Reporting guidelines for prognosis research

Gary S. Collins

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

e-mail: gary.collins@csm.ox.ac.uk
twitter: @GSCollins

30-March-2022
Outline

- Importance of reporting

- Current status of reporting of clinical prediction models
  - describe some of the key deficiencies regularly seen in both model development and validation studies

- Consequences of poor reporting

- Initiatives to improve reporting: the TRIPOD Statement
  - and upcoming guidance
Reporting

Reporting guidelines: www.equator-network.org
Purpose of a research article

- **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 → research integrity**
“Altruism and trust lie at the heart of research on human subjects. Altruistic individuals volunteer for research because they trust that their participation will contribute to improved health […] In return for the altruism and trust that make clinical research possible, the research enterprise has an obligation to conduct research ethically and to report it honestly”

[International Committee of Medical Journal Editors, *CMAJ* 2004]
Research waste from poor reporting

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

Research publication can both communicate and miscommunicate. Unless research is adequately reported, the time and resources invested in the conduct of research is wasted. Reporting guidelines such as CONSORT, STARD, PRISMA, and ARRIVE aim to improve the quality of research reports, but all are much less adopted and adhered to than they should be. Adequate reports of research should clearly describe which questions were addressed and why, what was done, what was shown, and what the findings mean. However, substantial failures occur in each of these elements. For example, studies of published trial reports showed that the poor description of interventions meant that 40–89% were non-replicable; comparisons of protocols with publications showed that most studies had at least one primary outcome changed, introduced, or omitted; and investigators of new trials rarely set their findings in the context of a systematic review, and cited a very small and biased selection of previous relevant trials. Although best documented in reports of controlled trials, inadequate reporting occurs in all types of studies—animal and other preclinical studies, diagnostic studies, epidemiological studies, clinical prediction research, surveys, and qualitative studies. In this report, and in the Series more generally, we point to a waste at all stages in medical research. Although a more nuanced understanding of the complex systems involved in the conduct, writing, and publication of research is desirable, some immediate action can be taken to improve the reporting of research. Evidence for some recommendations is clear: change the current system of research rewards and regulations to encourage better and more complete reporting, and fund the development and maintenance of infrastructure to support better reporting, linkage, and archiving of all elements of research. However, the 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.
What should be reported?

Methods

- “Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results” [ICMJE]
- Same principle should extend to all study methods
- Allow repetition (in principle) if desired

Results

- Main findings (corresponding to a pre-specified plan)
- Should not be misleading
  - avoiding any (un)intentional spin or overinterpretation
Why is clear and transparent reporting important?

“If reporting is inadequate — namely, information is missing, incomplete or ambiguous — assumptions have to be made, and, as a result, important findings could be missed and not acted upon”

Prediction Models
What are prediction models?

- A single factor associated with an outcome has limited predictive information for individualized prediction

- Prediction is therefore typically a multivariable problem

- A prediction model combines multiple factors to yield an individualized prediction, typically using
  - Logistic regression (short term outcomes)
  - Cox regression (survival, long term outcomes) account for censoring
  - Increasingly data-driven approaches based on ‘machine learning’

- Used to guide
  - e.g., further testing, treatment/lifestyle changes and other clinical decisions, patient/clinician communication, selection of participants into studies, ...
**Diagnostic vs. Prognostic Model Studies**

- **Diagnostic**: examine the relationship of lab or imaging test results, signs & symptoms in relation to whether a particular disease is absent or present.

- **Prognostic**: examine future outcomes in individuals with a certain health profile (demographics, disease and individual characteristics).
- **QRISK (NICE CG181)**
  - 10-year risk of developing cardiovascular disease
- **Nottingham prognostic index (NICE CG80)**
  - risk of recurrence and overall survival in breast cancer patients
- **GRACE/PURSUIT/PREDICT/TIMI (NICE CG94)**
  - adverse CVD outcomes (mortality, MI, stroke etc...) for patients with UA/NSTEMI
- **APGAR (NICE CG132/2)**
  - evaluate the prognosis of a newborn baby
- **ABCD2 NICE CG68)**
  - Stroke / transient ischaemic attack
- **SAPS/APACHE (NICE CG50)**
  - ICU scoring systems for predicting mortality
- **Thoracoscore (NICE CG121)**
  - NSCLC pre-operative risk of death
- **CRB65/CURB65 (NICE CG191)**
  - Pneumonia
- **Blatchford / Rockall scores (NICE QS38)**
  - Upper gastrointestinal bleeding
- **FRAX / QFracture (NICE CG146)**
  - 10-year risk of developing osteoporotic & hip fracture
‘Prediction’ is a hot (and getting hotter) topic

PubMed search (09-December-2021)
Landscape of clinical prediction models

1382 models for cardiovascular disease (Wessler 2021)
408 models for COPD (Bellou 2019)
363 models for incident CVD (Damen 2016)
327 models for toxicity prediction after radiotherapy (Takada 2022)
263 models for obstetrics (Kleijnrouweleer 2016)
259 models for general trauma patients (De Munter 2016)
232 models for covid-19 (Wyants 2020)
222 models for neurodevelopment outcomes in preterm/VLBW children (Linsell 2016)
212 models for vascular surgery (Li 2022)
160 models for CVD models for women (Baart 2019)
142 models for preterm infant mortality (Van Beek 2021)
142 models for pregnancy care in primary care (Wingermuhle 2018)
137 models for dementia (Goerdten 2019)
129 models for neonatal mortality (Mangold 2021)
128 models for intracranial haemorrhage in ICU (Simon-Pimmel 2021)
119 models for critical care prognosis in LMIC (Haniffa 2018)
102 models for traumatic brain injury (Perel 2006)
101 models for gastric cancer (Feng 2019)
99 models for non-specific neck pain (Wingermuhle 2018)
91 models for psychosis transition (Studerus 2017)
87 models for diabetes complications (Tan 2021)
84 models for acute kidney injury (Song 2021)
83 models for ovarian malignant (Oomini 2009)
83 models for acute stroke (Counsell 2001)
83 models for colorectal cancer with surgical resection (He 2019)
81 models for sudden cardiac arrest (Carrick 2020)
77 models for orthopaedic surgical outcomes (Ogink 2021)
74 models for contrast-induced acute kidney injury (Allen 2017)
73 models for 28/30 day hospital readmission (Zhou 2016)
69 models for predicting falls in community-dwelling older adults (Gade 2021)
69 models for predicting stillbirth (Townsend 2020)
68 models for living donor/liver transplant counselling (Haller 2022)
68 models for pre-eclampsia (De Kat 2019)

67 models for moderate/severe traumatic brain injury (Dijkstra 2019)
66 models for predicting outcomes in men with prostate cancer following radiation therapy (Raymond 2017)
66 models for mortality/functional outcome follow ischemic stroke (Fahey 2018)
64 models for heart failure (Rahimi 2014)
64 models for suicide/suicide attempt (Belshere 2019)
64 models for nephropathy in type 2 diabetes (Stieker 2021)
61 models for dementia (Hou 2019)
58 models for oral health (Du 2020)
59 models for orthopaedic surgery (Grooth 2022)
58 models for breast cancer (Phung 2019)
58 models for heart failure (DI Tanna 2020)
54 models for prostate cancer patients undergoing radical prostatectomy (Campbell 2017)
53 models for short-term CABG mortality (Karim 2017)
53 models for colorectal cancer (Mahar 2017)
52 models for pre-eclampsia (Townsend 2019)
52 models for colorectal cancer (Usher-Smith 2015)
52 models for child/adolescent mental health (Senior 2021)
50 models for metastatic castration-resistant prostate cancer (Pinart et a 2018)
48 models for osteoporotic fracture (Rubin 2013)
48 models for incident hypertension (Son 2017)
47 models for oesophageal or gastric cancer (Van den Boorn 2018)
47 models for chronic kidney disease (Echourffo-Tcheugui 2012)
47 models for acute pancreatitis (Zhou 2022)
46 models for melanoma (Kaiser 2020)
46 models for carotid revascularisation (Volkers 2017)
45 models for CVD risk in type 2 diabetes (Van Dieren 2011)
45 models for surgical outcomes (Elfangely 2021)
43 models for hospital readmission (Van Grootveld 2021)
43 models for mortality in critically ill (Keuning 2019)
43 models for lung cancer (Wu 2022)
43 models for type 2 diabetes (Collins 2011)
42 models for chronic diseases (Delipino 2022)
41 models for mortality in very premature infants (Medlock 2011)

+ many many more
Reporting & Prediction Models
Prognosis Studies and reporting guidelines

- **Prognostic factor studies** - which predictors contribute (associated with) to prediction of particular prognostic/diagnostic outcome – aim not to develop a model for individualized predictions.

- **Model development studies** – to develop a prediction model from data: identify important predictors; estimate predictor weights; construct a model for individualized predictions; quantify predictive performance; internal validation.

- **Model validation studies** – evaluate (validate) predictive performance of previously developed model in participant data other than the development set.

- **Model impact studies** – quantify effect/impact actually using model on participant/physician management and health outcomes – relative to not using the model -> comparative studies.

* Currently in the early stages of being updated/scope broadened

** Tailored guidance for AI; SPIRIT-AI (Protocols); CONSORT-AI (reports)
Reporting of prognostic model research

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

- **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 (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 discrimination**

* Bellou et al, BMJ 2019
Findings from multiple systematic reviews

- Poor reporting & poor methodological conduct
- Number of events often difficult to identify
  - candidate predictors (and number) not always easy to find
- How candidate predictors were selected
  - unclear in: 25% studies (Bouwmeester 2012); 69% studies (Haller 2022)
- How the multivariable model was derived
  - unclear in 77% of studies in cancer (Mallet 2010)
Findings from systematic reviews

- **Missing data rarely mentioned**
  - 41% Collins 2010; 45% Collins 2012; 64% Bellou
  - often an exclusion criteria (though often not specified)
  - complete-case usually carried out

- **Range of continuous predictors rarely reported**
  - ...and coding of binary/categorical predictors
  - applying a model ‘off-label’ – outside the range of a continuous predictor

- **Models often not reported in full (nor a link to any code)**
  - intercept missing (logistic regression); baseline survival missing (cox regression)
  - why build a model and not provide sufficient information for others to use it, including evaluating it on other data?

Either code to implement the model to get predictions for an individual or analysis code
Other conclusions from systematic reviews

- **Methodological shortcomings include**
  - large number of candidate predictors
  - small sample size (number of events)
  - calibration rarely assessed (and often done poorly, e.g., Hosmer-Lemeshow test)
    - not done in 85% studies (Altman: cancer); 74% (Collins: diabetes); 46% (Bouwmeester: general medical journals); 87% (He, colorectal cancer)
  - dichotomisation / categorisation of continuous predictors
    - 63% studies (Collins: diabetes); 70% studies (Mallet: cancer)
  - previously published models often ignored - waste?
  - inadequate or no validation
    - reliance on (inefficient) random-split to validate
  - **overfitting**

- **Lack of comparing competing models** (Collins & Moons BMJ 2014)
  - is the newly developed model better than any other models?

- **Unsurprisingly (and fortunately) very few models are used**
External validation studies*

- 16% of studies failed to cite the original article developing the model (N.B. >360 models for incident CVD)

- 60% of studies failed to make/discuss any case-mix comparison
  - Or discussion on the representativeness of the target population

- Tend to be small (few events, if reported at all) (48% < 100 events)

- Missing data rarely mentioned (54%)
  - 64% implicitly/explicitly conducted complete-case analyses
    - Loss of information and impact of representativeness
  - 9% used multiple imputation

- Overwhelming focus only on discrimination
  - 73% of external validation studies evaluated discrimination
  - only 32% assessed calibration (often incorrectly/weakly)
  - 24% presented ‘blank’ ROC curves (i.e., no cut-points labelled)
    - (see Verbakel et al J Clin Epidemiol 2020 and discussion with Janssens 2020)

*Collins et al BMC MRM 2014
"...substantial deficits in the reporting of risk prediction models for depression, with major concerns regarding the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis statement."

Twitter: @GSCollins
Not a fan of ROC curves (for model evaluation) at the best of times (see why here -> bit.ly/3GYjcaX)

 Почему присутствуют две разные ROC кривые в одной и той же статье и были изменены осями между двумя кривыми?

 BTW: No calibration curve presented. #sigh
Poor calibration (from weak modelling) \rightarrow misleading conclusions (spin)

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

* Zhou et al, J Dermatol 2021
Clear(ish) reporting, poor methods

Materials and Methods

Patient Eligibility

This retrospective study was approved by the Ethics Committee of Chang Gung Memorial Hospital (IRB no. 104-4097B). Patient records were anonymised and de-identified prior to the analysis. We included 21,614 (9710 men and 11,904 women) apparently asymptomatic individuals who had at least once voluntarily undergone an out-of-pocket tumour marker panel test between March 2003 and December 2012 consecutively at the Linkou branch of Chang Gung Memorial Hospital [2]. We excluded malignancies. All eligible indivi
duals were randomised to the training set. Moreover, for the training set, random
deresampling was applied [12–14] because of the extremely unbalanced data set used in this study. A cancer to noncancer ratio of 1:1 was adopted to randomise 67 individuals from the 6128 noncancer cases to the final training set. Consequently, the training set, which comprised 67 cases of newly diagnosed cancer and 67 noncancer cases, was used to train the machine learning models. For the women, 116 cases (58 newly diagnosed cancer cases and 58 noncancer cases) were randomised to the training set. In addition, one-third of all individuals were randomly allocated to the validation set to test the performance of the constructed models. The validation sets comprised 3097 cases (33 cases of newly diagnosed cancer and 3064 noncancer cases) for men and 3801 cases (29 cases of newly diagnosed cancer and 3772 noncancer cases) for women. The tumour types of occult cancer cases were also listed in the training and validation sets.
TRIPOD Statement

- **Consensus-based guidance for improving the quality of reporting of multivariable prediction model studies**
  - led by Collins, Moons, Altman, Reitsma
  - 21 experts (statisticians, epidemiologists, clinicians, journals editors)
    - Delphi survey, 3-day meeting in 2011

- **Focus on reporting**
  - but considerable attention on (highlighting good and bad) methodological conduct in the *Explanation & Elaboration* paper

- **Funded by Cancer Research UK, ZonMW, Medical Research Council, NIHR**
TRIPOD Statement

- Published simultaneously in 11 leading general and specialty journals (January 2015)
  - Ann Intern Med; BJOG; BMC Med; BMJ; Br J Cancer; Br J Surgery; Circulation; Diabet Med; Eur J Clin Invest; Eur Urol; J Clin Epidemiol
  - Editorials/comments in other journals
    - e.g., Am J Kidney Dis; Sci Transl Med; Clin Chem

- Guidance for authors, reviewers, editors and readers

- Checklist

- Explanation & Elaboration paper
  - Rationale; examples of good reporting; methodology summaries; 532 references
TRIPOD Statement

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

Predictive 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 the 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).

Editors’ Note: In order to encourage dissemination of the TRIPOD Statement, this article is freely accessible on the Annals of Internal Medicine Web site (www.annals.org) and will be also published in BJOG, British Journal of Obstetrics, British Journal of Surgery, BMC Medicine, BMJ, CMAJ, Current Epidemiology, Current Medical Education, Evidence-Based Medicine, Journal of Clinical Epidemiology, Journal of General Internal Medicine, Journal of the Royal Society of Medicine, and Multiple Sclerosis Journal. The TRIPOD Initiative is a multi-disciplinary and multi-national collaboration of individuals interested in improving the quality of reporting of prediction model studies. The TRIPOD Initiative is supported by The Collaboration for Diagnostic and Prognostic Research (CDPRO). For more information, visit www.tripod-statement.org.

In medicine, numerous decisions are made by care providers, often in shared decision making, on the basis of an estimated probability that a specific disease or condition is present (diagnostic setting) or a specific event will occur in the future (prognostic setting). In an individual, in the diagnostic setting, the probability that a disease or condition is present is influenced by the test result; in the prognostic setting, the probability that a future event will occur is influenced by the current status. Predictors are also referred to as covariates, risk indicators, prognostic factors, determinants, test results, or—even statistically—dependent variables. They may range from demographic characteristics (for example, age and sex), medical history-taking, and physical examination results to results from imaging, electrophysiological, blood, and other body fluid analyses.
Reporting guideline checklists

- Reminders of scientific content (like shopping lists)

- TRIPOD Reporting Checklist
  - Title & Abstract
  - Introduction
    - Background & Objectives
  - Methods
    - source of data, participants, outcomes, predictors
    - sample size, missing data
    - statistical analysis methods, risk groups
  - Results
    - participants
    - model development, specification, performance
  - Discussion
    - limitations, interpretation, implications
  - Other Information
    - supplementary information, funding
37 items covering
22 ‘topics’ that should be included on an articles describing the development or validation of a prediction model

D -> applies to development studies only
V -> applies to validation studies only

D; V -> applies to both development and validation studies
Pre-TRIPOD era: adherence to TRIPOD*

Specify type of model and all model building steps

Report performance measures with CIs

Participant characteristics

* Heus et al BMJ Med 2018
Pre (‘12-‘14) and post TRIPOD (‘16-‘17)*

- No discernible improvement in reporting (yet…)

- But improvements in assessment of model performance
  - e.g., Calibration (21% vs 87%)

- Handling of missing data,
  - e.g., multiple imputation (12% versus 50%)

- Limitations: Small sample size, short post TRIPOD time frame

* Najafabadi et al BMJ Open 2020
New guidance in preparation

- **TRIPOD-Cluster** [led by Thomas Debray/Carl Moons; UMC Utrecht]
  - Studies developing/validating models using ‘clustered’ data
    - (Large) multicentre data (e.g., cluster = centre/hospital)
    - Individual Participant Data from multiple studies (cluster = study)

- **TRIPOD-SRMA** [led by Kym Snell/Richard Riley, Keele]
  - Systematic reviews/meta-analysis of prediction model studies

- **TRIPOD-AI** [led by Collins (Oxford); Moons (Utrecht)]
  - Studies developing/validating models using machine learning

- **TRIPOD-P** [led by Paula Dhiman/Collins, Oxford]
  - Protocols for studies developing/validation prediction models

Reporting and critical appraisal

- **Evaluating the study methods / results** is a core component of evidence-based medicine
  - An important skill for any researcher

- **Risk of bias tools** attempt to assess (and rate) the study methods in a structured manner
  - Enables us to judge the study methods and interpret the findings accordingly

- **Poor reporting** makes risk of bias assessment more difficult
  - Rating will often be ‘unclear’
Prognosis Studies and risk of bias

**Prognostic factor studies** - which predictors contribute to prediction of particular prognostic/diagnostic outcome – aim not to develop a model for individualised predictions

**Model development studies** – to develop prediction model from data: identify important predictors; estimate predictor weights; construct model for individualised predictions; quantify predictive performance; internal validation

**Model validation studies** – test (validate) predictive performance of previously developed model in participant data other than development set

**Model impact studies** – quantify effect/impact actually using model on participant/physician management and health outcomes – relative to not using the model -> comparative studies.

*Wolff et al Ann Intern Med 2019*
"We advise against applying poorly developed, reported, or validated prediction models “

* Haller et al J Clin Epidemiol 2022
Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review

Constanza L. Andauro Navarro1,2,2,a, Johanna A. A. Damen1,2,2,b, Toshihiko Takada3, Steven W. J Nijman4, Paula Dhiman2,2,a, Jie Ma2,2,b, Gary S. Collins5,2,b, Ram Bajpai6,2,b, Richard D. Riley2,2,b, Karel G. M. Moons1,2,2,b and Lotty Hooft2,2,b

Abstract

Background: While many studies have consistently found incomplete reporting of regression-based prediction model studies, evidence is lacking for machine learning-based prediction model studies. We aim to systematically review the adherence of Machine Learning (ML)-based prediction model studies to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement.

Methodology: We conducted a systematic review to identify inclusion criteria for ML-based prediction model studies, and used the TRIPOD statement to evaluate the adherence of these studies. We identified 21 studies that met our inclusion criteria, and found a low level of adherence to the TRIPOD statement.

Results: The studies included a variety of ML algorithms, such as random forests, support vector machines, and neural networks. The adherence to the TRIPOD statement was low, with many studies lacking important information such as the model's performance metrics and the study's limitations.

Conclusion: More research is needed to improve the reporting of ML-based prediction models. We recommend that future studies adhere to the TRIPOD statement to ensure the transparency and reproducibility of these models.

Keywords: Machine Learning, Prediction models, TRIPOD Statement, Reporting, Systematic review.
Adherence to TRIPOD

Fig. 2. Overall adherence per TRIPOD item. Overall sample n=112.
"Most of the issues discussed here could be avoided through more robust designs & high-quality reporting, although several hurdles must be overcome before<br>deep learning<br>breast & cervical

We recommend [...] greater emphasis on standardized adoption of existing & developing guidelines for reporting of certain key methodological & model presentation criteria was inadequate. [...] lack of consistent adherence to reporting guidelines"

tinyurl.com/bdft4w9r

Adhere to @TRIPODStatement tinyurl.com/3bva4h

#DataScience #machine


#MachineLearning in vascular surgery: a systematic review & critical appraisal tinyurl.com/aufcfwz2

- use @TRIPODStatement to report your prediction model study tinyurl.com/3pmpmfyh

TRIPOD Guidance for #ML underway tinyurl.com/4t5ujakp

#statstwitter #ml4h #mltwitter
“Most studies on prediction models developed using machine learning show poor methodological quality and are at high risk of bias"
Machine learning studies

- **Beware of the hype**
  - Reported performance is often too good to be true

- **Often little or no difference in performance in (typically) noisy (low-signal-to-noise) health care problems**
  - Clear benefits in high signal-to-noise settings (e.g., imaging)

- **Need the same robust development and evaluation of non-machine learning studies (principally the same)**
  - Some very good studies but *many* poor studies
    - as there are *many* poor statistical based prediction model studies

- **Need complete and transparent reporting**
  - TRIPOD is relevant though updated and tailored guidance is underway (checklist/preprint in summer 2022🚀)
    - Collins & Moons Lancet 2019; Collins et al BMJ Open 2021 for protocol
TRIPOD-AI challenge: model availability

Models based on regression can typically be written down—Regression coefficients + intercept/baseline survival—Allows independent researchers to validate and recalibrate (to their setting)

ML are typically 'blackbox'—We can't write down a Random Forest

How can independent researchers evaluate these models?

Impact on reproducibility

Issues of proprietary

• Protecting scientific innovation

• Commercial exploitation

Artificial Intelligence Algorithms for Medical Prediction Should Be Nonproprietary and Readily Available

To the Editor Wang and colleagues describe the challenges that arise for deep learning and other black-box machine learning algorithms for medical prediction. The authors rightfully hint at the fact that reliable performance of predictive analytics in health care is far from guaranteed by discussing data quantity, data quality, model generalizability and interoperability.

Ben Van Calster, PhD
Ewout W. Steyerberg, PhD
Gary S. Collins, PhD

Author Affiliations: Department of Development and Regeneration, KU Leuven, Leuven, Belgium (Van Calster); Department of Biomedical Data Sciences, Leiden University Medical Center (LUMC), Leiden, the Netherlands (Van Calster, Steyerberg); Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom (Collins); Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom (Collins).
e.g.,

- Make it available on a repository (e.g., GitHub)
- Grant access to get predictions for your data set
- Gain access to the code by setting-up non-disclosure agreements

Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist

Here we present the MI-CLAIM checklist, a tool intended to improve transparent reporting of AI algorithms in medicine.


The application of artificial intelligence (AI) in medicine is an old idea — but methods for data in the past involved programming computers with precise rules or rules extracted from human experts, which resulted in deterministic, rule-based systems. The study of AI in medicine has grown tremendously in the past few years due to increasingly available datasets from medical practice, including clinical images, genetics, and electronic health records, as well as the maturity of methods that use data to teach computers. The use of data labeled by clinical experts to teach machine learning algorithms, probabilities, and statistical models is called supervised machine learning. Successful use of these new machine learning approaches include targeted real-time early warning systems for adverse events, the detection of diagnostic pathology, the classification of pathology and other images, the prediction of the near-term future state of patients with chronic medical problems, patient discharge disposition, and more.

Table 2 | Frameworks to share code, software dependencies and deep-learning models

<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
</tr>
</thead>
</table>

Reproducibility (Part 6): choose appropriate tier of transparency

- Tier 1: complete sharing of the code
- Tier 2: allow a third party to evaluate the code for accuracy/fairness; share the results of this evaluation
- Tier 3: release of a virtual machine (binary) for running the code on new data without sharing its details
- Tier 4: no sharing
Reporting, code, data and the potential for scientific fraud

Consider the following hypothetical scenario...

- **A model has been developed**
  - maybe multiple models for comparison (RF, LR, ANN, SVM, XGBoost)

- **A paper has been published describing their development**

- **None of the models are presented in the paper**

- **The models (and data) are not made available in a software repository (e.g., via Github)**

- **Table of ‘AUC’s is reported**
  - the paper concludes (with associated ‘spin’) one of more models as having excellent predictive accuracy

- **The paper is published**
Some examples
Prediction models: An opportunity to take centre(ish) stage, but…

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

Laure Wynants,1,2 Ben Van Calster,2,3 Gary S Collins,4,5 Richard D Riley,6 Georg Heinze,7 Ewoud Schuit,8,9 Marc M J Bonten,8,10 Darren L Dahly,11,12 Johanna A A Damen,8,9 Thomas P A Debray,8,9 Valentijn M T de Jong,8,9 Maarten De Vos,2,13 Paula Dhiman,4,5 Maria C Haller,7,14 Michael O Harhay,15,16 Liesbet Henckaerts,17,18 Pauline Heus,8,9 Nina Kreuzberger,19 Anna Lohmann,20 Kim Luijken,20 Jie Ma,19 Glen P Martin,21 Constanza L Andaur Navarro,8,9 Johannes B Reitsma,8,9 Jamie C Sergeant,22,23 Chunhu Shi,24 Nicole Skoetz,19 Luc JM Smits,1 Kym I E Snell,9 Matthew Sperrin,25 René Spijkert,8,9,26 Ewout W Steyerberg,7 Toshihiko Takada,8 Ioanna Tzoulaki,27,28 Sander M J van Kuijk,29 Florien S van Royen,8 Jan Y Verbakel30,31 Christine Wallisch,7,32,33 Jack Wilkinson,21 Robert Wolff,14 Lotty Hooft,8,9 Karel G M Moons,8,9 Maarten van Smeden8

ABSTRACT

OBJECTIVE
To review and appraise the validity and usefulness of published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of becoming infected with covid-19 or being admitted to hospital with the disease.

STUDY SELECTION
Studies that developed or validated a multivariable covid-19 related prediction model.

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

- **169 studies describing 232 prediction models**
  - 7 risk scores, 118 diagnostic; 107 prognostic
  - Mixture of modelling procedures

- **Reported c-index values ranged from**
  - 0.71 to 0.99 (risk scores)
  - 0.65 to 0.99 (diagnostic models)
  - 0.54 to 0.99 (prognostic models)

- **Calibration rarely assessedreported (and often incorrectly)**

- **Table of participant characteristics sometimes missing**

- **“This review indicates that almost all published prediction models are poorly reported”**

- **Bottom line: 226 at high risk of bias; 6 at unclear risk of bias**

** Latest update (forthcoming) now includes >500 models
Risk of bias assessment
COVID Example 1* (generally poor)

- **Sample size for development (after splitting data into train/test)**
  - 239 individuals, 57 events for model development with 75 predictors
  - Using sample size formula (pmsampsize) indicates 1285 individuals and 306 events were actually required. **No sample size calculation in the paper reported.**

- **Sample size for testing**
  - 60 individuals with ~14 events (**not reported**)

- **Overfitting not addressed neither adjusting performance for optimism or shrinkage of regression coefficients**

- **Weak / flawed assessment of calibration**
  - e.g., Hosmer-Lemeshow test, didn’t present calibration plot

- **No mention of missing data**
  - presumably an unspecified exclusion criteria
  - yet 75 predictors examined

- **Assumption of linearity of the continuous predictors**

- **No model reported (just a nomogram)**
  - e.g., no intercept/regression coefficients

---

**Conclusion**

“The machine-learning model, nomogram, and online-calculator might be useful to assess the onset of severe and critical illness among COVID-19 patients and triage at hospital admission”

Small validation sample size → misleading conclusions

Sample size:
- n=279
- Number of outcome events= 7

No calibration

Red flag – should’ve been picked up during editorial process / peer review
232 #covid clinical prediction models (up until July 2020) have rated (generally, with some exceptions) to be at high risk of bias (see tinyurl.com/upyxf6s)

Hundreds of models later, things aren't getting much better - this from today:

#statstwitter #mltwitter #epitwitter
Summary

- Prediction models increasingly seen as useful tools for identifying individuals at increased risk => target treatments / interventions
  - increasingly recommended in clinical guidelines

- Many components to prediction model study (study design, missing data, continuous predictors, model evaluation) – easy to get one or more of these ‘wrong’

- Prediction model studies are often done badly and poorly reported (including ‘spin’)
  - Obvious flaws in poor reporting often go unmissed during peer review -> plethora of poorly developed/reported (potentially harmful) models

- **TRIPOD Statement** available to help authors, reviewers and editors to help with full and transparent reporting (important for PROBAST* risk of bias assessment)
  - New reporting guidelines for machine learning (TRIPOD-AI), systematic reviews (TRIPOD-SRMA) and protocols (TRIPOD-P) in preparation

Thank you for listening

www: www.tripod-statement.org
twitter: @TRIPODStatement

Risk of bias: www.probast.org

www.prognosisresearch.com

Journal: BMC Diagnostic & Prognostic Research
https://diagnprognres.biomedcentral.com

Topic Group 6 (prediction models): www.stratos-initiative.org

Reporting guidelines: www.equator-network.org
@EQUATORNetwork