UNIVERSITY OF BIRMINGHAM

Diagnostic and prognostic prediction model studies using artificial intelligence: the importance of transparency

Gary S. Collins, PhD
Professor of Medical Statistics

03-October-2025

Disclosures

I have no financial interests or relationships to disclose

Research and Publication

- Medical research **should advance scientific knowledge** - directly or indirectly - lead to improvements in treatment or prevention of disease
  - Good research question, design, conduct and fully reported

- 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

- **Good reporting** is an **essential** activity in doing **good research**
  - Open science, reproducibility and research(er) integrity

- Avoiding mis/over-interpretation of study findings (e.g., 'spin'/hype)

Declaration of Helsinki

Methods →

Protocols →

Funding
COIs
Expertise
Benefits →

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.

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.

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.

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.

← **Registration**

← **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

Research waste* from poor reporting

Research: increasing value, reducing waste 5

Reducing waste from incomplete or unusable reports of biomedical research

Paul Glasziou, Douglas G Altman, Patrick Bossuyt, Isabelle Boutron, Mike Clarke, Steven Julious, Susan Michie, David Moher, Elizabeth Wager

- "**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**"

Glasziou et al, Lancet 2014

* Research that has limited or no value

Reporting guidelines

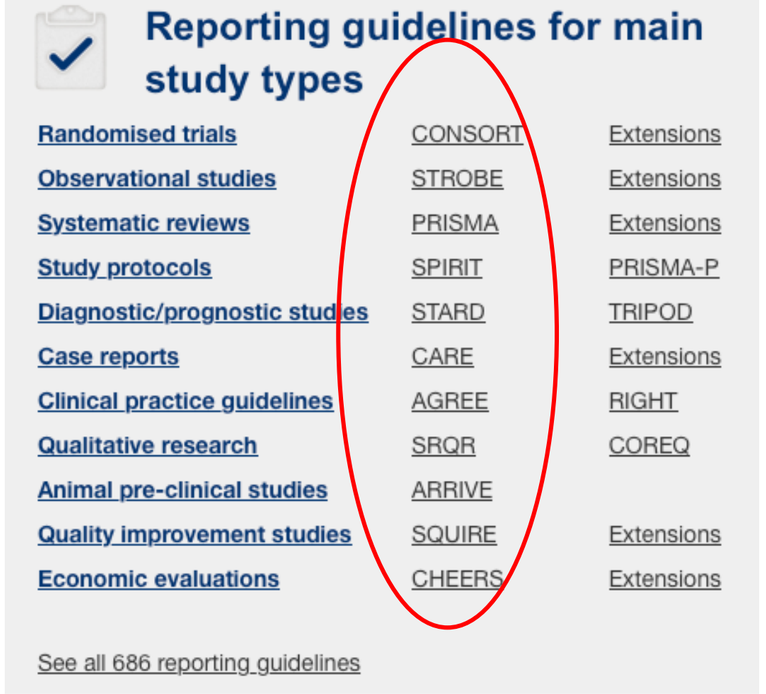
www.equator-network.org

equator network

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.

- They are a **minimum** set of essential items when reporting a study
  - **Reminders** (in the form of a checklist) of **scientific content** for authors
  - Recommendations and guidance, not requirements
- 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



Journal Instructions to authors

How Do I?

Determine My Article Type

JAMA

Categories of Articles

Research

Article Type	Description	Requirements
Original Investigation full info	Clinical trial Meta-analysis Intervention study Cohort study Case-control study Epidemiologic assessment Survey with high response rate Cost-effectiveness analysis Decision analysis Study of screening and diagnostic tests Other observational study	• 3000 words • ≤5 tables and/or figures • Structured abstract • Key Points • Data Sharing Statement • Follow EQUATOR Reporting Guidelines

Statistical issues

thebmj

Reporting guidelines

Reporting guidelines promote clear reporting of methods and results to allow critical appraisal of the manuscript. We ask that all manuscripts be written in accordance with the appropriate reporting guideline. Please submit as supplemental material the appropriate reporting guideline checklist showing on which page of your manuscript each checklist item appears. A complete list of guidelines can be found in the website of the Equator Network. Below is the list of most often used checklists but others may apply.

For a **clinical trials**, use the CONSORT checklist and also include a structured abstract that follows the CONSORT extension for abstract checklist, the CONSORT flowchart and, where applicable, the appropriate CONSORT extension statements (for example, for cluster RCTs, pragmatic trials, etc.). A completed TIDieR checklist is also helpful as this helps to ensure that trial interventions are fully described in ways that are reproducible, usable by other clinicians, and clear enough for systematic reviewers and guideline writers.

For **systematic reviews or meta-analysis** of randomised trials and other evaluation studies, use the PRISMA checklist and flowchart and use the PRISMA structured abstract checklist when writing the structured abstract.

For **studies of diagnostic accuracy**, use the STARD checklist and flowchart.

For **observational studies**, use the STROBE checklist and any appropriate extension STROBE extensions.

For **genetic risk prediction studies,** use GRIPS.

For **economic evaluation studies**, use CHEERS.

For **studies developing, validating or updating a prediction model**, use TRIPOD.

For articles that include explicit statements of the quality of evidence and strength of recommendations, we prefer reporting using the GRADE system.

For studies using data from electronic health records, please use CODE-EHR.

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 subheadings 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.

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

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

Other transparency incentives?

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 Pol[1], Matthew D. F. McInnes[1,2]*, William Petrcich[2], Adam S. Tunis[1], Ramez Hanna[1]

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

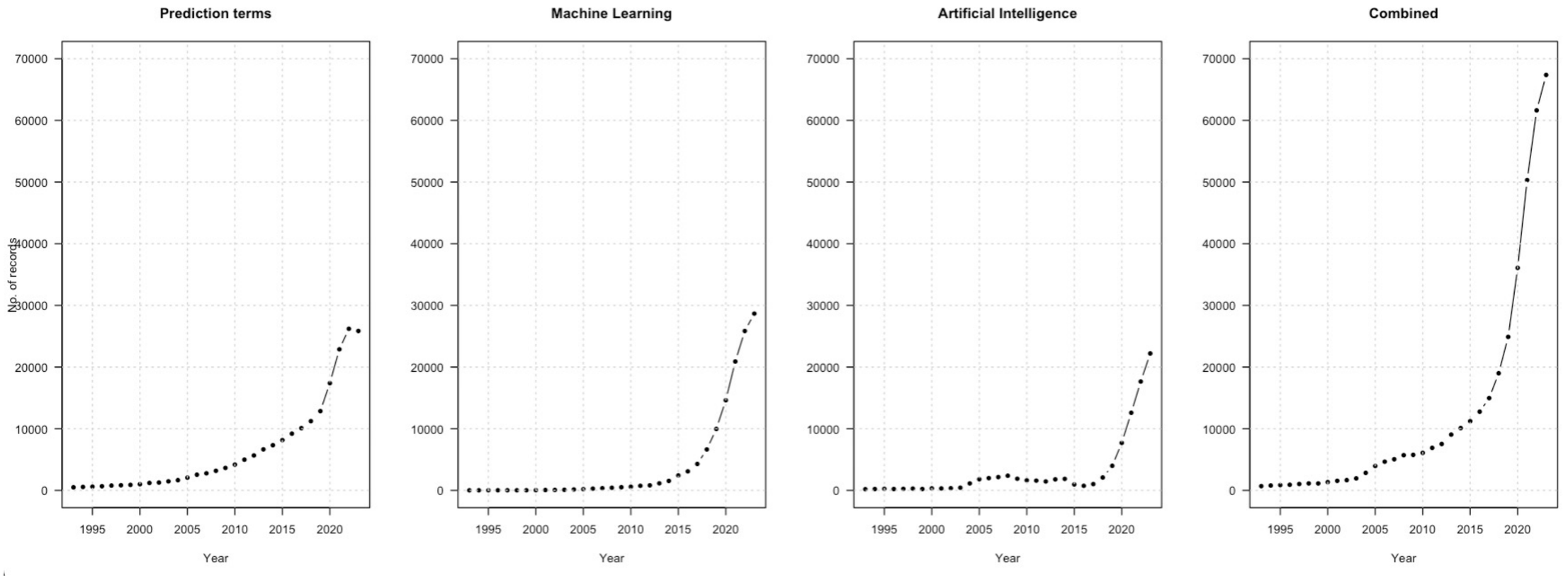
"There is a positive correlation between the quality and the completeness of a reported systematic review or meta-analysis with citation rate which persists when adjusted for journal IF and journal 5-year IF"

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

What is predictive AI?

- Applying machine learning methods to combine patient level information (e.g., demographics, symptoms, biomarkers, PROMs, imaging, omics) to estimate their individualised probability/risk
    - of the presence of a particular health condition (diagnostic)
    - whether a particular outcome will occur in the future (prognostic)
- Starting to see predictive AI being used to estimate PROMs, e.g., HRQoL, fatigue, pain
- Supporting (shared) clinical decision-making, such as whether to refer patients for further testing, monitor disease deterioration or treatment effects, or initiate treatment or lifestyle changes

Predictive AI* is a hot topic



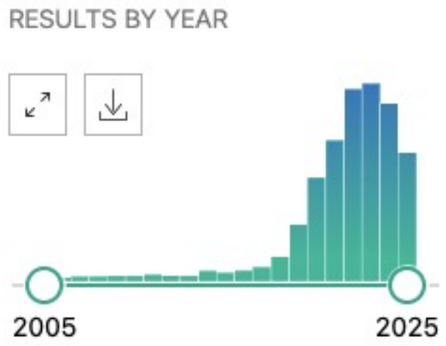
*models that estimate an individual risk/probability to aid diagnosis/prognosis

ModelMania: e.g., predictive AI using the SEER data

- **SEER is a population-based cancer registry from the US**
  - **Covering ~48% of the US population**

- **>2800 papers (indexed on PubMed) developing/validating a cancer prediction model using the SEER data**

- **521 papers published in 2024 (577 in 2023, 562 in 2022, 408 in 2021, 298 in 2020) using the SEER data**
  - 10 papers per week in 2024
  - 373 papers to date in 2025
  - >2300 papers in the last 5 years

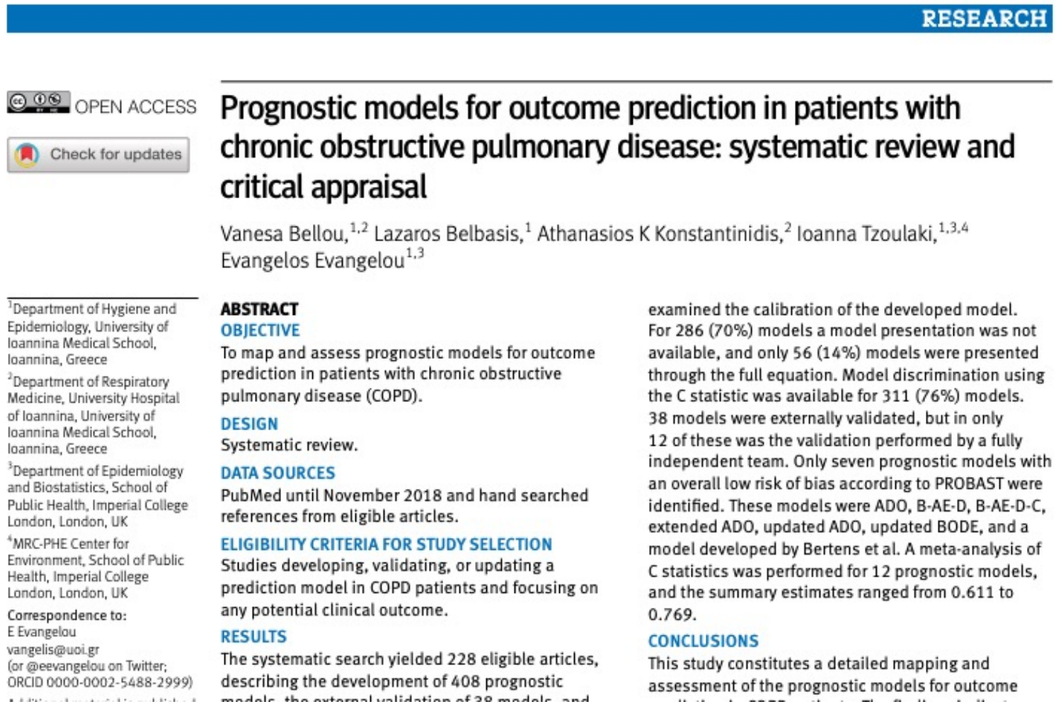


(risk score*[tiab] OR nomogram*[tiab] OR prediction model*[tiab] OR prognostic model*[tiab] OR predictive model*[tiab]) AND SEER[tiab] AND 2024[dp]

Reporting of prediction models: ‘pre-ML/AI’ 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, no estimates of the calibration slope
- **24% did not evaluate model discrimination (e.g., AUC)**



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]

[1]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece
[2]Department of Respiratory Medicine, University Hospital of Ioannina, University of Ioannina Medical School, Ioannina, Greece
[3]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
[4]MRC-PHE Center for Environment, School of Public Health, Imperial College London, London, UK
Correspondence to:
E Evangelou
vangelis@uoi.gr
(or @eevangelou 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 38 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

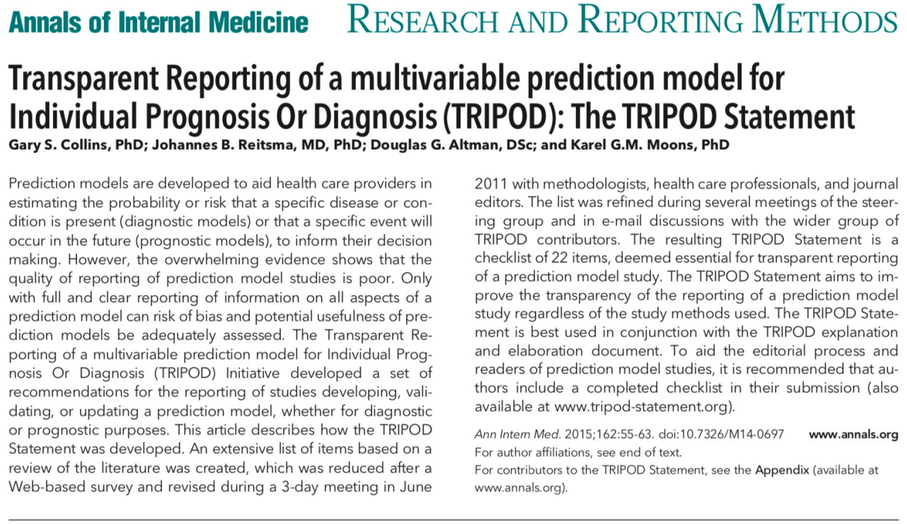
- **Started in 2010, published in Jan 2015, in 11 journals**

- **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 conduct/'how to' (best practice)
  - Discusses common methodological issues / flaws

- **Widely cited / included in journal author instructions**
  - Statement paper >10,000 times; E&E paper >4,000 times

- **Needed tailoring for the AI/ML community (TRIPOD+AI)**
  - e.g., examples, terminology, model presentation & availability, fairness, open science, PPI
  - Harmonise the two fields (statistics / machine learning)



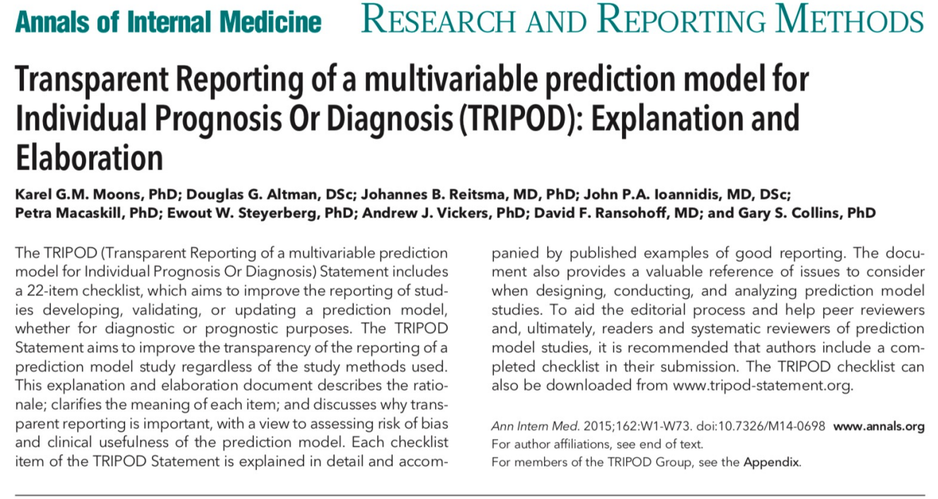
Annals of Internal Medicine RESEARCH AND REPORTING METHODS

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 accompanied 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.

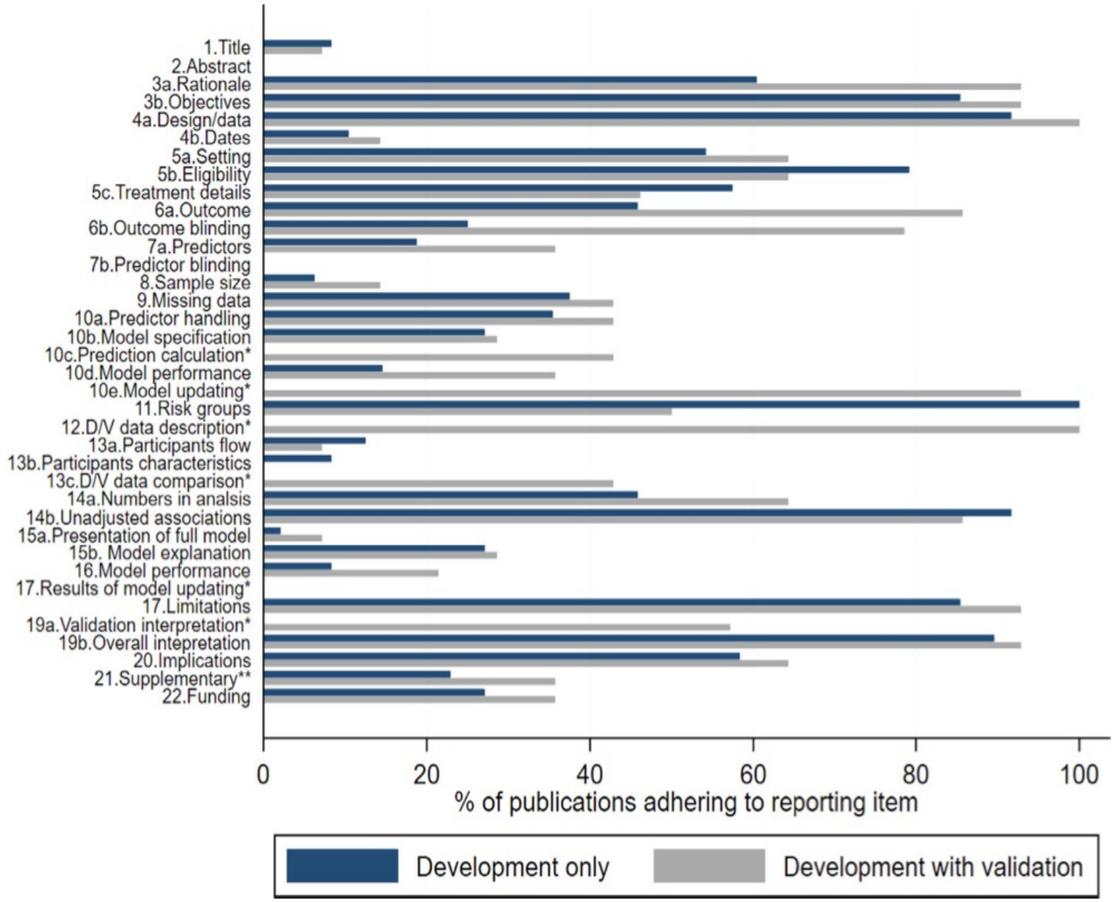
Ann Intern Med. 2015;162:W1-W73. doi:10.7326/M14-0698 **www.annals.org**
For author affiliations, see end of text.
For members of the TRIPOD Group, see the **Appendix**.

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

Oxford (oncology)

Utrecht (general medical journals)

Completeness of reporting



Oxford
(oncology)

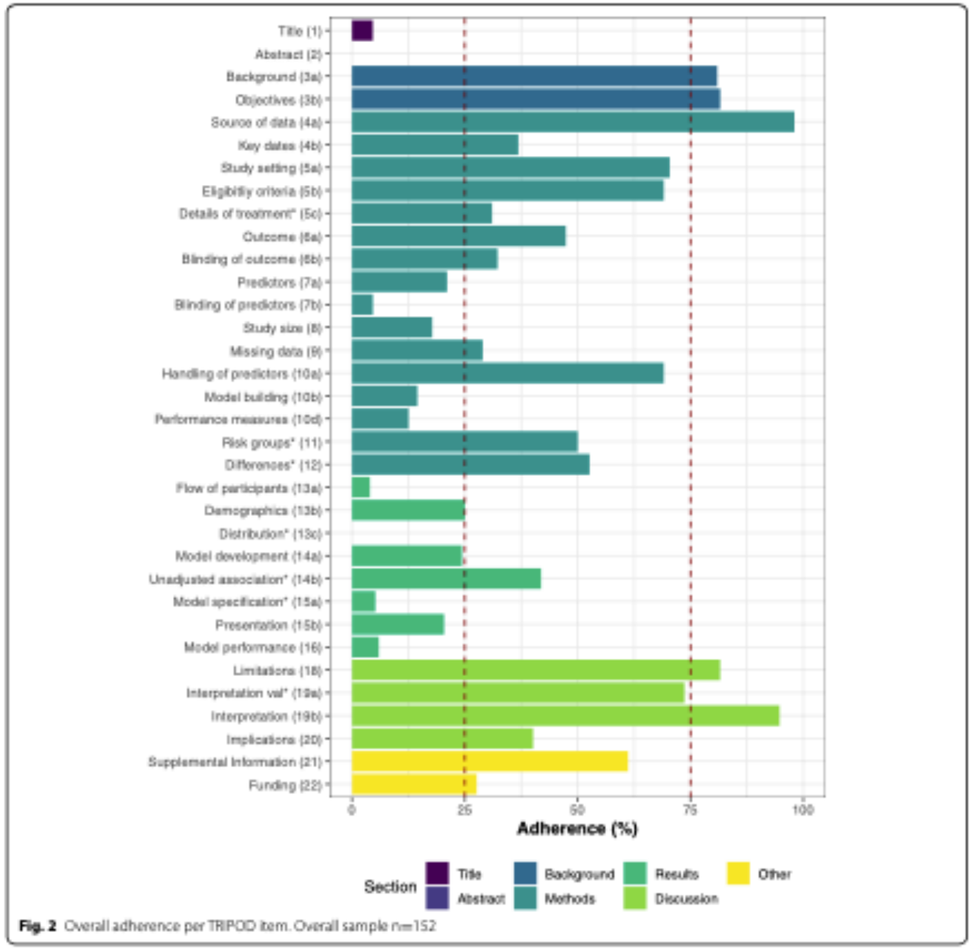
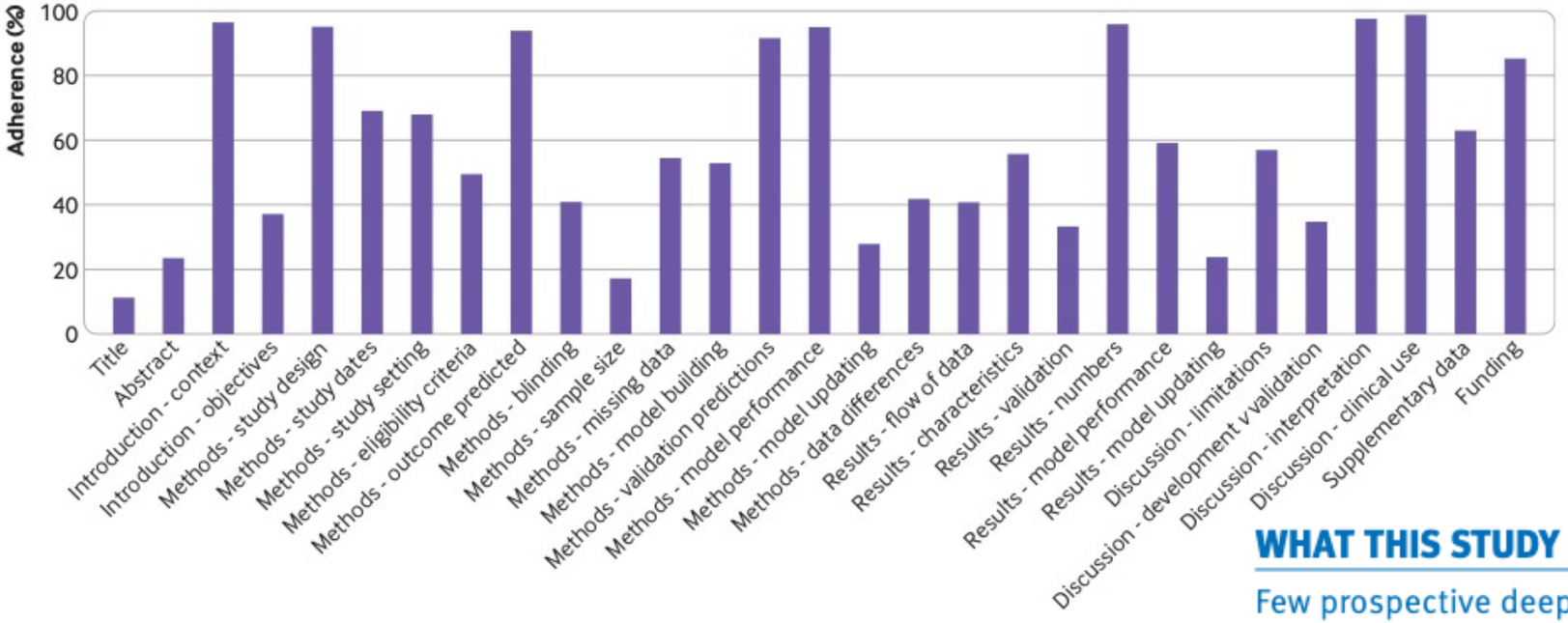


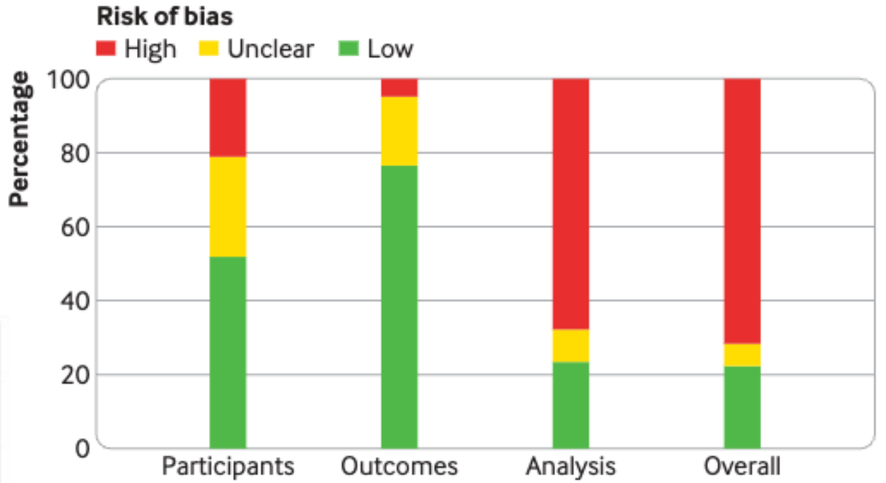
Fig. 2 Overall adherence per TRIPOD item. Overall sample n=152

Utrecht
(general medical journals)

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 Maruthappu[3]





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

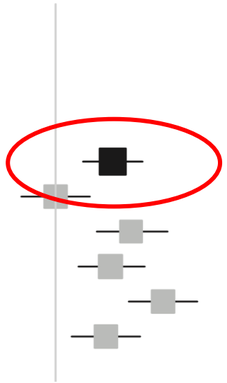
Data and code availability are lacking in most studies, and human comparator groups are often small

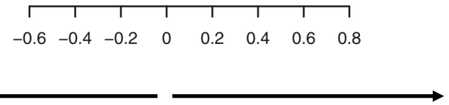
Future studies should diminish risk of bias, enhance real world clinical relevance, improve reporting and transparency, and appropriately temper conclusions

Nagendran et al, BMJ 2020

Why transparency matters: risk of bias ('off the shelf' machine learning)

	Diff logit(AUC) (95% CI)	N
Overall		
– Any ML vs LR	0.25 (0.12;0.38)	282
– Tree vs LR	0.00 (–0.15;0.15)	42
– RF vs LR	0.33 (0.18;0.49)	59
– SVM vs LR	0.24 (0.10;0.39)	43
– ANN vs LR	0.47 (0.32;0.62)	52
– Other ML vs LR	0.22 (0.07;0.37)	86





Logistic regression better

Machine learning better

ELSEVIER

Check for updates

Journal of Clinical Epidemiology

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[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

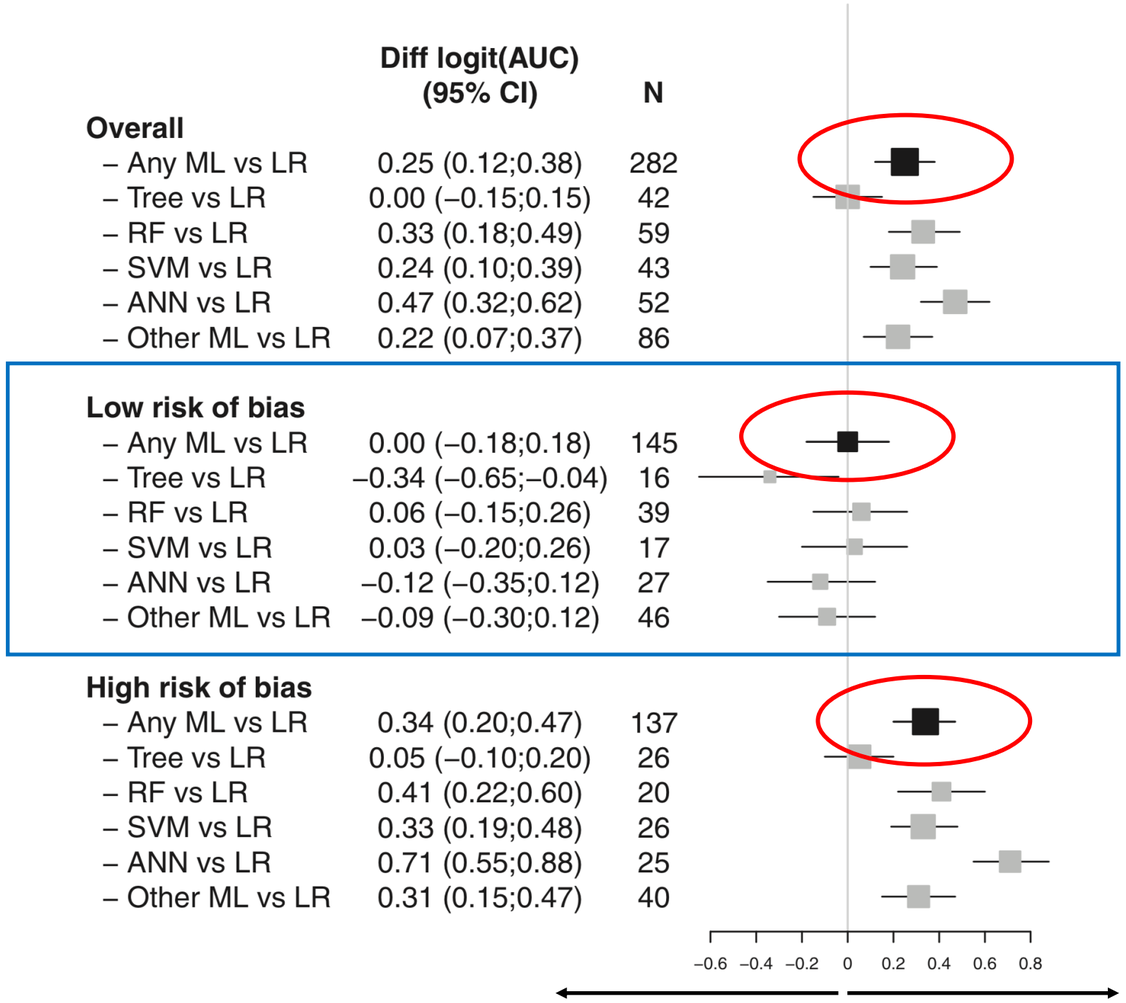
[a]*Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium*
[b]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK*
[c]*Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
[d]*Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands*
[e]*Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium*
[f]*Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK*

Accepted 5 February 2019; Published online 11 February 2019

- 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

Christodolou et al J Clin Epidemiol 2019

Why transparency matters: risk of bias ('off the shelf' machine learning)



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

Christodolou et al J Clin Epidemiol 2019

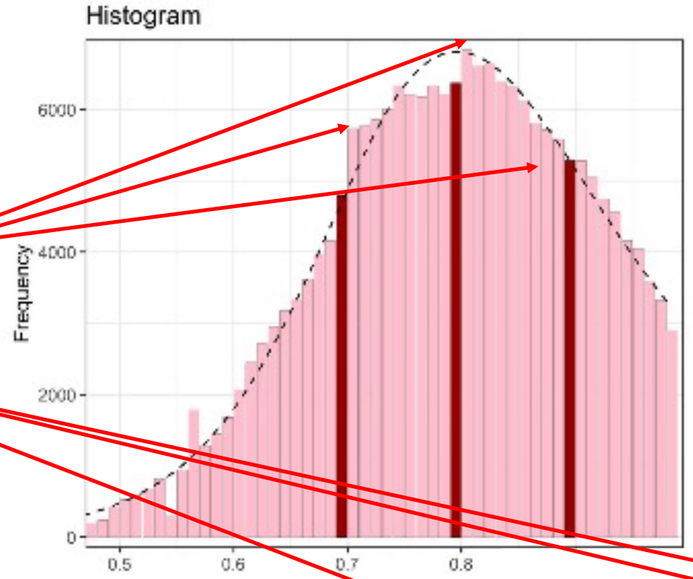
White et al BMC Med 2023

Questionable research practices

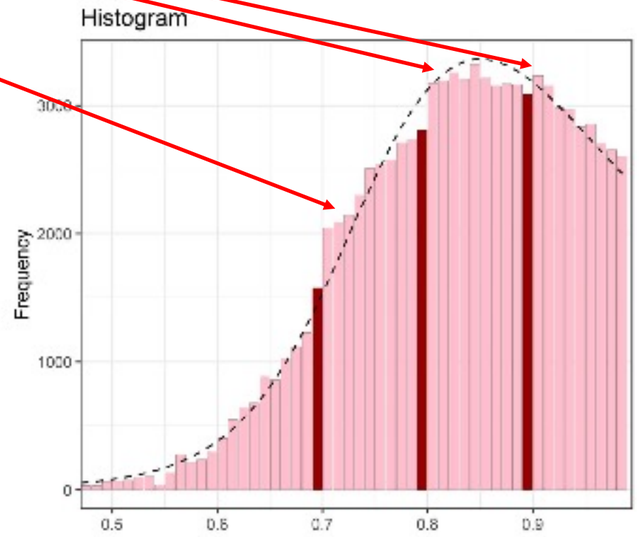
- The distribution of 306,888 AUC values (from ~97k abstracts on PubMed)
  - Clear excesses above the thresholds of 0.7, 0.8 and 0.9 and shortfalls below the thresholds

- Evidence (or suggestive) of AUC hacking

- Emphasising the need for registration, protocols, and clear and transparent reporting



Largest AUCs



Open science practices

- Increasing expectation to adhere to open science principles*
  - Protocol and study registration rare
    - Yet the norm in clinical trials
  - Data sharing statements are often expected
    - …and should go beyond 'available upon reasonable request'
    - Current reality…data is rarely shared
  - Some journals are increasingly requiring code sharing statements (e.g., BMJ [from May 2024])
    - Code to implement models uncommon
    - Hampers independent evaluation

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

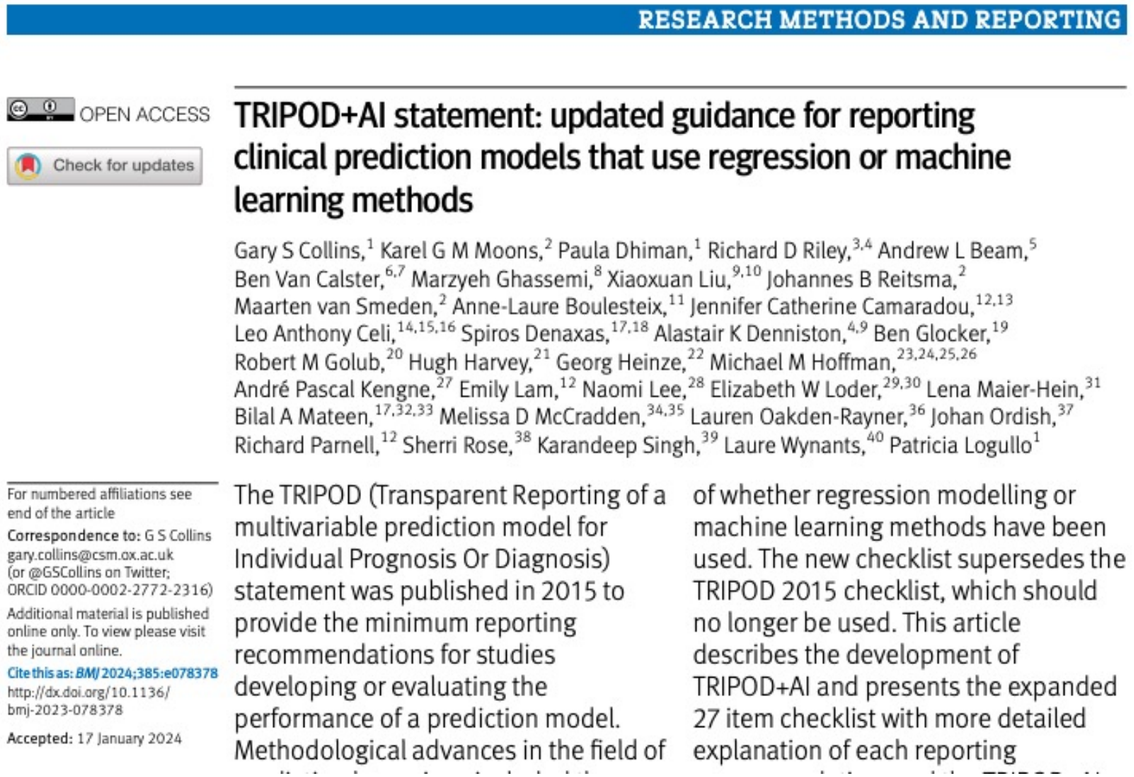
Open science practice	Frequency	% (95 CI)
Data sharing statement	35	76% (61–87%)
Available upon request	21	46% (31–61%)
Explicitly not shared	6	13% (5–26%)
Links to a website (e.g., SEER)	3	7% (1–18%)
Reported as available in the article but not	2	4% (0–15%)
Available (in supplementary material)	2	4% (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

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

TRIPOD+AI

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

Check for updates

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 Logullo[1]

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

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 TRIPOD+AI

- Followed guidance set out by the **EQUATOR Network**
- Over **200 international experts** participated in the Delphi survey
  - >27 countries covering six continents
- **28 experts** participated in an online **consensus meeting** 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

Funded by Cancer Research UK and Health Data Research UK

TRIPOD+AI

Version: 11-January-2024

Section/Topic	Item	Development / evaluation[1]	Checklist item	Reported on page
TITLE				
Title	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				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	
INTRODUCTION				
Background	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	
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	
METHODS				
Data	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	
Participants	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	
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	
Outcome	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	
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and	

TRIPOD+AI

- New checklist of reporting **27 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, to appear in 2026)

- TRIPOD+AI are **recommendations** on what to report and **not a 'how to'** on design, analysis or use for critical appraisal (see PROBAST+AI, Moons et al, BMJ 2025)

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

TRIPOD+AI

- The **clinical decision (and point in the clinical pathway)** the model is **intended to support**
  - Why is the model needed?
  - Who is the intended user? Healthcare professional, patient?

- Clear **description** and **provenance of the data** being used
  - **Rationale, richness** and **representativeness**
  - Data quality, and presence and handling of any **missing data**
  - How the data are being used to train/test
  - **Sample size** considerations (for both training and testing)

- Aspects of **fairness are embedded throughout** the guidance
  - ensuring we don't introduce tools that widen (or create) disparities in health care provision in certain sociodemographic groups (e.g., ethnicity, socioeconomic status)

- How to use the prediction model
  - Any **restrictions on use** (i.e., freely available, proprietary)

- Inclusion of an item on **‘patient and public involvement’ (PPI)**

(Kuo et al, eClinicalMedicine, 2024)

- 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

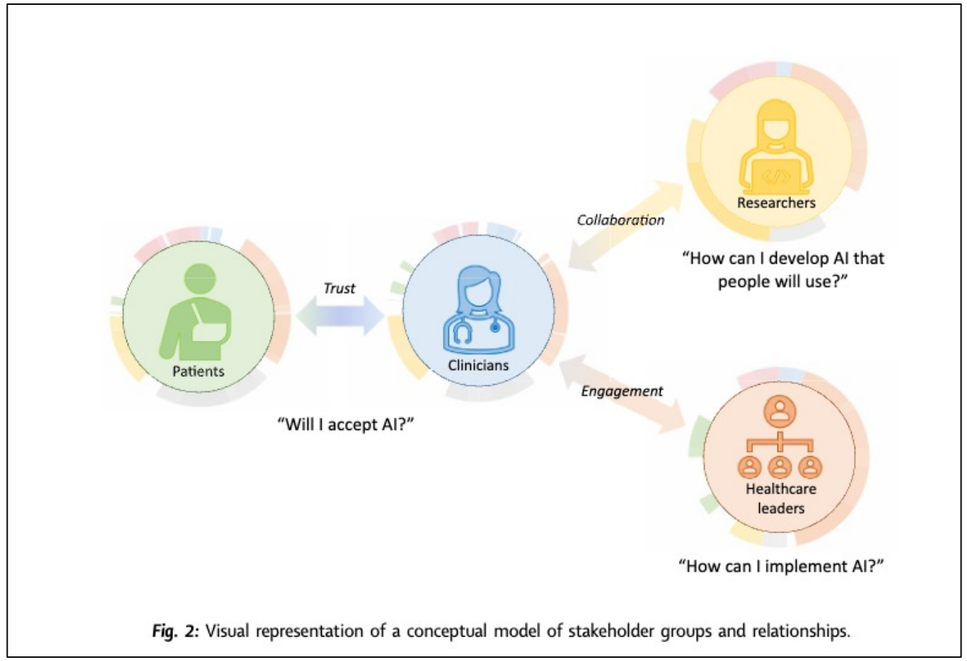


Fig. 2: Visual representation of a conceptual model of stakeholder groups and relationships.

TRIPOD+AI

- Introduction of an **‘open science’** section with reporting recommendations for

  - Funding (and role of funder)
  - Conflicts of interest
  - Study registration
  - Study protocols (TRIPOD-P in preparation)
  - Data availability
  - Code availability (analytical code and model code)
    - Acknowledging difficulties in this area (e.g., proprietary issues)
    - Any conditions/licences/hardware requirements needed to implement the AI in clinical practice

  - <u>Items that are unable to be shared should be also be declared</u>

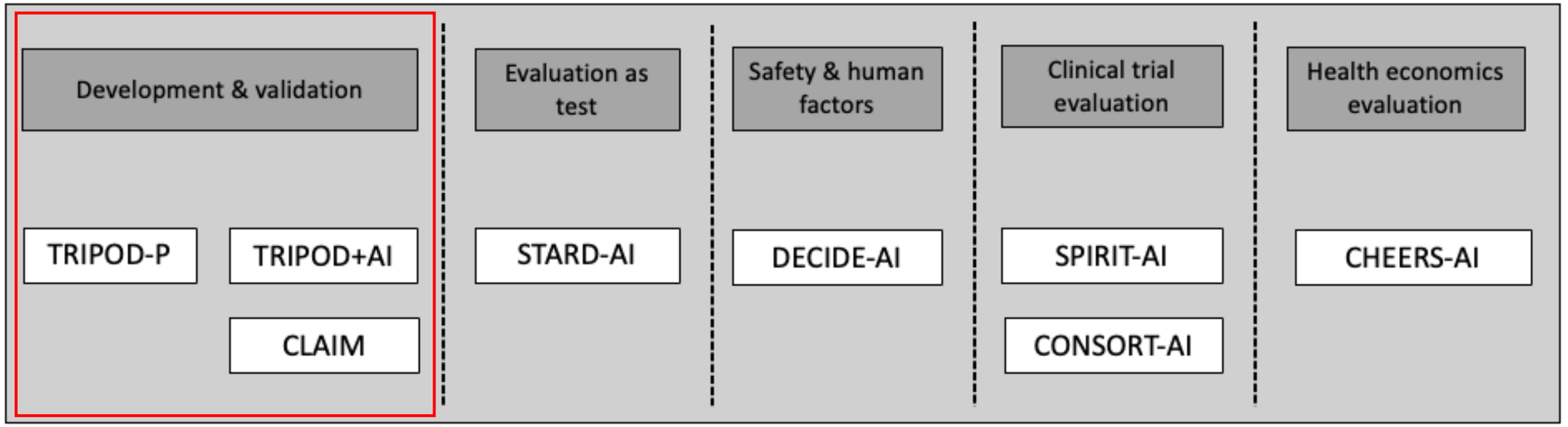
Expanded guidance

TRIPOD+AI

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*

Supplementary table 1; Collins et al BMJ 2024

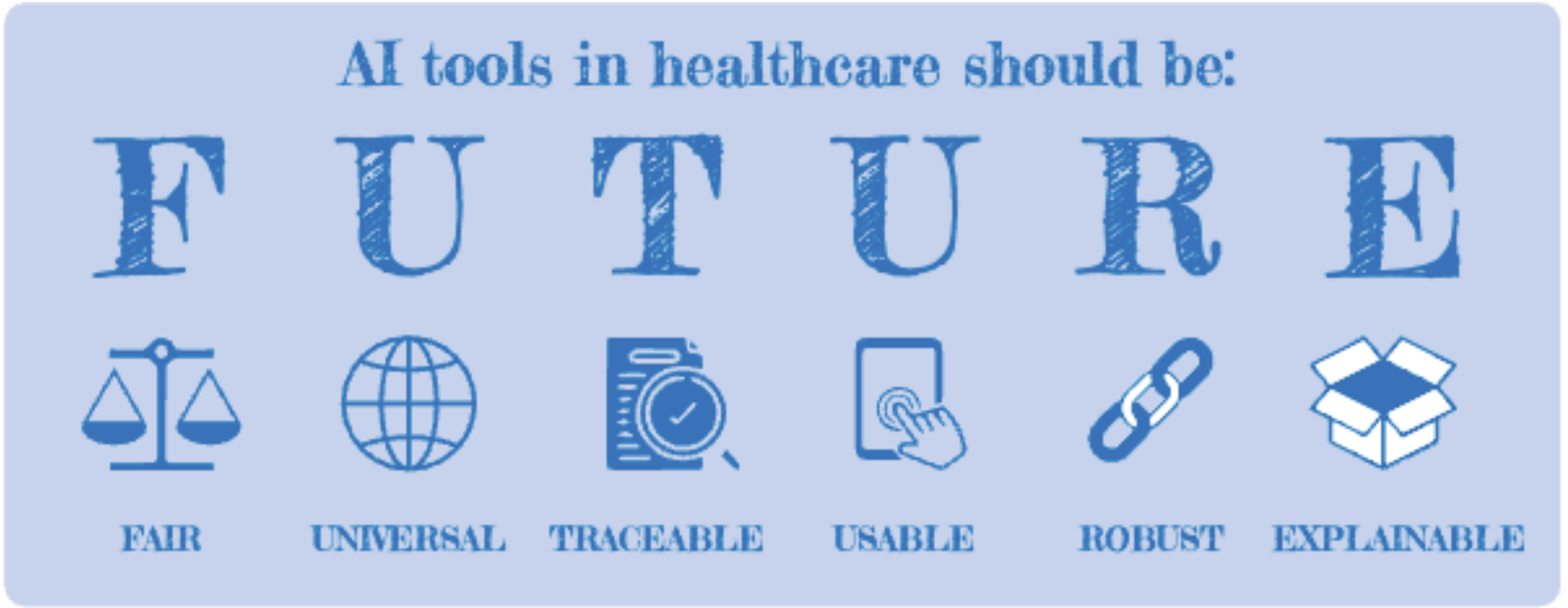


Reporting guideline	Phase of AI model development, testing or evaluation
TRIPOD-P	Protocols for AI model development, validation and updating studies (Dhiman et al, Nat Mach Intell 2023)
TRIPOD+AI	Studies describing the development, validation and updating of an AI model (Collins et al, BMJ 2024)
CLAIM-2024	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); CANGARU; responsible LLM use, Cacciamani et al, forthcoming);

FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Karim Lekadir,[1,2] Alejandro F Frangi,[3,4] Antonio R Porras,[5] Ben Glocker,[6] Celia Cintas,[7] Curtis P Langlotz,[8] Eva Weicken,[9] Folkert W Asselbergs,[10,11] Fred Prior,[12] Gary S Collins,[13] Georgios Kaissis,[14] Gianna Tsakou,[15] Irène Buvat,[16] Jayashree Kalpathy-Cramer,[17] John Mongan,[18] Julia A Schnabel,[19] Kaisar Kushibar,[1] Katrine Riklund,[20] Kostas Marias,[21] Lameck M Amugongo,[22] Lauren A Fromont,[23] Lena Maier-Hein,[24] Leonor Cerdá-Alberich,[25] Luis Martí-Bonmatí,[26] M Jorge Cardoso,[27] Maciej Bobowicz,[28] Mahsa Shabani,[29] Manolis Tsiknakis,[21] Maria A Zuluaga,[30] Marie-Christine Fritzsche,[31] Marina Camacho,[1] Marius George Linguraru,[32] Markus Wenzel,[9] Marleen De Bruijne,[33] Martin G Tolsgaard,[34] Melanie Goisauf,[35] Mónica Cano Abadía,[35] Nikolaos Papanikolaou,[36] Noussair Lazrak,[1] Oriol Pujol,[1] Richard Osuala,[1] Sandy Napel,[37] Sara Colantonio,[38] Smriti Joshi,[1] Stefan Klein,[33] Susanna Aussó,[39] Wendy A Rogers,[40] Zohaib Salahuddin,[41] Martijn P A Starmans[33];
on behalf of the FUTURE-AI Consortium



- Set of 30 ‘best’ practices addressing technical, clinical, socio-ethical, and legal dimensions – underpinned by transparency

- The guideline addresses the entire AI lifecycle, from design and development to validation and deployment, ensuring alignment with real world needs and ethical requirements

- Continuous risk assessment and mitigation are fundamental, addressing biases, data variations, and evolving challenges during the AI lifecycle

Summary

- AI is a major driver of innovative technology with enormous potential to improve patient outcomes, decision-making, workflow efficiency
  - …but it has the potential to harm, create healthcare disparities or widen existing ones

- Trustworthy AI needs thorough evaluation using high methodological standards, with complete & accurate reporting

- Lots of evidence that AI research is poorly designed, conducted and reported

- The use of tools like TRIPOD+AI, CLAIM-2024, STARD-AI, CONSORT-AI, DECIDE-AI and PROBAST+AI can play a pivotal role to improve trust in AI research at various stages in the research pipeline