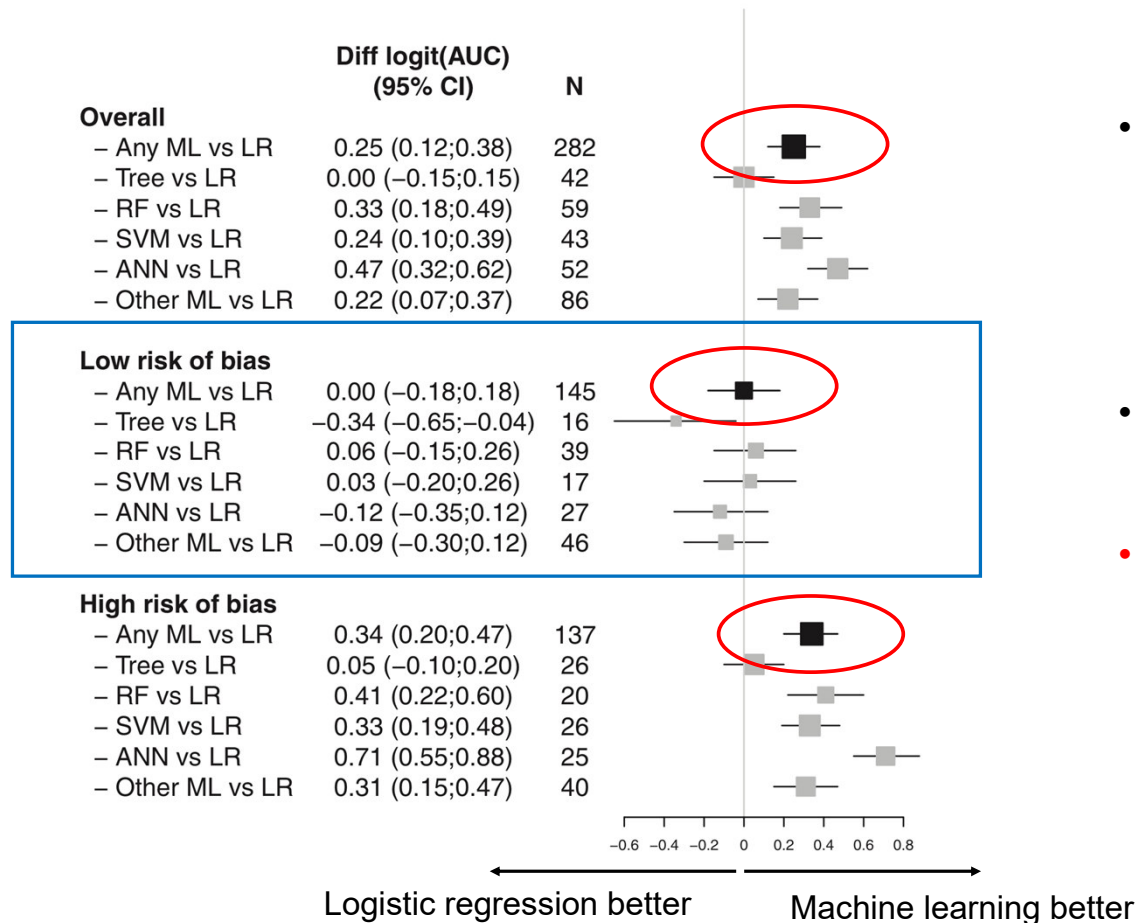
^eDepartment of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33 box 7001, Leuven, 3000 Belgium

^fNuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK

Accepted 5 February 2019; Published online 11 February 2019

Christodoulou et al J Clin Epidemiol 2019

Why methods & transparency matters: risk of bias ('off the shelf' ML)

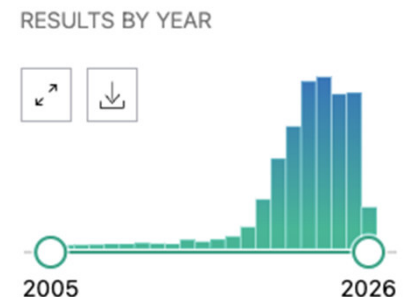


- Complete and transparent reporting aids risk of bias assessment
 - Were the design/methods robust?
 - Need authors to transparently tell readers all the key details
- Impacts on how we interpret study findings and conclusions
- (unfortunately) hype sells
 - Not good for patients
 - Need good design/robust methods & transparency for trustworthy research

ModelMania: e.g., using the SEER data

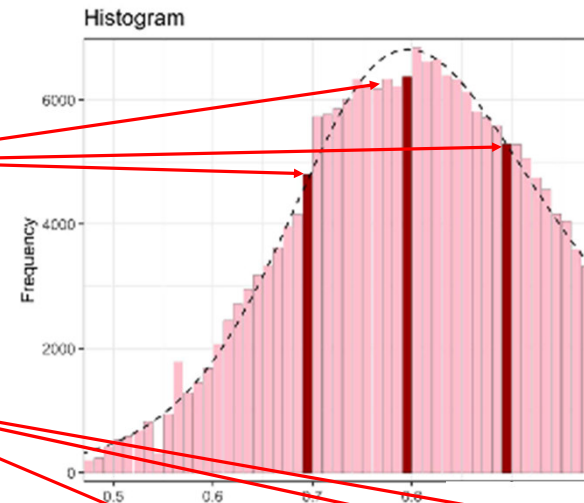
- **SEER is a population-based cancer registry from the US**
- **>2000 papers (indexed on PubMed) developing/evaluating a cancer prediction model using SEER**
- **524 papers published in 2025**
 - ~2600 papers in the last 5 years
- **‘Fast-churn’ research concerns (Tsegaye et al, metaRxiv 2026)**
 - Studies predominantly carried out by researchers outside the US
 - Not reported: intended setting, country
 - Not reported: clinical decision-making role
 - Not reported: suitability of using SEER
 - External validation in (assumed) target population/country rare

(risk score*[tiab] OR nomogram*[tiab] OR prediction model*[tiab] OR prognostic model*[tiab] OR predictive model*[tiab]) AND SEER[tiab] AND 2024[dp]

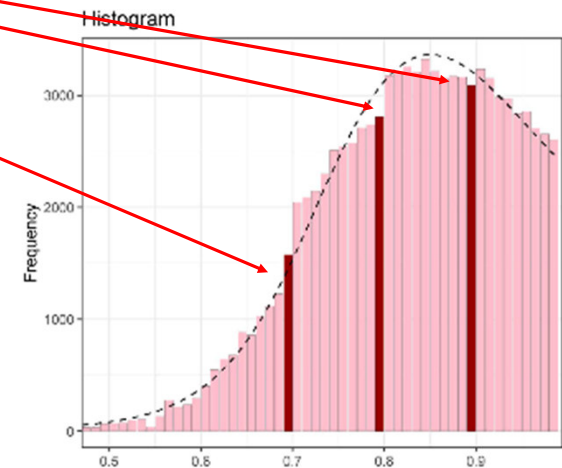


Questionable research practices

- The distribution of 306,888 AUC values (from ~97k abstracts on PubMed)
 - Clear excesses above the thresholds of 0.7, 0.8 and 0.9 and shortfalls below the thresholds
- Evidence (or suggestive) of AUC hacking?
- Emphasising the need for registration, protocols, and clear and transparent reporting



Largest AUCs



Study protocols

- Compared to the systematic reviews of final results publications of prediction model studies that have highlighted poor reporting and methodological weaknesses in studies
- We observed generally good reporting and the conduct of protocols describing plans for prediction model development or validation



Journal of Clinical Epidemiology 169 (2024) 111287

Journal of
Clinical
Epidemiology

ORIGINAL ARTICLE

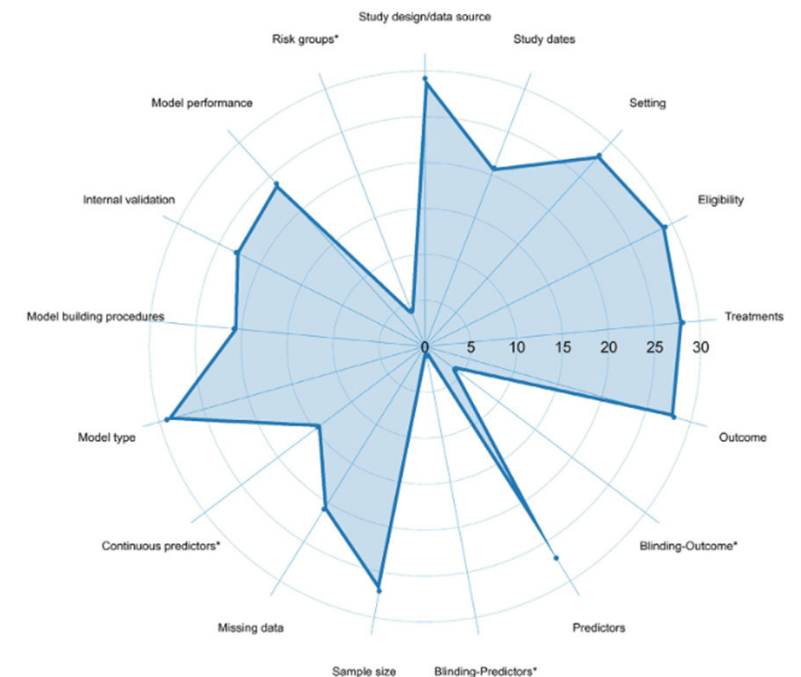
Prediction model protocols indicate better adherence to recommended guidelines for study conduct and reporting

Paula Dhiman^{a,*}, Jie Ma^a, Shona Kirtley^a, Elizabeth Mouka^{a,1}, Caitlin M. Waldron^{a,1}, Rebecca Whittle^{a,b}, Gary S. Collins^a

^aNuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, University of Oxford, Oxford, OX3 7LD, UK

^bNIHR Blood and Transplant Research Unit in Data Driven Transfusion Practice, Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

Accepted 15 February 2024; Published online 20 February 2024



Open science practices

- Increasing expectation to adhere to open science principles*
 - **Protocol** and **study registration** rare
 - Yet an expectation for clinical trials
 - Journals are increasingly requiring **analytical code sharing or statements** (e.g., BMJ [from May 2024])
 - Code to implement models uncommon
 - Hampers independent evaluation (Van Calster et al JAMIA 2019)
 - **Data sharing** statements are often expected
 - ...and should go beyond 'available upon reasonable request'
 - Current reality...data is rarely shared

* or give an explicit and meaningful justification for not adhering to open science (e.g., ethical/legal reasons, proprietary)

Table 3. Summary of studies adhering to open science principles: research practices ($n = 46$)

Open science practice	Frequency	% (95 CI)
Data sharing statement	35	76% (61–87%)
Available upon request	21	46% (31–61%)
Explicitly not shared	6	13% (5–26%)
Links to a website (e.g., SEER)	3	7% (1–18%)
Reported as available in the article but not	2	4% (0–15%)
Available (in supplementary material)	2	4% (0–15%)
'Not applicable'	1	2% (0–12%)
Code sharing statement	12	26% (14–41%)
GitHub	8	17% (8–31%)
Available upon request	2	4% (0–15%)
Other (e.g., supplementary material)	2	4% (0–15%)
Protocol availability	1	2% (0–12%)
Study registration	1	2% (0–12%)
Reporting guideline used	8	17% (8–31%)
MI-CLAIM and CONSORT-AI	1	2% (0–12%)
STARD	1	2% (0–12%)
STROBE	1	2% (0–12%)
TREND	1	2% (0–12%)
TRIPOD	4	9% (2–21%)

Collins et al, J Clin Epidemiol 2024

A published protocol...the peer review

“The study was well designed.

But it was written by a new hand.

First, the authors should offer their results and conclusion sections in the manuscript.

Second, seldom study referred table or figure in the introduction section.

Third, future tense should not be widely used through the manuscript.

I personally suggested the reject of the manuscript.”

Open access

Protocol

BMJ Open Development and validation of clinical prediction models for breast cancer incidence and mortality: a protocol for a dual cohort study

Ashley Kieran Clift ^{1,2} Julia Hippisley-Cox ² David Dodwell ³
Simon Lord,⁴ Mike Brady,⁴ Stavros Petrou ² Gary S. Collins ⁵

Study registration

- clinicaltrials.gov is being used more to register clinical prediction model studies
- Planned studies prioritize developing new models over evaluating existing models
- 27% of study registrations published their planned prediction model study within 10 years of registration
- Nonpublication of clinical prediction model studies is a source of research waste
- Better visibility of planned models is needed to reduce waste from duplication
- Methods to follow-up study outcomes and model uptake can proactively reduce waste



Journal of Clinical Epidemiology 173 (2024) 111433

Journal of
Clinical
Epidemiology

ORIGINAL RESEARCH

Planned but ever published? A retrospective analysis of clinical prediction model studies registered on clinicaltrials.gov since 2000

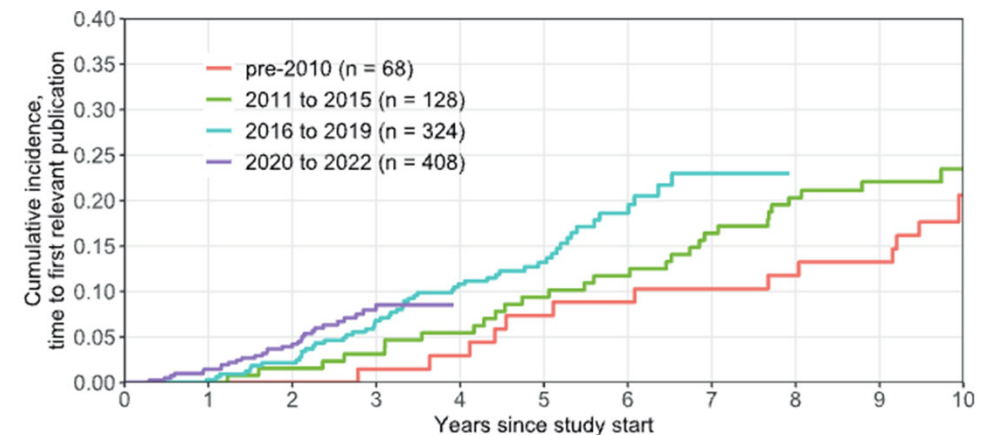
Nicole White^{a,*}, Rex Parsons^a, David Borg^{a,b}, Gary Collins^c, Adrian Barnett^a

^aAustralian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health and Social Work, Queensland University of Technology, Kelvin Grove, Queensland, Australia

^bSchool of Exercise and Nutrition Sciences, Queensland University of Technology, Kelvin Grove, Queensland, Australia

^cCentre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford, United Kingdom

Accepted 12 June 2024; Published online 17 June 2024



Data provenance

- 2 large, publicly available Kaggle datasets, on stroke and diabetes
- Provenance of both datasets could not be verified
 - evidence they are likely to be simulated or fabricated
- 124 clinical prediction model articles
 - evidence of (potential) use in clinical practice
 - 1 model cited in a medical device patent
 - models were cited in 86 review articles
 - 2+ articles have been retracted, others under investigation

Gibson et al, medRxiv 2026

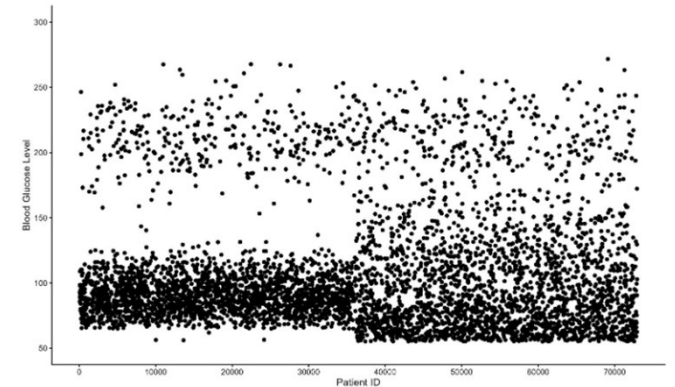


Figure 2. Average blood glucose level plotted against the patient identifier for the stroke dataset. The plot shows a clear shift in the distribution of glucose at half the maximum ID variable.

nature

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [news](#) > article

NEWS | 15 April 2026

Dozens of AI disease-prediction models were trained on dubious data

The models are designed to predict someone's risk of diabetes or stroke. A few might already have been used on patients.

By [Mohana Basu](#)

Retraction Note | [Open access](#) | Published: 07 April 2026

Retraction Note: A comprehensive explainable AI approach for enhancing transparency and interpretability in stroke prediction

[Marwa El-Geneedy](#) [✉](#), [Hossam El-Din Moustafa](#), [Hatem Khater](#), [Seham Abd-Elsamee](#) & [Samah A. Gamel](#)

Scientific Reports **16**, Article number: 11622 (2026) | [Cite this article](#)

1083 Accesses | [Metrics](#)

[i](#) The [Original Article](#) was published on 18 July 2025

Retraction of: *Scientific Reports* <https://doi.org/10.1038/s41598-025-11263-9>, published online 18 July 2025

The Editors have retracted this Article.

The Authors have used a publicly available dataset (The Stroke Prediction Dataset). Concerns have been raised regarding the provenance and validity of this dataset¹. The Authors have been unable to provide any further information regarding the provenance or accuracy of the data included in this dataset. As this dataset was used for the training and validation of the proposed model in this Article, the Editors no longer have confidence in the reliability of the results and conclusions of this Article.

Hossam El-Din Moustafa and Samah A. Gamel agree with the retraction. Marwa El-Geneedy, Hatem Khater, and Seham Abd-Elsamee disagree with the retraction.

Retraction Note | [Open access](#) | Published: 31 March 2026

Retraction Note: An intelligent learning system based on electronic health records for unbiased stroke prediction

[Muhammad Asim Saleem](#), [Ashir Javeed](#), [Wasan Akarathanawat](#), [Aurauma Chutinet](#), [Nijasri Charnnarong Suwanwela](#), [Pasu Kaewplung](#) [✉](#), [Surachai Chaitusaney](#), [Sunchai Deelertpaiboon](#), [Wattanasak Srisiri](#) & [Watit Benjapolakul](#) [✉](#)

Scientific Reports **16**, Article number: 10546 (2026) | [Cite this article](#)

1253 Accesses | 1 Altmetric | [Metrics](#)

[i](#) The [Original Article](#) was published on 04 October 2024

Retraction of: *Scientific Reports* <https://doi.org/10.1038/s41598-024-73570-x>, published online 04 October 2024

The Editors have retracted this Article.

The Authors have used a publicly available dataset (The Stroke Prediction Dataset). Concerns have been raised regarding the provenance and validity of this dataset¹. The Authors have been unable to provide any further information regarding the provenance or accuracy of the data included in this dataset. As this dataset was used for the training and validation of the proposed model in this Article, the Editors no longer have confidence in the reliability of the results and conclusions of this Article.

Watit Benjapolakul and Muhammad Asim Saleem disagree with the retraction. None of the remaining Authors have responded to correspondence from the Editors about the retraction.

Retraction Note | [Open access](#) | Published: 13 April 2026

Retraction Note: Predictive modelling and identification of key risk factors for stroke using machine learning

[Ahmad Hassan](#), [Saima Gulzar Ahmad](#), [Ehsan Ullah Munir](#), [Imtiaz Ali Khan](#) & [Naeem Ramzan](#) [✉](#)

Scientific Reports **16**, Article number: 12197 (2026) | [Cite this article](#)

1187 Accesses | 5 Altmetric | [Metrics](#)

[i](#) The [Original Article](#) was published on 20 May 2024

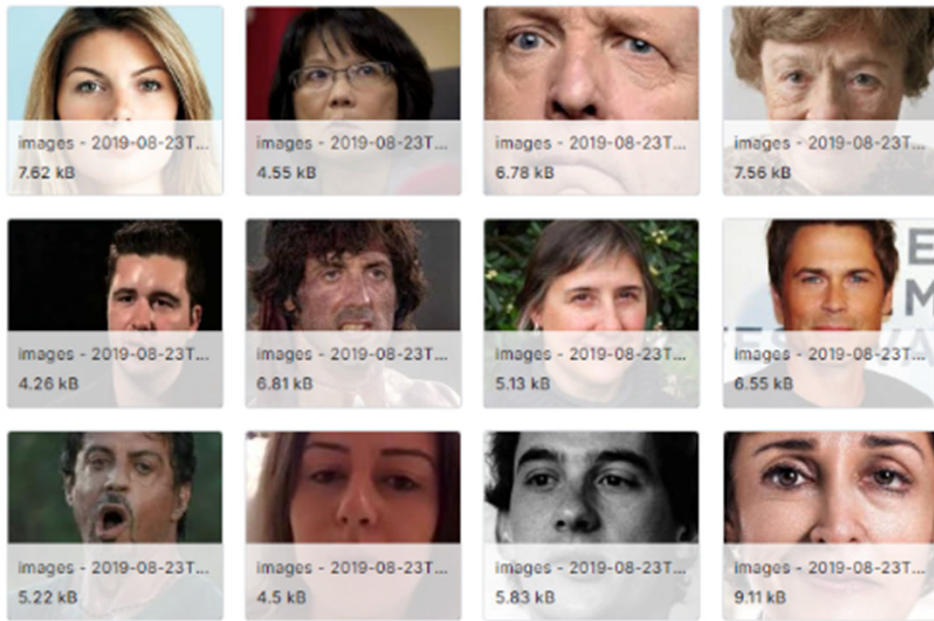
Retraction of: *Scientific Reports* <https://doi.org/10.1038/s41598-024-61665-4>, published online 20 May 2024

The Editors have retracted this Article.

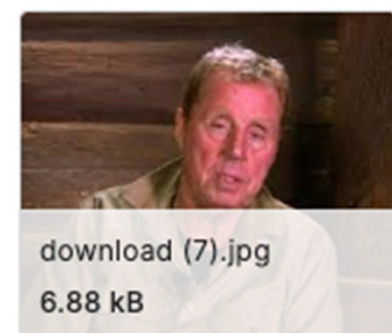
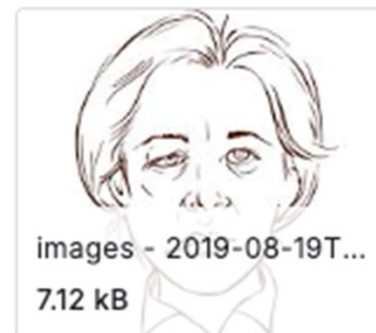
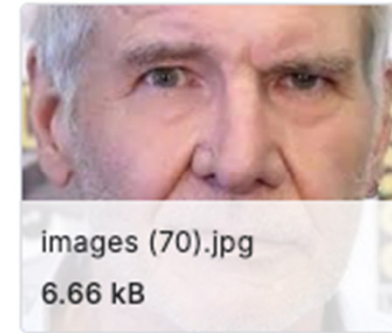
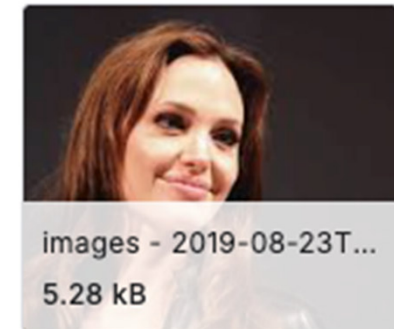
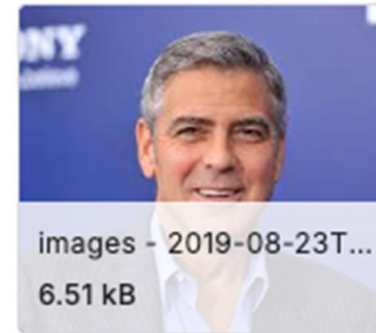
The Authors have used a publicly available dataset (The Stroke Prediction Dataset). Concerns have been raised regarding the provenance and validity of this dataset [1]. The Authors have been unable to provide any further information regarding the provenance or accuracy of the data included in this dataset. As this dataset was used to train and assess the machine-learning models examined in this Article, the Editors no longer have confidence in the reliability of the results and conclusions of this Article.

All Authors disagree with the retraction.

'Comically bad' datasets used to train clinical models for stroke and diabetes



A dataset on Kaggle purportedly showing people who have had a stroke includes images of Sylvester Stallone from *Rambo* and other celebrities. [Source](#)



retractionwatch.com/2026/05/18/kaggle-dataset-clinical-models-stroke-diabetes/

RETRACTED ARTICLE | [Open access](#) | Published: 27 December 2025

RETRACTED ARTICLE: Real-time monitoring system for early stroke detection based on fog computing and enhanced deep learning techniques

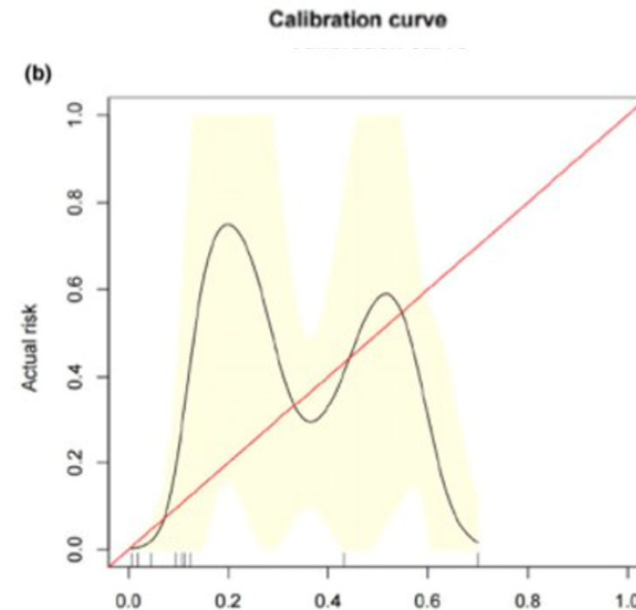
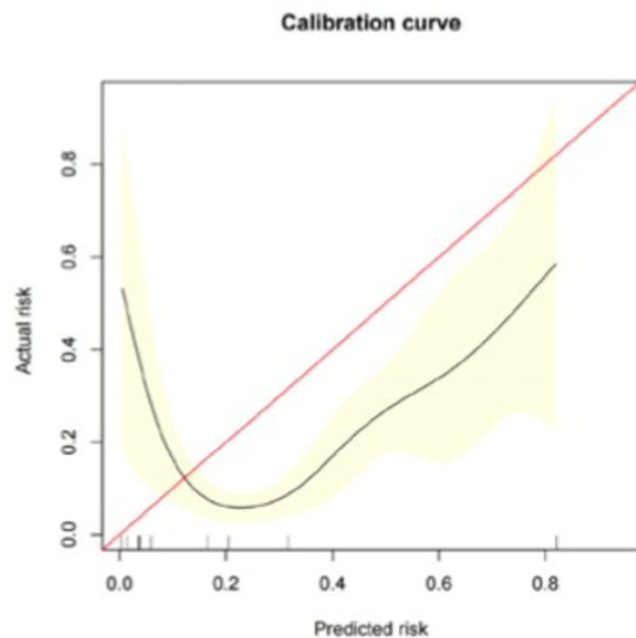
[Alaa M. Mohamed](#) [✉](#), [Hanan M. Amer](#), [Asmaa H. Rabie](#), [Ahmed I. Saleh](#) & [Mohy-Eldin A. Abo-Elsoud](#)[Scientific Reports](#) **15**, Article number: 44671 (2025) | [Cite this article](#)[Save article](#)4643 Accesses | 2 Citations | 13 Altmetric | [Metrics](#) This article has been [updated](#)

The Editors have retracted this Article.

After publication, concerns were raised regarding the provenance and reliability of the public datasets used in the study. The dataset “face images of acute stroke and non-acute stroke” can no longer be verified through its original source. The dataset “facial droop and facial paralysis image” [1] appears to contain identifiable facial images; however, information regarding ethical oversight of data collection or an associated data governance framework could not be confirmed, nor is it clear that clinical diagnoses were validated. Since these datasets were used for the training and validation of the proposed model, the Editors no longer have confidence in the contents of the Article. Since the consent procedures for inclusion of identifiable information in the dataset are unclear, the content of this Article has been removed to protect the privacy of the individuals.

Hanan M. Amer, Asmaa H. Rabie, and Mohy-Eldin A. Abo-Elsoud disagree with the retraction. Alaa M. Mohamed and Ahmed I. Saleh did not explicitly state whether or not they agree with this retraction.

Overinterpretation ('spin')



“The calibration curve showed a good agreement between the predictive risk and the actual probability”



TRIPOD+AI is an international initiative to improve the completeness and transparency of reporting in studies developing clinical prediction models involving artificial intelligence driven by machine learning (and regression)

RESEARCH METHODS AND REPORTING

OPEN ACCESS



TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods

Gary S Collins,¹ Karel G M Moons,² Paula Dhiman,¹ Richard D Riley,^{3,4} Andrew L Beam,⁵ Ben Van Calster,^{6,7} Marzyeh Ghassemi,⁸ Xiaoxuan Liu,^{9,10} Johannes B Reitsma,² Maarten van Smeden,² Anne-Laure Boulesteix,¹¹ Jennifer Catherine Camaradou,^{12,13} Leo Anthony Celi,^{14,15,16} Spiros Denaxas,^{17,18} Alastair K Denniston,^{4,9} Ben Glocker,¹⁹ Robert M Golub,²⁰ Hugh Harvey,²¹ Georg Heinze,²² Michael M Hoffman,^{23,24,25,26} André Pascal Kengne,²⁷ Emily Lam,¹² Naomi Lee,²⁸ Elizabeth W Loder,^{29,30} Lena Maier-Hein,³¹ Bilal A Mateen,^{17,32,33} Melissa D McCradden,^{34,35} Lauren Oakden-Rayner,³⁶ Johan Ordish,³⁷ Richard Parnell,¹² Sherri Rose,³⁸ Karandeep Singh,³⁹ Laure Wynants,⁴⁰ Patricia Logullo¹

For numbered affiliations see end of the article.
Correspondence to: G S Collins
gary.collins@csm.ox.ac.uk
(or @GSCollins on Twitter;
ORCID 0000-0002-2772-2316)
Additional material is published online only. To view please visit the journal online.
Cite this as: *BMJ* 2024;385:e078378
<http://dx.doi.org/10.1136/bmj-2023-078378>
Accepted: 17 January 2024

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement was published in 2015 to provide the minimum reporting recommendations for studies developing or evaluating the performance of a prediction model. Methodological advances in the field of whether regression modelling or machine learning methods have been used. The new checklist supersedes the TRIPOD 2015 checklist, which should no longer be used. This article describes the development of TRIPOD+AI and presents the expanded 27 item checklist with more detailed explanation of each reporting

- Supplementary material includes an Explanation & Elaboration 'light' with bullet points to guide reporting
- Longer Explanation & Elaboration paper currently being written with detailed guidance/education (to appear in 2026)

Developing TRAPOD+AI

- Followed guidance set out by the **EQUATOR Network** (Moher et al PLoS Med 2010)
 - (informed by on-going work [at the time] developing recommendations for consensus-based methods – the ACCORD statement, Gattrell et al, PLoS Med 2024)
- Over **200 international experts** participated in the Delphi survey
 - >27 countries covering six continents
- **28 experts** participated in a **consensus meeting** (held online) in July 2022
- Researchers (statisticians/data scientists, epidemiologists, machine learning researchers/scientists, clinicians, radiologists, and ethicists), healthcare professionals, journal editors, funders, policymakers, healthcare regulators, patients, and the general public

Section/Topic	Item	Development / evaluation ¹	Checklist item	Reported on page
TITLE				
<i>Title</i>	1	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				
<i>Abstract</i>	2	D;E	See TRIPOD+AI for Abstracts checklist	
INTRODUCTION				
<i>Background</i>	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	
	3b	D;E	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)	
	3c	D;E	Describe any known health inequalities between sociodemographic groups	
<i>Objectives</i>	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	
METHODS				
<i>Data</i>	5a	D;E	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	
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	
<i>Participants</i>	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	
	6b	D;E	Describe the eligibility criteria for study participants	
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	
<i>Data preparation</i>	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	
<i>Outcome</i>	8a	D;E	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	
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	
<i>Predictors</i>	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and	



- New checklist of reporting **recommendations which are agnostic to modelling approach** to cover prediction model studies using **any regression or machine learning** method*
- **Harmonisation of nomenclature** between regression and machine learning communities
- The new **TRIPOD+AI checklist supersedes the TRIPOD-2015** checklist, which should no longer be used (explanatory/explanation paper still useful; updated version currently in preparation)
- Particular **emphasis on ‘fairness’** to raise awareness and ensure reports mention whether specific methods were used to address fairness. Aspects of **fairness are embedded throughout** the checklist, e.g.,
 - **Diverse** and **representative data** (STANDING Together, Lancet Digital Health)
 - **Performance evaluated in key subgroups** (e.g., defined by personal, social or clinical attributes)

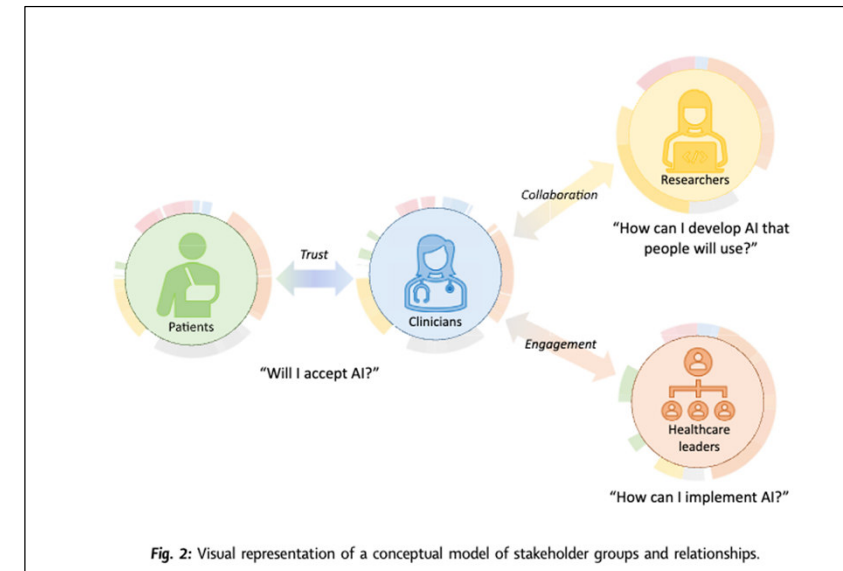
* does not cover generative AI, but TRIPOD-LLM now available (Gallifant et al, Nat Med 2025);
Interactive checklist available on tripod-llm.vercel.app

TRAPOD+AI

- Asking for the **clinical decision** the model is **intended to support**
 - Why is the model needed?
- Clear **description of the provenance of the data** being used
 - **Rationale, richness** and **representativeness**
 - Data quality and handling of any **missing data**
 - strategies for **missingness at deployment**
 - How the data are being used to train and evaluate/test the model
 - **Sample size** considerations (for both training and testing)
- Rationale for the modelling approach (and details) including description of any tuning processes
- Modification of the 'model performance' item recommending authors **evaluate model performance in key subgroups** (e.g., defined by personal, social or clinical attributes)
- How to use the prediction model
 - Any **restrictions on use** (i.e., freely available, proprietary)

TRAPOD+AI

- Inclusion of a new item on ‘**patient and public involvement**’ (PPI)
- Raising awareness and prompting authors to provide details on any PPI during the design, conduct, reporting (and interpretation) or dissemination of the study
- Increasingly expected in healthcare research
 - Often a requirement for funding
 - Some journals (e.g., BMJ) require an explicit PPI statement
- If there was no PPI in any aspect, then clearly state so



(Kuo et al, eClinicalMedicine, 2024)



Patient and public involvement

A group of patient partners was engaged during the design phase to provide feedback on prediction time horizons of interest, presentation of both risk predictions simultaneously [kidney failure and death], and how to visualise them (KDpredict app and figures of this report). A qualitative study is underway on how patients, care givers, and providers understand risk.

Liu et al, BMJ 2024



- Inclusion of a new **'open science'** section with sub-items on
 - Funding (and role of funder in the research)
 - Conflicts of interest
 - Study registration
 - Study protocols
 - TRIPOD-P in preparation
 - Data availability
 - Acknowledging not all data can be shared
 - Code availability (analytical code and model code)
 - Acknowledging difficulties in this area (e.g., proprietary issues)
 - Any conditions/licences/hardware requirements
 - TRIPOD-Code in preparation
 - Materials that are unable to be shared should be also noted, with reasons

Expanded guidance



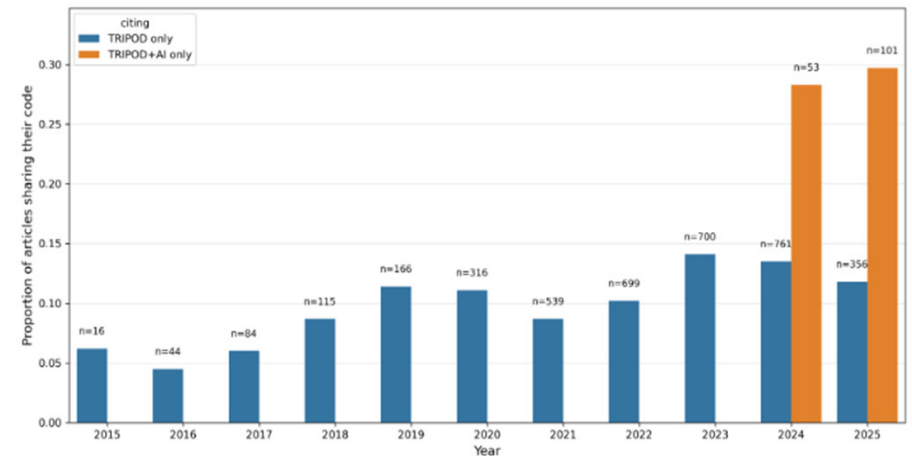
Version: 7-February-2024

Section/Topic	Item	Checklist item
	9c D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors - For predictors that require a subjective interpretation (e.g., interpreting the results from an imaging test), the qualifications and demographic characteristics of the predictor assessors should be reported - If the measurement and interpretation require (additional) training or specific instructions, then these should be reported. This could be reported in the supplementary material
Sample size	10 D;E	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 - Describe how the sample size was determined – this should be done separately for determining the sample size needed for model development and the sample size needed to evaluate the performance of the model irrespective of whether data are being prospectively collected or using existing data - Provide details and all estimates used in any sample size calculation - If no formal sample size calculation was done, e.g., all available data were used, provide a justification whether the size of the data was sufficient to answer the research question
Missing data	11 D;E	Describe how missing data were handled. Provide reasons for omitting any data - Missing data is an omnipresent problem. Authors should report for each predictor being considered for inclusion in the model the number of missing values - The handling of missing values should be reported, including any assumptions for the reason of the missingness - If individuals (or predictors) have been omitted due to the missing values, this should be reported, and reasons given - If missing values have been imputed, then full details of the method for imputing any missing values should be reported - If missing values have been imputed confirm it was done separately for the training and any test data (i.e., avoiding leakage)
Analytical methods	12a D	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 - Describe how the available data were used to develop the model and to evaluate model performance, including whether and how the data were partitioned, and the reasons for partitioning the data (e.g., model development, hyperparameter tuning, evaluating model performance, internal-external cross-validation) - If the data has been partitioned, report whether sample size requirements (see item 10) were considered during the partitioning, and whether the size of the partitioned data are sufficient to carry out the analyses and answer the research question - If the data has been partitioned into training (including any hyperparameter tuning data) and test data, confirm that there has been no data leakage

Code sharing: TRIPOD Code

- Among 3,967 eligible articles, 12.2% (n=482) included code sharing statements*
- Code sharing increased over time, reaching 15.8% in 2025, and was higher among TRIPOD+AI citing studies than TRIPOD citing studies (29.7% vs. 11.8% in 2025, excluding articles citing both)
- Sharing prevalence varied widely by journal and country, with higher sharing observed in journals that discuss code availability in their submission guidelines.
- Next steps: TRIPOD-Code; guidance for
 - reporting code sharing in papers
 - preparing code repositories
 - Information often missing, e.g., README files, license, instructions on use

Sounack et al, Nat Med 2026 (to appear)



Pollard et al. *Diagnostic and Prognostic Research* (2026) 10:4
<https://doi.org/10.1186/s41512-025-00217-4>

Diagnostic and
Prognostic Research

PROTOCOL

Open Access

Protocol for development of a reporting guideline (TRIPOD-Code) for code repositories associated with diagnostic and prognostic prediction model studies

Tom Pollard^{1,2*}, Thomas Sounack^{3†}, Catherine A. Gao⁴, Leo Anthony Celi^{1,2}, Charlotta Lindvall^{3,5}, Hyeonhoon Lee^{6,7}, Hyung-Chul Lee^{6,7}, Karel G. M. Moons⁸ and Gary S. Collins^{9,10}

Incentive? Completeness and transparency of reporting

RESEARCH ARTICLE

Is Quality and Completeness of Reporting of Systematic Reviews and Meta-Analyses Published in High Impact Radiology Journals Associated with Citation Rates?

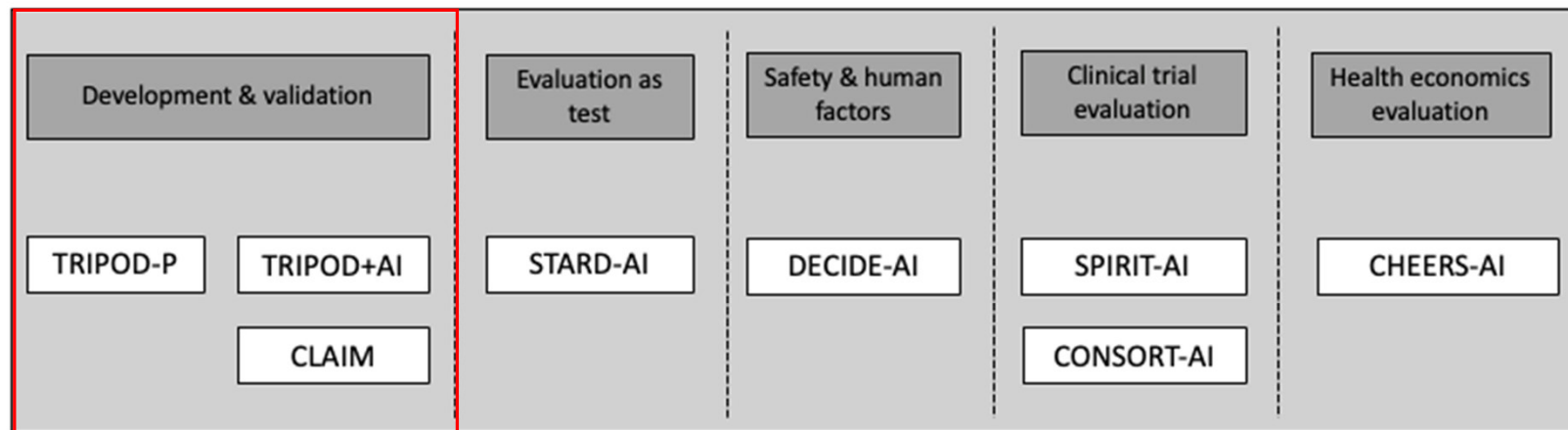
Christian B. van der Pol¹, Matthew D. F. McInnes^{1,2*}, William Petrcich², Adam S. Tunis¹, Ramez Hanna¹

¹ Department of Radiology, University of Ottawa, Ottawa, Ontario, Canada, ² Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

“There is a positive correlation between the quality and the completeness of a reported systematic review or meta-analysis with citation rate which persists when adjusted for journal IF and journal 5-year IF”

Assumption: the better reported a study is, the more likely the findings will be used to improve patient care & outcomes and influence future research

Other reporting guidelines
for different points in the
predictive AI pipeline



Reporting guideline	Phase of AI model development, testing or evaluation
TRIPOD-P	Protocols for AI model development, validation and updating studies (Dhiman et al, Nat Mach Intell 2023)
TRIPOD+AI	Studies describing the development, validation and updating of an AI model (Collins et al, BMJ 2024)
CLAIM-2024	Studies describing the development, validation of a medical imaging AI model (Tejani et al, Radiol AI 2024)
STARD-AI	Studies describing the diagnostic test accuracy of an AI intervention (Sounderajah et al, Nat Med 2025)
DECIDE-AI	Studies describing early stage (safety, human factors) evaluation of an AI intervention (Vasey et al, Nat Med 2023)
SPIRIT-AI	Protocols for the intervention studies evaluating an AI intervention (Rivera et al, BMJ 2020)
CONSORT-AI	Trial reports evaluating the effectiveness of an AI intervention (Liu et al, Nat Med 2020)
CHEERS-AI	Studies describing the health economic evaluation of AI interventions (Elvidge et al, Val Health 2024)

Generative AI: TRIPOD-LLM (Gallifant et al, Nat Med 2025); CHART - chatbots for health advice, (Huo et al, BMJ 2025); TREGAI - ethics for generative AI (Liu et al, arxiv 2013);

Summary

- Trustworthy AI needs to be thoroughly evaluated, followed by complete & accurate reporting
- Lots of evidence that predictive AI research is poorly designed, conducted, reported and interpreted
- The use of tools like TRIPOD+AI, DECIDE-AI, CLAIM-2024, STARD-AI, CONSORT-AI and PROBAST+AI can play a pivotal role to improve
 - reporting transparency
 - design, conduct, reporting and interpretation
 - trust in AI research at various stages in the research pipeline