

# The importance of transparency in predictive AI: the role of reporting guidelines

**Gary Collins**  
**Professor of Medical Statistics**

Director of the UK EQUATOR Centre  
Director of the Centre for Statistics in Medicine  
University of Oxford

21-March-2025

# 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 reporting
- **Scientific manuscripts** should **present sufficient data** so that a reader can **fully evaluate** the information and reach their own conclusions about the result
- **Avoiding misinterpretation** of study findings (e.g., spin/hype)

# 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
  - **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**
- **Good reporting is an essential part of good research**
  - open science, reproducibility and research(er) integrity



# 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

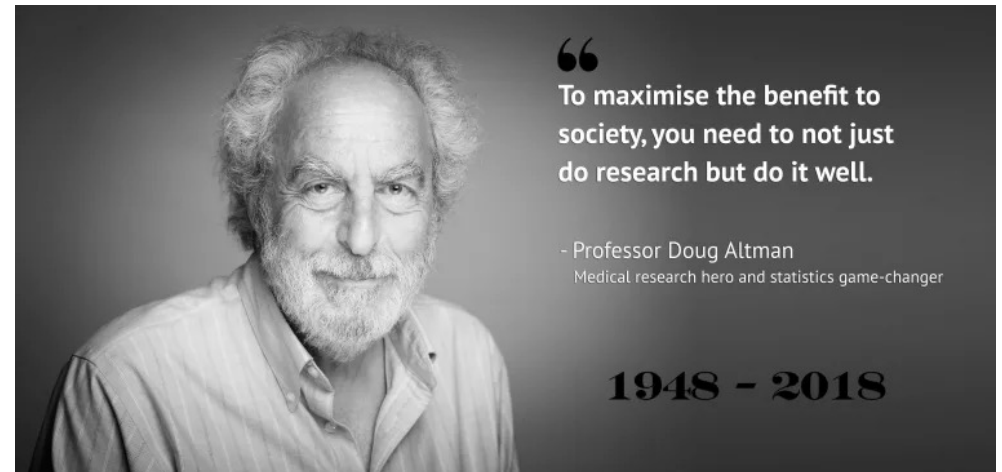
# Transparency & Reporting



[www.equator-network.org](http://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

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**”

\* Research that has limited or no value

# 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

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

[See all 659 reporting guidelines](#)



## CONSORT Statement extension for reporting abstracts of randomized controlled trials

This extension to the CONSORT Statement provides a minimum list of essential items, that authors should consider when reporting the main results of a randomized trial in any journal or conference abstract.

CONSORT for Abstract Checklist

[www.consort-statement.org](http://www.consort-statement.org)

| Item               | Description   |
|--------------------|---|
| Title              | Identification of the study as randomized   |
| Authors *          | Contact details for the corresponding author  |
| Trial design       | Description of the trial design (e.g. parallel, cluster, non-inferiority)                                   |
| Methods            |   |
| Participants       | Eligibility criteria for participants and the settings where the data were collected                        |
| Interventions      | Interventions intended for each group   |
| Objective          | Specific objective or hypothesis  |
| Outcome            | Clearly defined primary outcome for this report   |
| Randomization      | How participants were allocated to interventions  |
| Blinding (masking) | Whether or not participants, care givers, and those assessing the outcomes were blinded to group assignment |
| Results            |   |
| Numbers randomized | Number of participants randomized to each group   |
| Recruitment        | Trial status  |
| Numbers analysed   | Number of participants analysed in each group   |
| Outcome            | For the primary outcome, a result for each group and the estimated effect size and its precision            |
| Harms              | Important adverse events or side effects  |
| Conclusions        | General interpretation of the results   |
| Trial registration | Registration number and name of trial register  |
| Funding            | Source of funding   |



[PRISMA 2020 statement](#) [PRISMA Extensions](#) [PRISMA Translations](#) [PRISMA Endorsement](#)

Welcome to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) website

Key documents

[PRISMA 2020 checklist](#)

[PRISMA 2020 flow diagram](#)

[PRISMA 2020 statement paper](#)

Here you can access information about the PRISMA reporting guidelines, which are designed to help authors transparently report why their systematic review was done, what methods they used, and what they found.



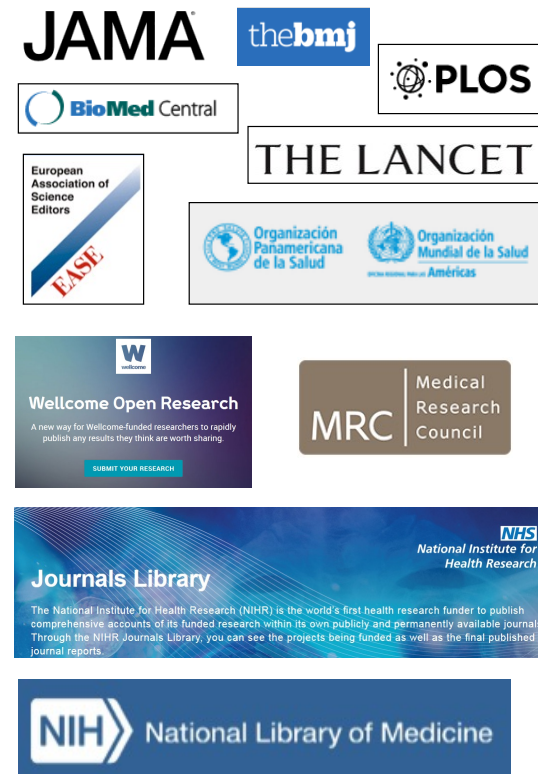
### PRISMA 2020 Checklist

| Section and Topic             | Item # | Checklist item   | Location where item is reported |
|-------------------------------|--------|--|---------------------------------|
| <b>TITLE</b>                  |        |  |                                 |
| Title                         | 1      | Identify the report as a systematic review.  |                                 |
| <b>ABSTRACT</b>               |        |  |                                 |
| Abstract                      | 2      | See the PRISMA 2020 for Abstracts checklist.   |                                 |
| <b>INTRODUCTION</b>           |        |  |                                 |
| Rationale                     | 3      | Describe the rationale for the review in the context of existing knowledge.  |                                 |
| Objectives                    | 4      | Provide an explicit statement of the objective(s) or question(s) the review addresses.   |                                 |
| <b>METHODS</b>                |        |  |                                 |
| Eligibility criteria          | 5      | Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.  |                                 |
| Information sources           | 6      | Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.  |                                 |
| Search strategy               | 7      | Present the full search strategies for all databases, registers and websites, including any filters and limits used.   |                                 |
| Selection process             | 8      | Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.                     |                                 |
| Data collection process       | 9      | Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process. |                                 |
| Data items                    | 10a    | List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.                        |                                 |
|                               | 10b    | List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.   |                                 |
| Study risk of bias assessment | 11     | Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.                                    |                                 |
| Effect measures               | 12     | Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.  |                                 |
| Synthesis methods             | 13a    | Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).   |                                 |
|                               | 13b    | Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.  |                                 |
|                               | 13c    | Describe any methods used to tabulate or visually display results of individual studies and syntheses.   |                                 |
|                               | 13d    | Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.  |                                 |
|                               | 13e    | Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).   |                                 |
|                               | 13f    | Describe any sensitivity analyses conducted to assess robustness of the synthesized results.   |                                 |
| Reporting bias assessment     | 14     | Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).  |                                 |
| Certainty assessment          | 15     | Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.  |                                 |



# Reporting Guidelines/EQUATOR endorsed by

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



# Journal Instructions to authors

How Do I?

## Determine My Article Type

### Categories of Articles



#### Research

| Article Type                               | Description                             | Requirements  |
|--|---|---|
| <b>Original Investigation</b><br>full info | Clinical trial                          | <ul style="list-style-type: none"><li>• 3000 words</li><li>• ≤5 tables and/or figures</li><li>• <b>Structured abstract</b></li><li>• <b>Key Points</b></li><li>• <b>Data Sharing Statement</b></li><li>• <b>Follow EQUATOR Reporting Guidelines</b></li></ul> |
|  | 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               |   |

#### Statistical issues



#### 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 **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 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](http://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](http://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/](http://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/](http://www.equator-network.org/home/)) and the NLM's Research Reporting Guidelines and Initiatives ([www.nlm.nih.gov/services/research\\_report\\_guide.html](http://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



# 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?

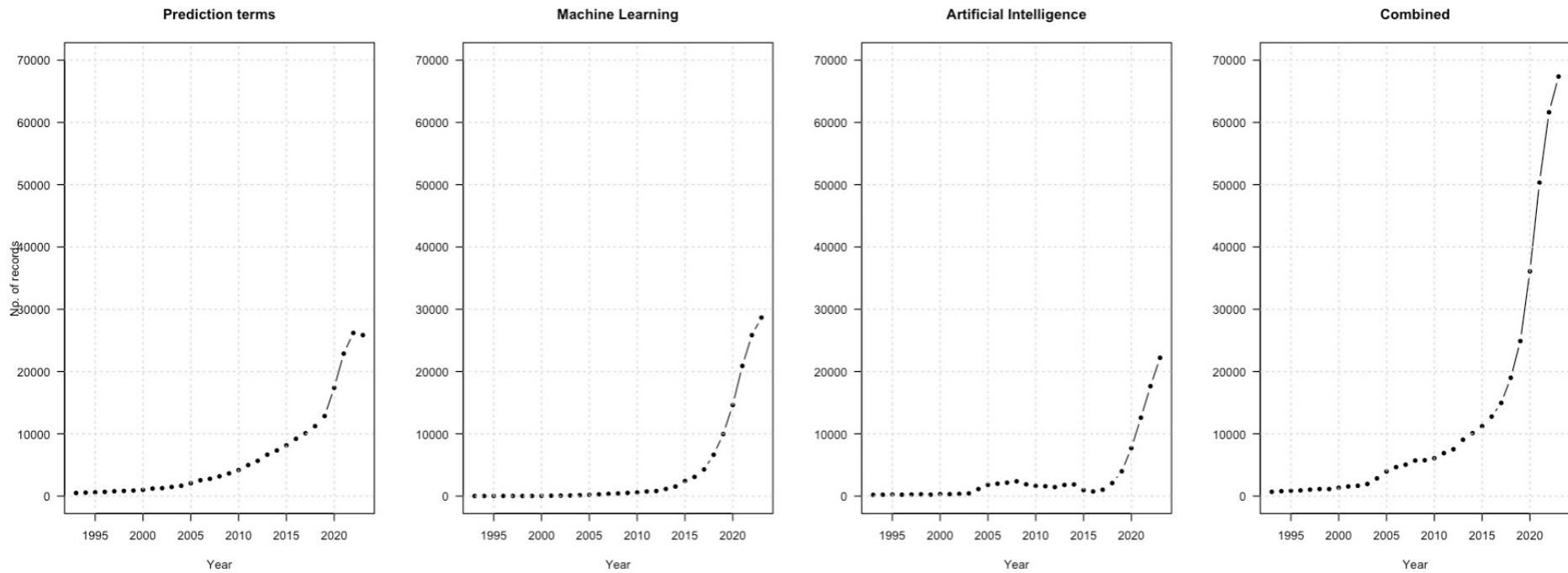
Christian B. van der Pol<sup>1</sup>, Matthew D. F. McInnes<sup>1,2\*</sup>, William Petrcich<sup>2</sup>, Adam S. Tunis<sup>1</sup>, Ramez Hanna<sup>1</sup>

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 patients outcomes and influence future research**

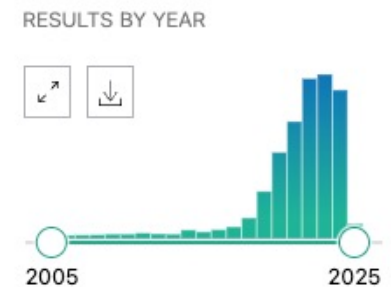
# Prediction is a hot topic



# ModelMania: e.g., prediction using the SEER data

- **SEER is a population-based cancer registry from the US**
  - Covering ~48% of the US population
- **>2000 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
  - >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’ era (regression models)

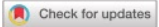
Example: 228 articles [development of 408 prognostic models for patients with chronic obstructive pulmonary disease]

- **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



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

Vanessa Bellou,<sup>1,2</sup> Lazaros Belbasis,<sup>1</sup> Athanasios K Konstantinidis,<sup>2</sup> Ioanna Tzoulaki,<sup>1,3,4</sup> Evangelos Evangelou<sup>1,3</sup>

<sup>1</sup>Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece  
<sup>2</sup>Department of Respiratory Medicine, University Hospital of Ioannina, University of Ioannina Medical School, Ioannina, Greece  
<sup>3</sup>Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK  
<sup>4</sup>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

RESEARCH ARTICLE

Open Access

## Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting

Gary S Collins\*, Susan Mallett, Omar Omar and Ly-Mee Yu

## Prognostic models in obstetrics: available, but far from applicable

C. Emily Kleinrouweler, MD, PhD; Fiona M. Cheong-See, MRCOG; Gary S. Collins, PhD; Anneke Kwee, MD, PhD; Shakila Thangaratinam, PhD; Khalid S. Khan, MSc, MRCOG; Ben Willem J. Mol, MD, PhD; Eva Pajkrt, MD, PhD; Karel G. M. Moons, PhD; Ewoud Schuit, PhD

http://www.biomedcentral.com/1471-2288/14/40

RESEARCH ARTICLE

Open Access

## External validation of multivariable prediction models: a systematic review of methodological conduct and reporting

Gary S Collins<sup>1\*</sup>, Joris A de Groot<sup>2</sup>, Susan Dutton<sup>1</sup>, Omar Omar<sup>1</sup>, Milensu Shanyinde<sup>1</sup>, Abdelouahid Tajar<sup>1</sup>, Meryn Voysey<sup>1</sup>, Rose Wharton<sup>1</sup>, Ly-Mee Yu<sup>1</sup>, Karel G Moons<sup>2</sup> and Douglas G Altman<sup>1</sup>

Curr Osteoporos Rep (2012) 10:199–207  
DOI 10.1007/s11914-012-0108-1

EVALUATION AND MANAGEMENT (M KLEEREKOPER, SECTION EDITOR)

## Fracture Risk Assessment: State of the Art, Methodologically Unsound, or Poorly Reported?

Gary S. Collins · Karl Michaëlsson

OPEN ACCESS Freely available online

PLOS MEDICINE

## Reporting and Methods in Clinical Prediction Research: A Systematic Review

Walter Bouwmeester<sup>1\*</sup>, Nicolaas P. A. Zuithoff<sup>1\*</sup>, Susan Mallett<sup>2</sup>, Mirjam I. Geerlings<sup>1</sup>, Yvonne Vergouwe<sup>1,3</sup>, Ewout W. Steyerberg<sup>3</sup>, Douglas G. Altman<sup>4</sup>, Karel G. M. Moons<sup>1\*</sup>

<sup>1</sup>Julius Center for Health Care Research, University of Oxford, Oxford, UK



ELSEVIER

Journal of Clinical Epidemiology ■ (2012) ■

**Journal of  
Clinical  
Epidemiology**

REVIEW ARTICLE

A systematic review finds prediction models for chronic kidney were poorly reported and often developed using inappropriate methods

Gary S. Collins\*, Omar Omar, Milensu Shanyinde, Ly-Mee Yu

Centre for Statistics in Medicine, Wolfson College Annex, University of Oxford, Linton Road, Oxford OX2 6UD, UK

*Cancer Investigation*, 27:235–243, 2009  
ISSN: 0735-7907 print / 1532-4192 online  
Copyright © Informa Healthcare USA, Inc.  
DOI: 10.1080/07357900802572110

SPECIAL ARTICLE

## Prognostic Models: A Methodological Framework and Review of Models for Breast Cancer

Douglas G. Altman

Centre for Statistics in Medicine, University of Oxford, Oxford, UK

# 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)
  - Opportunity to highlight good methodology
  - Opportunity to flag methodological issues
- **Widely cited / included in journal author instructions**
  - Statement paper >9000 times; E&E paper >4000 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](http://www.tripod-statement.org)).

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

*Ann Intern Med.* 2015;162:W1-W73. doi:10.7326/M14-0698 [www.annals.org](http://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

Journal of Clinical Epidemiology  
Volume 138, October 2021, Pages 60-72

Original Article

## Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved

Paula Dhiman<sup>a,b</sup>, Jie Ma<sup>a</sup>, Constanza L. Andaur Navarro<sup>c</sup>, Benjamin Speich<sup>a,d</sup>, Garrett Bullock<sup>a</sup>, Johanna A.A. Damen<sup>e</sup>, Shona Kirtley<sup>a</sup>, Lotty Hooft<sup>c</sup>, Richard D. Riley<sup>f</sup>, Ben Van Calster<sup>g,h,i</sup>, Karel G.M. Moons<sup>c</sup>, Gary S. Collins<sup>a,b</sup>

Research | Open Access | Published: 08 April 2022

## Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review

Paula Dhiman<sup>a</sup>, Jie Ma<sup>a</sup>, Constanza L. Andaur Navarro<sup>c</sup>, Benjamin Speich<sup>a</sup>, Garrett Bullock<sup>a</sup>, Johanna A. A. Damen<sup>e</sup>, Lotty Hooft<sup>c</sup>, Shona Kirtley<sup>a</sup>, Richard D. Riley<sup>f</sup>, Ben Van Calster<sup>g,h,i</sup>, Karel G. M. Moons<sup>c</sup> & Gary S. Collins<sup>a</sup>

*BMC Medical Research Methodology*, 22, Article number: 101 (2022) | Cite this article

Home > Diagnostic and Prognostic Research > Article

## Risk of bias of prognostic models developed using machine learning: a systematic review in oncology

Diagnostic and Prognostic Research

Research | Open Access | Published: 07 July 2022  
6, Article number: 13 (2022)

Paula Dhiman<sup>a</sup>, Jie Ma<sup>a</sup>, Constanza L. Andaur Navarro<sup>c</sup>, Benjamin Speich<sup>a</sup>, Garrett Bullock<sup>a</sup>, Johanna A. A. Damen<sup>e</sup>, Lotty Hooft<sup>c</sup>, Shona Kirtley<sup>a</sup>, Richard D. Riley<sup>f</sup>, Ben Van Calster<sup>g,h,i</sup>, Karel G. M. Moons<sup>c</sup> & Gary S. Collins<sup>a</sup>

Journal of Clinical Epidemiology  
Volume 157, May 2023, Pages 120-133

Review Article

## Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review

Paula Dhiman<sup>a,b</sup>, Jie Ma<sup>a</sup>, Constanza L. Andaur Navarro<sup>c</sup>, Benjamin Speich<sup>a,d</sup>, Garrett Bullock<sup>a</sup>, Johanna A.A. Damen<sup>e</sup>, Lotty Hooft<sup>c</sup>, Shona Kirtley<sup>a</sup>, Richard D. Riley<sup>f</sup>, Ben Van Calster<sup>g,h,i</sup>, Karel G.M. Moons<sup>c</sup>, Gary S. Collins<sup>a,b</sup>

Oxford  
(oncology)

Home > BMC Medical Research Methodology > Article

## Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review

BMC Medical Research Methodology

Research | Open Access | Published: 13 January 2022  
22, Article number: 12 (2022)

Constanza L. Andaur Navarro<sup>a</sup>, Johanna A. A. Damen<sup>b</sup>, Toshihiko Takada<sup>c</sup>, Steven W. J. Nijman<sup>d</sup>, Paula Dhiman<sup>e</sup>, Jie Ma<sup>f</sup>, Gary S. Collins<sup>g</sup>, Ram Bajpai<sup>h</sup>, Richard D. Riley<sup>i</sup>, Karel G. M. Moons<sup>j</sup> & Lotty Hooft<sup>k</sup>

Journal of Clinical Epidemiology  
Volume 154, February 2023, Pages 8-22

Review Article

## Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models

Constanza L. Andaur Navarro (Doctoral Student)<sup>a,b</sup>, Johanna A.A. Damen (Assistant Professor)<sup>a,b</sup>, Maarten van Smeden (Associate Professor)<sup>c</sup>, Toshihiko Takada (Assistant Professor)<sup>d</sup>, Steven W.J. Nijman (Doctoral Student)<sup>e</sup>, Paula Dhiman (Research Fellow)<sup>f</sup>, Jie Ma (Medical Statistician)<sup>g</sup>, Gary S. Collins (Professor)<sup>g,h</sup>, Ram Bajpai (Research Fellow)<sup>g</sup>, Richard D. Riley (Professor)<sup>g</sup>, Karel G.M. Moons (Professor)<sup>a,b</sup>, Lotty Hooft (Professor)<sup>a,b</sup>

Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review

BMJ 2021 ; 375 doi: https://doi.org/10.1136/bmj.n2281 (Published 20 October 2021)  
Cite this as: BMJ 2021;375:n2281

Article | Related content | Metrics | Responses | Peer review

Constanza L Andaur Navarro<sup>a</sup>, Johanna A A Damen<sup>b</sup>, Toshihiko Takada<sup>c</sup>, Steven W J Nijman<sup>d</sup>, Paula Dhiman<sup>e</sup>, Jie Ma<sup>f</sup>, Gary S Collins<sup>g</sup>, Ram Bajpai<sup>h</sup>, Richard D Riley<sup>i</sup>, Karel G M Moons<sup>j</sup>, professor<sup>1,2</sup>, Lotty Hooft, professor<sup>1,2</sup>

Journal of Clinical Epidemiology  
Volume 158, June 2023, Pages 99-110

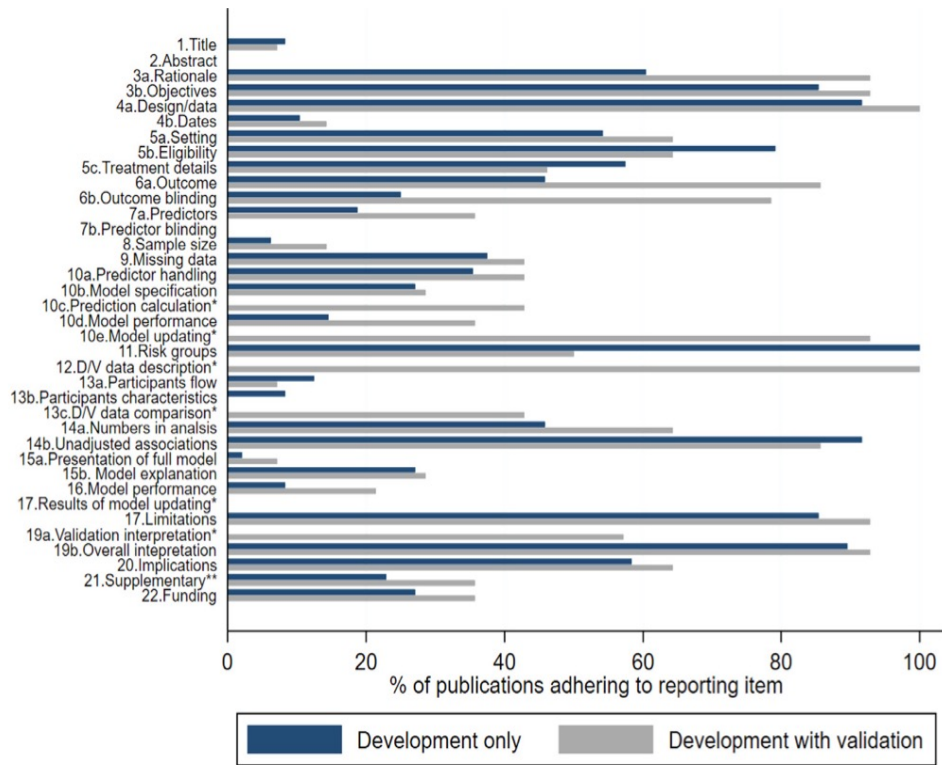
Review Article

## Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models

Constanza L. Andaur Navarro<sup>a,b</sup>, Johanna A.A. Damen<sup>a,b</sup>, Toshihiko Takada<sup>c</sup>, Steven W.J. Nijman<sup>d</sup>, Paula Dhiman<sup>e</sup>, Jie Ma<sup>f</sup>, Gary S. Collins<sup>g</sup>, Ram Bajpai<sup>h</sup>, Richard D. Riley<sup>i</sup>, Karel G.M. Moons<sup>a,b</sup>, Lotty Hooft<sup>a,b</sup>

Utrecht  
(general medical journals)

# Adherence to TRIPOD



**Oxford  
(oncology)**

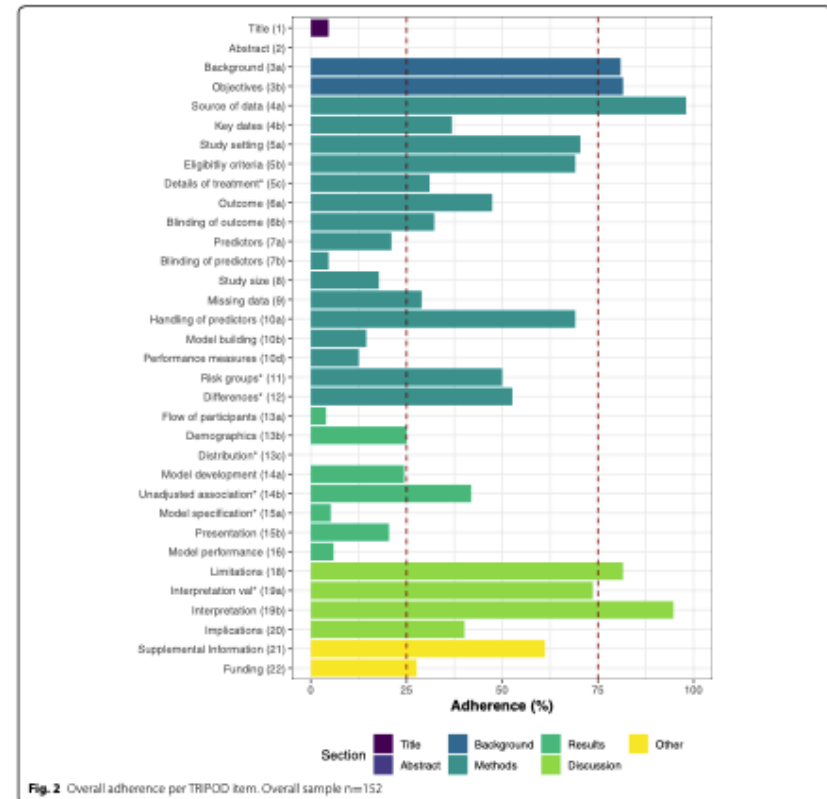


Fig. 2 Overall adherence per TRIPOD item. Overall sample n=152

**Utrecht  
(general medical journals)**



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

### Abstract

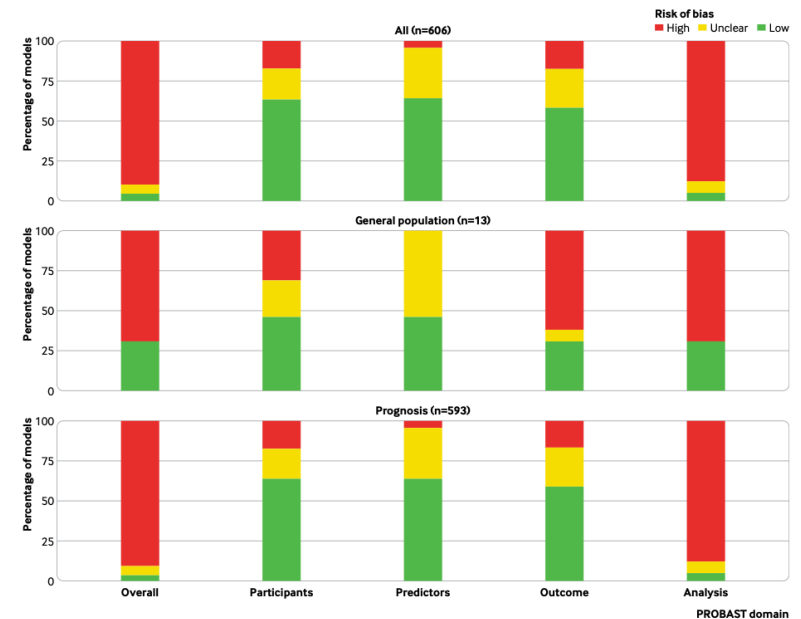
#### OBJECTIVE

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.

# Reporting of machine learning research

## Reporting concerns identified include

- Characteristics of the data
- Small sample size
- Handling of missing data
- Description of model development
- Details on hyperparameter tuning
- Details on model validation
- Performance evaluation
  - Often a focus on discrimination, or measures of accuracy
  - Calibration overlooked
- Model availability
  - Where is the model?
  - How to use it



Journal of Clinical Epidemiology 110 (2019) 12–22

**Journal of  
Clinical  
Epidemiology**

## REVIEW

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

Evangelia Christodoulou<sup>a</sup>, Jie Ma<sup>b</sup>, Gary S. Collins<sup>b,c</sup>, Ewout W. Steyerberg<sup>d</sup>,  
Jan Y. Verbakel<sup>a,e,f</sup>, Ben Van Calster<sup>a,d,\*</sup>

<sup>a</sup>Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium

<sup>b</sup>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

<sup>c</sup>Oxford University Hospitals NHS Foundation Trust, Oxford, UK

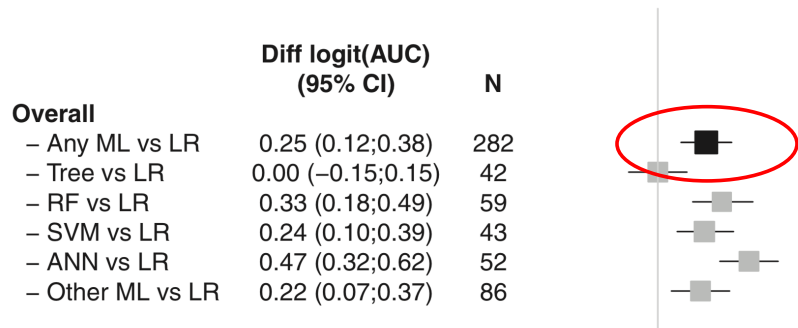
<sup>d</sup>Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands

<sup>e</sup>Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium

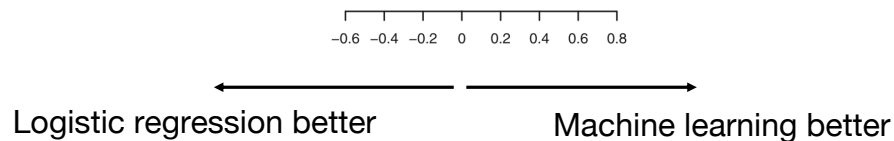
<sup>f</sup>Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK

Accepted 5 February 2019; Published online 11 February 2019

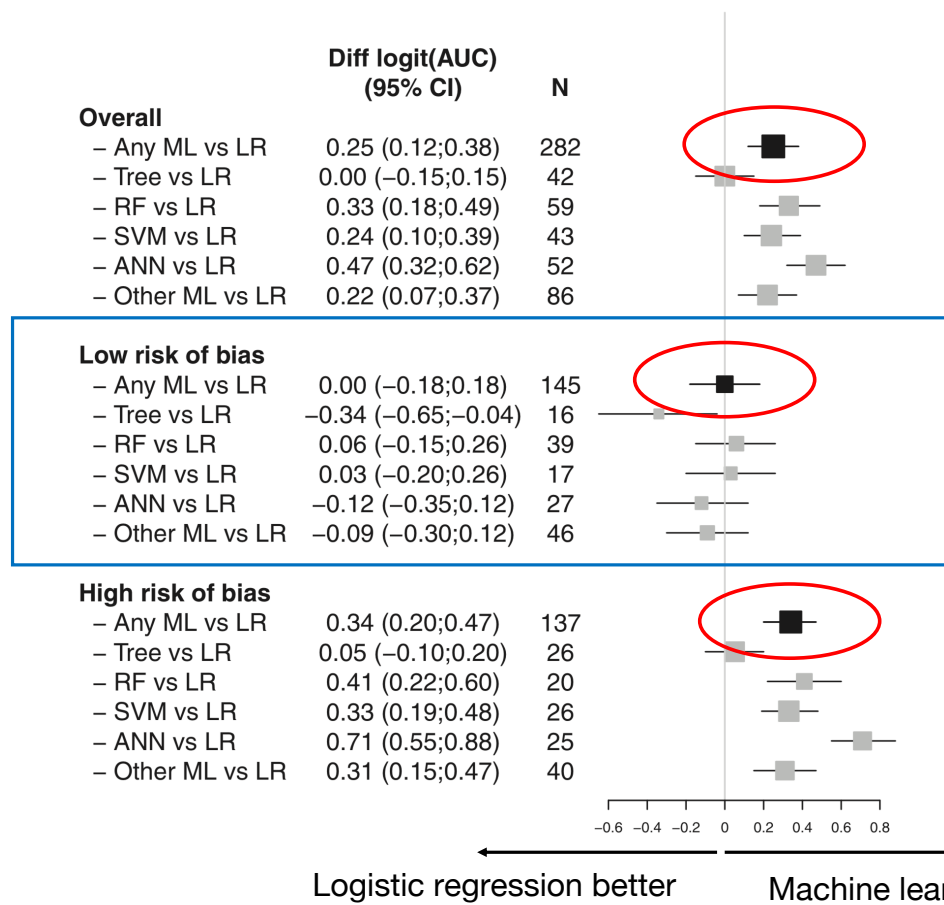
# Why it matters: risk of bias ('off the shelf' ML)



- 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



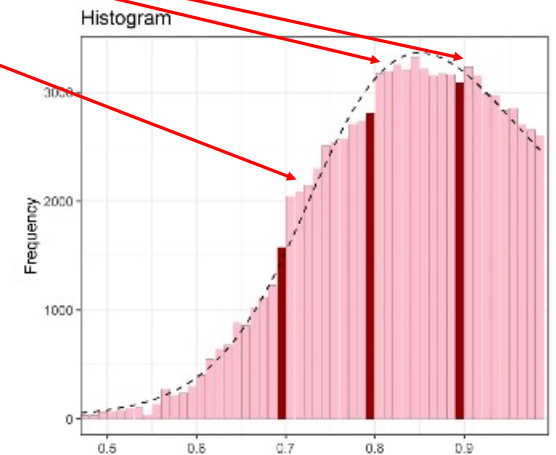
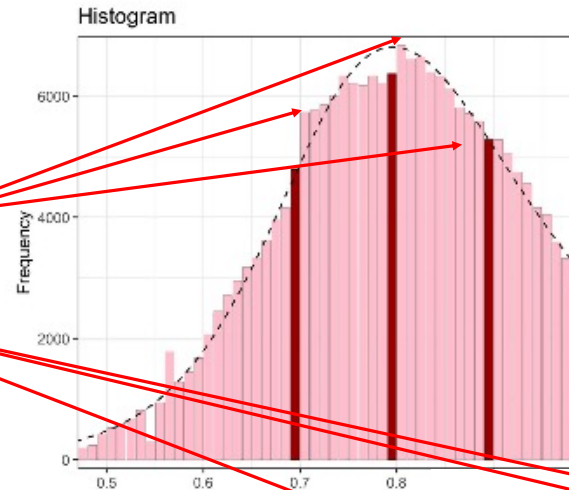
# Why methods matter: risk of bias ('off the shelf' ML)



- 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

# 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 an expectation for trials
  - Some 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

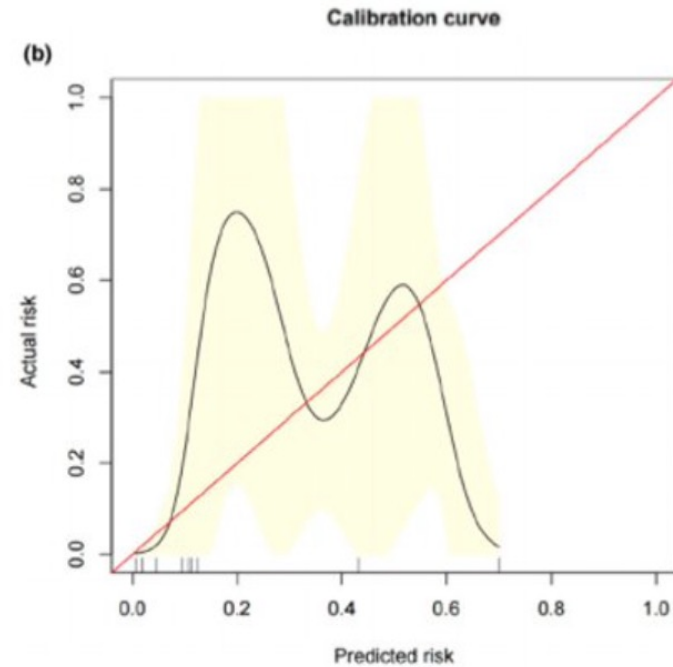
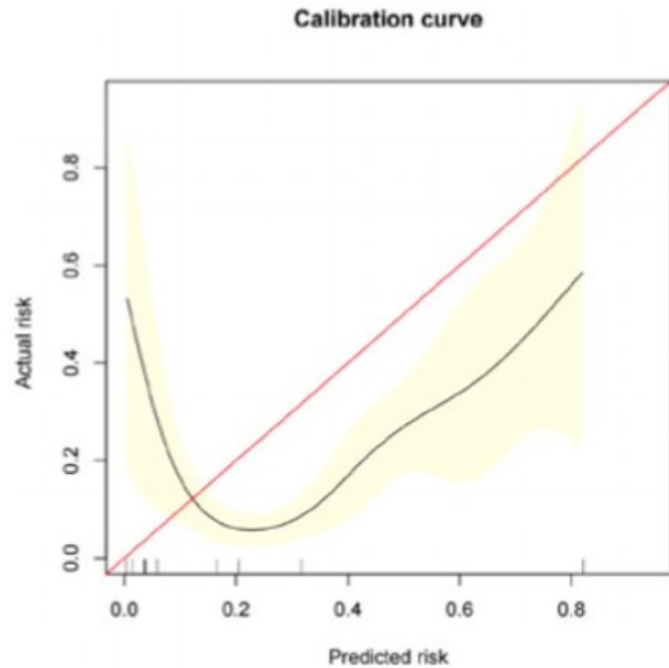
\* or give an explicit and meaningful justification for not adhering to open science (e.g., ethical/legal reasons, proprietary)

**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

# Overinterpretation ('spin')



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

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

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 2025)



# Developing TRAPPOD+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 <sup>1</sup> | Checklist item   | Reported on page |
|-------------------------|------|---------------------------------------|--|------------------|
| <b>TITLE</b>            |      |                                       |  |                  |
| <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   |                  |
| <b>ABSTRACT</b>         |      |                                       |  |                  |
| <i>Abstract</i>         | 2    | D;E                                   | See TRIPOD+AI for Abstracts checklist  |                  |
| <b>INTRODUCTION</b>     |      |                                       |  |                  |
| <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)  |                  |
| <b>METHODS</b>          |      |                                       |  |                  |
| <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  |                  |

# TRIPOD+AI

- 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 explicitly cover generative AI, but TRIPOD-LLM now available (Gallifant et al, Nat Med 2025);  
Interactive website ([tripod-llm.vercel.app](https://tripod-llm.vercel.app))

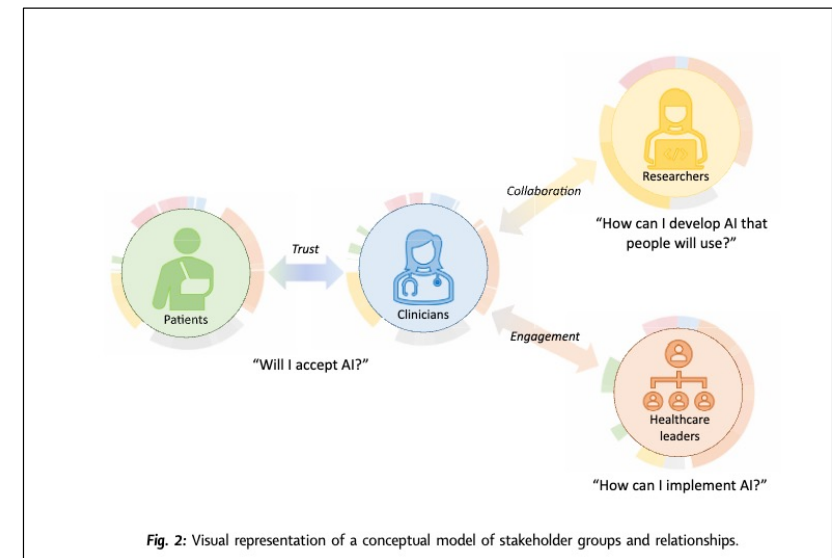
# TRAPOD + AI

- The **clinical decision** the model is **intended to support**
  - Why is the model needed?
- Clear **description and provenance of the data** being used
  - **Rationale, richness** and **representativeness**
  - Data quality and handling of any **missing data**
  - How the data are being used to train/test
  - **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

# TRIPOD + AI

- Inclusion of a new **'open science'** section with sub-items on
  - 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
    - TRIPOD-Code in preparation
  - Items that are unable to be shared should be declared

# Expanded guidance



Version: 7-February-2024

| Section/Topic      | Item | Checklist item  |
|--------------------|------|---|
|                    | 9c   | D;E<br><b>If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors</b> <ul style="list-style-type: none"> <li>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</li> <li>If the measurement and interpretation require (additional) training or specific instructions, then these should be reported. This could be reported in the supplementary material</li> </ul>  |
| Sample size        | 10   | D;E<br><b>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</b> <ul style="list-style-type: none"> <li>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</li> <li>Provide details and all estimates used in any sample size calculation</li> <li>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</li> </ul>  |
| Missing data       | 11   | D;E<br><b>Describe how missing data were handled. Provide reasons for omitting any data</b> <ul style="list-style-type: none"> <li>Missing data is an omnipresent problem. Authors should report for each predictor being considered for inclusion in the model the number of missing values</li> <li>The handling of missing values should be reported, including any assumptions for the reason of the missingness</li> <li>If individuals (or predictors) have been omitted due to the missing values, this should be reported, and reasons given</li> <li>If missing values have been imputed, then full details of the method for imputing any missing values should be reported</li> <li>If missing values have been imputed confirm it was done separately for the training and any test data (i.e., avoiding leakage)</li> </ul>  |
| Analytical methods | 12a  | D<br><b>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</b> <ul style="list-style-type: none"> <li>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)</li> <li>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</li> <li>If the data has been partitioned into training (including any hyperparameter tuning data) and test data, confirm that there has been no data leakage</li> </ul> |



# Other reporting guidelines in the AI landscape

# AI driven healthcare studies

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

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

## Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, and the SPIRIT-AI and CONSORT-AI Working Group\*

The CONSORT 2010 statement provides minimum guidelines for reporting randomised trials. Its widespread use has been instrumental in ensuring transparency in the evaluation of new interventions. More recently, there has been a growing recognition that interventions involving artificial intelligence (AI) need to undergo rigorous, prospective

*Lancet Digital Health* 2020; 2: e537-548  
Published Online September 9, 2020

nature medicine

Consensus Statement <https://doi.org/10.1038/s41591-024-03425-5>

## The TRIPOD-LLM reporting guideline for studies using large language models

Received: 24 July 2024  
Accepted: 21 November 2024  
Published online: 8 January 2025

Check for updates

Jack Gallifant<sup>1,2,3</sup>, Majid Afshar<sup>4,29</sup>, Saleem Ameen<sup>15,6,29</sup>, Yindalon Aphinyanaphongs<sup>7,29</sup>, Shan Chen<sup>3,8,29</sup>, Giovanni Cacciamani<sup>9,10,29</sup>, Dina Demner-Fushman<sup>11,29</sup>, Dmitriy Dligach<sup>12,29</sup>, Roxana Daneshjou<sup>13,14,29</sup>, Christinne Fernandes<sup>1,29</sup>, Lasse Hyldig Hansen<sup>15,29</sup>, Adam Landman<sup>16,29</sup>, Lisa Lehmann<sup>16,29</sup>, Liam G. McCoy<sup>17,29</sup>, Timothy Miller<sup>18,29</sup>, Amy Moreno<sup>19,29</sup>, Nikolaj Munch<sup>15,29</sup>, David Restrepo<sup>1,20,29</sup>, Guergana Savova<sup>18,29</sup>, Renato Umeton<sup>21,29</sup>, Judy Wawira Gichoya<sup>22,29</sup>, Gary S. Collins<sup>23,24</sup>, Karel G. M. Moons<sup>25,26</sup>, Leo A. Celi<sup>1,27,28</sup> & Danielle S. Bitterman<sup>30</sup>✉

CONSENSUS STATEMENT <https://doi.org/10.1038/s41591-022-01772-9>

nature medicine

Check for updates

## Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI

Baptiste Vasey<sup>1,2,3</sup>✉, Myura Nagendran<sup>4</sup>, Bruce Campbell<sup>5,6</sup>, David A. Clifton<sup>2</sup>, Gary S. Collins<sup>7</sup>, Spiros Denaxas<sup>8,9,10,11</sup>, Alastair K. Denniston<sup>12,13,14</sup>, Livia Faes<sup>14</sup>, Bart Geerts<sup>15</sup>, Mudathir Ibrahim<sup>1,16</sup>, Xiaoxuan Liu<sup>12,13</sup>, Bilal A. Mateen<sup>8,17,18</sup>, Piyush Mathur<sup>19</sup>, Melissa D. McCradden<sup>20,21</sup>, Lauren Morgan<sup>22</sup>, Johan Ordish<sup>23</sup>, Campbell Rogers<sup>24</sup>, Suchi Saria<sup>25,26</sup>, Daniel S. W. Ting<sup>27,28</sup>, Peter Watkinson<sup>3,29</sup>, Wim Weber<sup>30</sup>, Peter Wheatstone<sup>31</sup>, Peter McCulloch<sup>1</sup> and the DECIDE-AI expert group\*

# DECIDE-AI

Focussed on **early-stage AI studies** which are important stepping stones towards large-scale (costly) comparative trials.

The objective of DECIDE-AI is to improve reporting of clinical AI studies along four main axes:

- the performance of the AI systems when **first used with humans** in small-scale, actual clinical settings
- the **safety profile** of the AI systems prior to large-scale utilisation
- the **human factors** (ergonomic) evaluation of the AI systems
- the **preparatory** steps towards large-scale (costly) **randomised controlled trials**

## CONSENSUS STATEMENT

<https://doi.org/10.1038/s41591-022-01772-9>

nature  
medicine

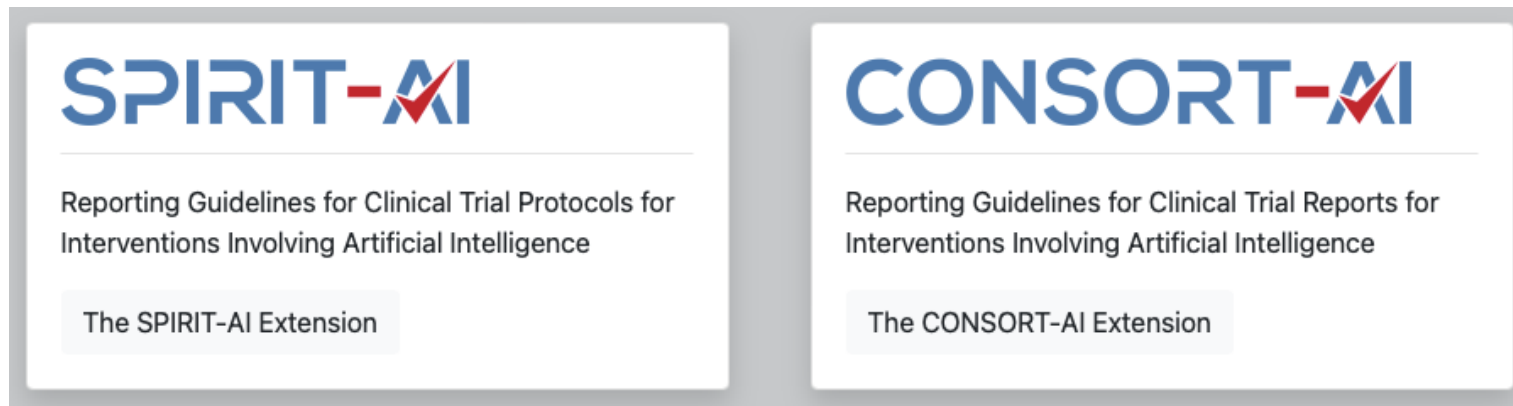


## Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI

Baptiste Vasey<sup>1,2,3</sup>✉, Myura Nagendran<sup>4</sup>, Bruce Campbell<sup>5,6</sup>, David A. Clifton<sup>2</sup>, Gary S. Collins<sup>7</sup>, Spiros Denaxas<sup>8,9,10,11</sup>, Alastair K. Denniston<sup>12,13,14</sup>, Livia Faes<sup>14</sup>, Bart Geerts<sup>15</sup>, Mudathir Ibrahim<sup>1,16</sup>, Xiaoxuan Liu<sup>12,13</sup>, Bilal A. Mateen<sup>8,17,18</sup>, Piyush Mathur<sup>19</sup>, Melissa D. McCradden<sup>20,21</sup>, Lauren Morgan<sup>22</sup>, Johan Ordish<sup>23</sup>, Campbell Rogers<sup>24</sup>, Suchi Saria<sup>25,26</sup>, Daniel S. W. Ting<sup>27,28</sup>, Peter Watkinson<sup>3,29</sup>, Wim Weber<sup>30</sup>, Peter Wheatstone<sup>31</sup>, Peter McCulloch<sup>1</sup> and the DECIDE-AI expert group\*

A growing number of artificial intelligence (AI)-based clinical decision support systems are showing promising performance in preclinical, in silico evaluation, but few have yet demonstrated real benefit to patient care. Early-stage clinical evaluation is important to assess an AI system's actual clinical performance at small scale, ensure its safety, evaluate the human factors surrounding its use and pave the way to further large-scale trials. However, the reporting of these early studies remains inadequate. The present statement provides a multi-stakeholder, consensus-based reporting guideline for the Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI). We conducted a two-round, modified Delphi process to collect and analyze expert opinion on the reporting of early clinical evaluation of AI systems. Experts were recruited from 20 pre-defined stakeholder categories. The final composition and wording of the guideline was determined at a virtual consensus meeting. The checklist and the Explanation & Elaboration (E&E) sections were refined based on feedback from a qualitative evaluation process. In total, 123 experts participated in the first round of Delphi, 138 in the second round, 16 in the consensus meeting and 16 in the qualitative evaluation. The DECIDE-AI reporting guideline comprises 17 AI-specific reporting items (made of 28 subitems) and ten generic reporting items, with an E&E paragraph provided for each. Through consultation and consensus with a range of stakeholders, we developed a guideline comprising key items that should be reported in early-stage clinical studies of AI-based decision support systems in healthcare. By providing an actionable checklist of minimal reporting items, the DECIDE-AI guideline will facilitate the appraisal of these studies and replicability of their findings.

Vasey et al, BMJ/Nat Med 2023



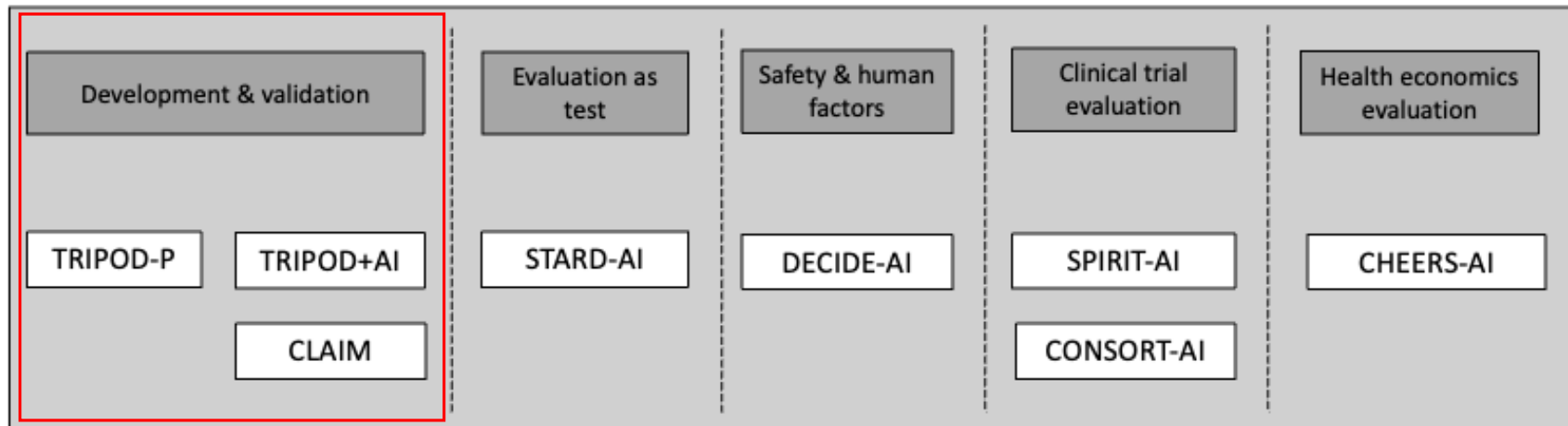
The **SPIRIT-AI** and **CONSORT-AI Working Group** is an international collaboration of methodologists, statisticians, healthcare professionals, computer scientists, industry representatives, journal editors, policy-makers, health informaticists, experts in law and ethics, regulators, patients and funders.

# SPIRIT-AI

- The **SPIRIT-AI extension** is a set of **recommendations for clinical trial protocols** evaluating interventions with an AI component.
- SPIRIT-AI includes 15 new items which should be routinely reported in addition to the core SPIRIT 2013 items.
- The checklist recommends that investigators provide
  - **clear descriptions of the AI intervention**
  - **prior evidence** supporting the validation of the AI intervention
  - the **proposed trial setting** in which the AI intervention will be evaluated
  - **specifying** how **the input and outputs** of the AI intervention will be handled
  - description of the **intended human-AI interaction** during the trial

# CONSORT-AI

- **The CONSORT-AI extension** is a set of recommendations for clinical trial reports evaluating interventions with an AI component.
- The checklist includes 14 new items, which were considered sufficiently important for AI interventions, that should be routinely reported in addition to the core CONSORT 2010 items
- CONSORT-AI recommends that investigators provide
  - a **clear description of the AI intervention**
  - including **instructions** and **skills required for use**
  - **handling of the input/output data** of the AI algorithm
  - the **human-AI interaction**
  - results of any **error cases analyses**



| 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 (forthcoming)                                 |
| 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, forthcoming); TREGAI - ethics for generative AI (Liu et al, arxiv 2013); CANGARU; responsible use, Cacciamani et al, forthcoming);





OPEN ACCESS

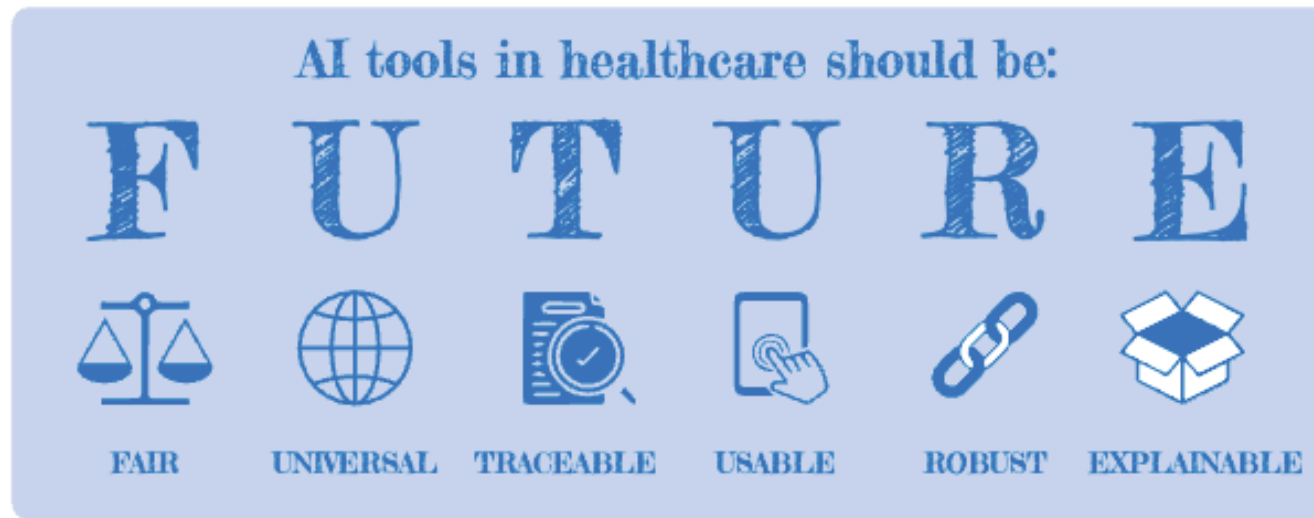


Check for updates

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

Karim Lekadir,<sup>1,2</sup> Alejandro F Frangi,<sup>3,4</sup> Antonio R Porras,<sup>5</sup> Ben Glocker,<sup>6</sup> Celia Cintas,<sup>7</sup> Curtis P Langlotz,<sup>8</sup> Eva Weicken,<sup>9</sup> Folkert W Asselbergs,<sup>10,11</sup> Fred Prior,<sup>12</sup> Gary S Collins,<sup>13</sup> Georgios Kaissis,<sup>14</sup> Gianna Tsakou,<sup>15</sup> Irène Buvat,<sup>16</sup> Jayashree Kalpathy-Cramer,<sup>17</sup> John Mongan,<sup>18</sup> Julia A Schnabel,<sup>19</sup> Kaiser Kushibar,<sup>1</sup> Katrine Riklund,<sup>20</sup> Kostas Marias,<sup>21</sup> Lameck M Amugongo,<sup>22</sup> Lauren A Fromont,<sup>23</sup> Lena Maier-Hein,<sup>24</sup> Leonor Cerdá-Alberich,<sup>25</sup> Luis Martí-Bonmatí,<sup>26</sup> M Jorge Cardoso,<sup>27</sup> Maciej Bobowicz,<sup>28</sup> Mahsa Shabani,<sup>29</sup> Manolis Tsiknakis,<sup>21</sup> Maria A Zuluaga,<sup>30</sup> Marie-Christine Fritzsche,<sup>31</sup> Marina Camacho,<sup>1</sup> Marius George Linguraru,<sup>32</sup> Markus Wenzel,<sup>9</sup> Marleen De Bruijne,<sup>33</sup> Martin G Tolsgaard,<sup>34</sup> Melanie Goisauf,<sup>35</sup> Mónica Cano Abadía,<sup>35</sup> Nikolaos Papanikolaou,<sup>36</sup> Noussair Lazrak,<sup>1</sup> Oriol Pujol,<sup>1</sup> Richard Osuala,<sup>1</sup> Sandy Napel,<sup>37</sup> Sara Colantonio,<sup>38</sup> Smriti Joshi,<sup>1</sup> Stefan Klein,<sup>33</sup> Susanna Aussó,<sup>39</sup> Wendy A Rogers,<sup>40</sup> Zohaib Salahuddin,<sup>41</sup> Martijn P A Starmans<sup>33</sup>; 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**
- AI has the potential to harm, create healthcare disparities or widen existing one
- Trustworthy AI needs thorough evaluation using high methodological standards, followed by complete & accurate reporting
- Lots of evidence that AI research is poorly designed, conducted and reported
- The use of tools like TRIPOD+AI, CLAIM-2024, 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