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**

Obligation

"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

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

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.

Lancet 2014; 383: 267-76

Published Online January 8, 2014 http://dx.doi.org/10.1016/S0140-6736(13)62228-X

See Perspectives page 209

This is the fifth in a Series of five papers about research

Centre for Research in Evidence Based Practice, Bond University, Robina, QLD, Australia (Prof P Glasziou FRACGP); Centre for Statistics in Medicine, University of Oxford, Oxford, UK (Prof D G Altman DSc); Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands (Prof P Bossuyt PhD); INSERM,

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"

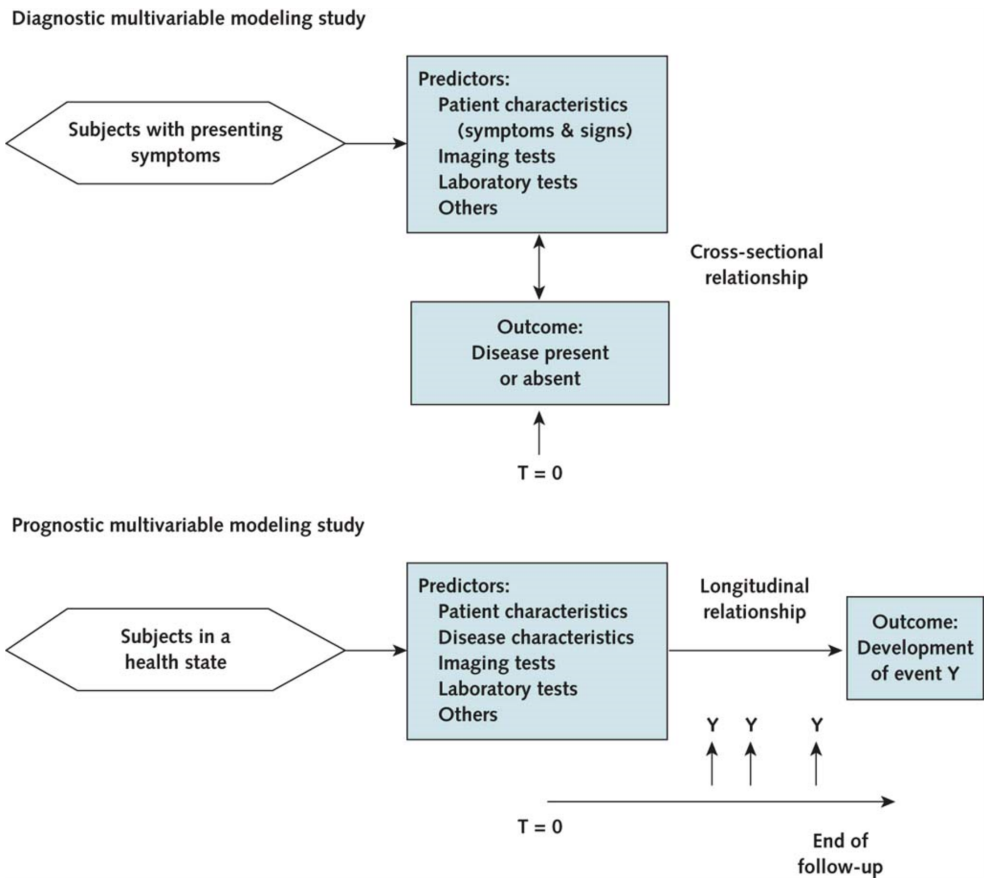
Needleman et al, *J Dent Res* 2008

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



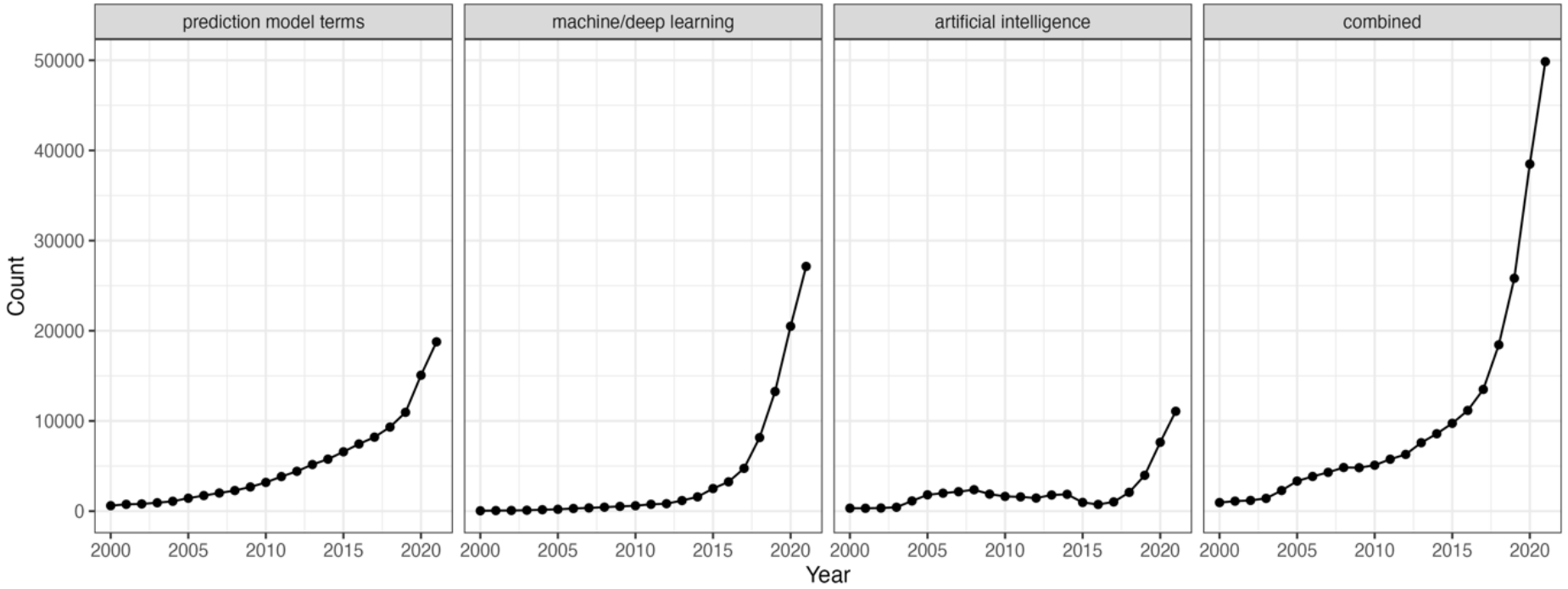
- **Diagnostic:** examine the relationship of lab or imaging test results, signs & symptoms in relation to whether a particular disease is absent or present

- **Prognostic:** examine future outcomes in individuals with a certain health profile (demographics, disease and individual characteristics)

(UK) NICE Clinical Guidelines

- **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 (Kleinrouweler 2016)
258 models for general trauma patients (De Munter 2016)
232 models for covid-19 (Wynants 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 (Wingbermuhle 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 (Wingbermuhle 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 malignancy (Geomini 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 (Dijkland 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 (Belsher 2019)
64 models for nephropathy in type 2 diabetes (Slieker 2021)
61 models for dementia (Hou 2019)
58 models for oral health (Du 2020)
59 models for orthopaedic surgery (Groot 2022)
58 models for breast cancer (Phung 2019)
58 models for heart failure (Di Tanna 2020)
54 models for prostate cancer patients undergoing radical proctectomy (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 20180
48 models for osteoporotic fracture (Rubin 2013)
48 models for incident hypertension (Sun 2017)
47 models for oesophageal or gastric cancer (Van den Boorn 2018)
47 models for chronic kidney disease (Echouffo-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 Grootven 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 (Delpino 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** - whi… with) to prediction of particular pr… to develop a model for individuali… — REMARK Statement*
- **Model development studies** – to develop a prediction model from… data: identify important predictors… model for individualized prediction… internal validation — TRIPOD Statement
- **Model validation studies** – eva… performance of previously develo… development set — TRIPOD Statement
- **Model impact studies** – quantify effect/impact actually using model on… participant/physician managemen… using the model -> comparative s… — CONSORT Statement**

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

** Tailored guidance for AI; SPIRIT-AI (Protocols); CONSORT-AI (reports)

SPIRIT-AI

CONSORT-AI

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)

  → overfitting
  - 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

- **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

Today's prediction model review: "There were several

Another recent one from psychiatry: Only 1 of the models was found to be at low risk of bias & the

Another review of prediction models - this time for bladder/kidney cancer bit.ly/31El5JG

"...prognostic prediction models for #COVID19 were evaluated according to the @TRIPODStatement; we found the reporting completeness to be poor " (tinyurl.com/2ax2nxa5) - which tallies with earlier findings critically appraising covid prediction models (tinyurl.com/2vn3s4ub)

Review Article

Page 1 of 13

Reporting of coronavirus disease 2019 prognostic models: the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis statement

Liuqing Yang[1,2], Qiang Wang[1,2], Tingting Cui[1,2], Jinxin Huang[1,2], Naiyang Shi[1,2], Hui Jin[1,2]

[1]Department of Epidemiology and Health Statistics, School of Public Health, Southeast University, Nanjing, China; [2]Key Laboratory o Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing, China

Contributions: (I) Conception and design: L Yang, H Jin; (II) Administrative support: H Jin; (III) Provision of study materials or patients: H Jin, Q Wang, T Cui; (IV) Collection and assembly of data: T Cui, Q Wang, J Huang; (V) Data analysis and interpretation: L Yang, N Shi, J Huang; (VI Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

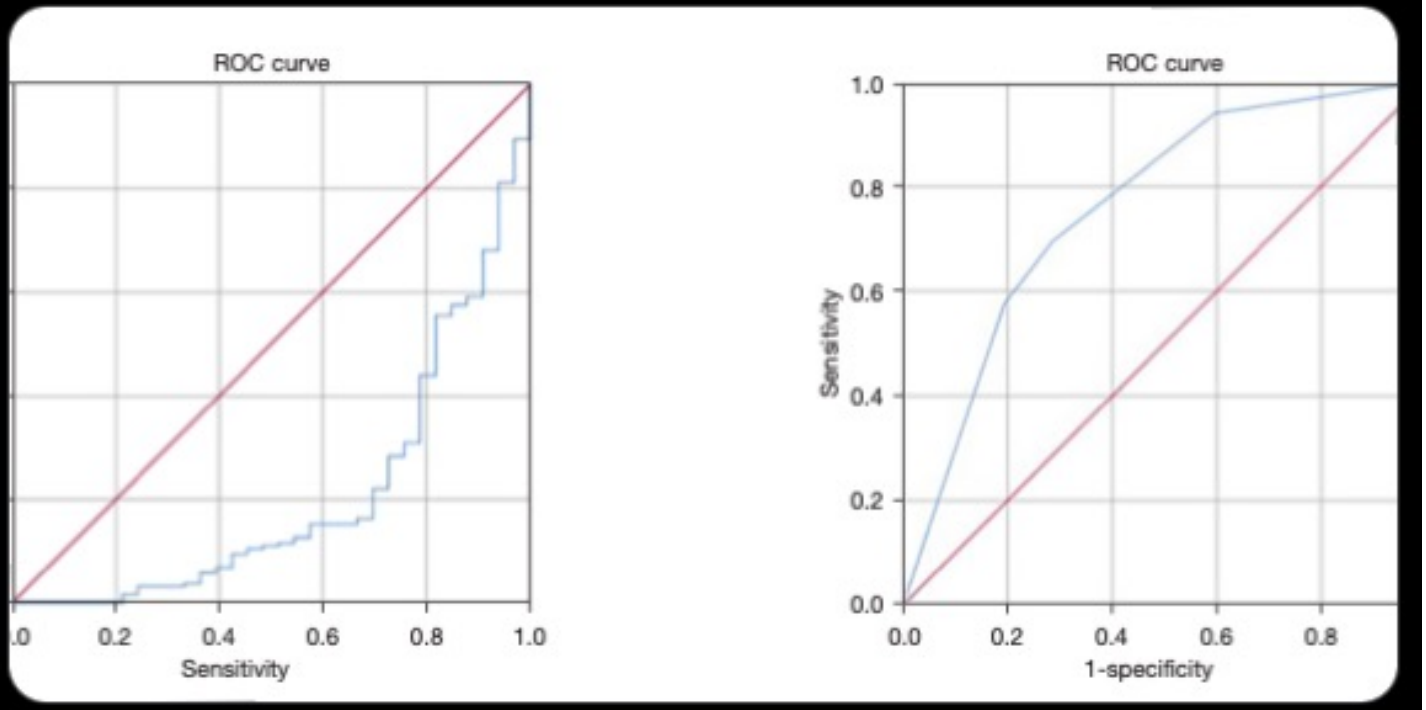
Correspondence to: Hui Jin. Department of Epidemiology and Health Statistics, School of Public Health, Southeast University, 87# Dingjiaqiao Nanjing 210009, China. Email: jinhui_hld@163.com.

Twitter: @GSCollins

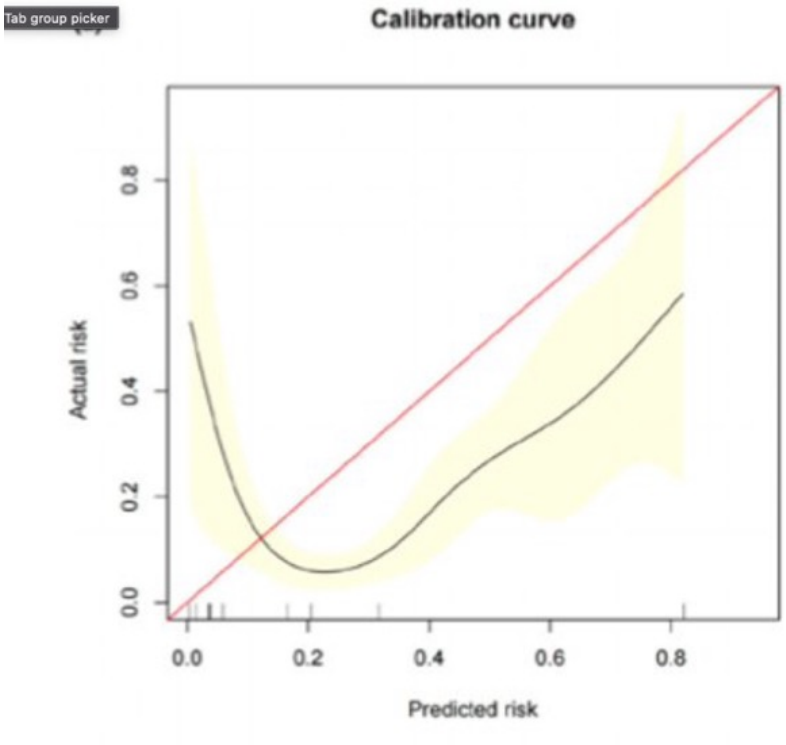
Not a fan of ROC curves (for model evaluation) at the best of times (see why here -> bit.ly/3GYjcaX)

👇...but why present two different ROC curves side-by-side in the same paper and switch the axes between the two curves?

BTW: No calibration curve presented. #sigh



Poor calibration (from weak modelling) → misleading conclusions (spin)



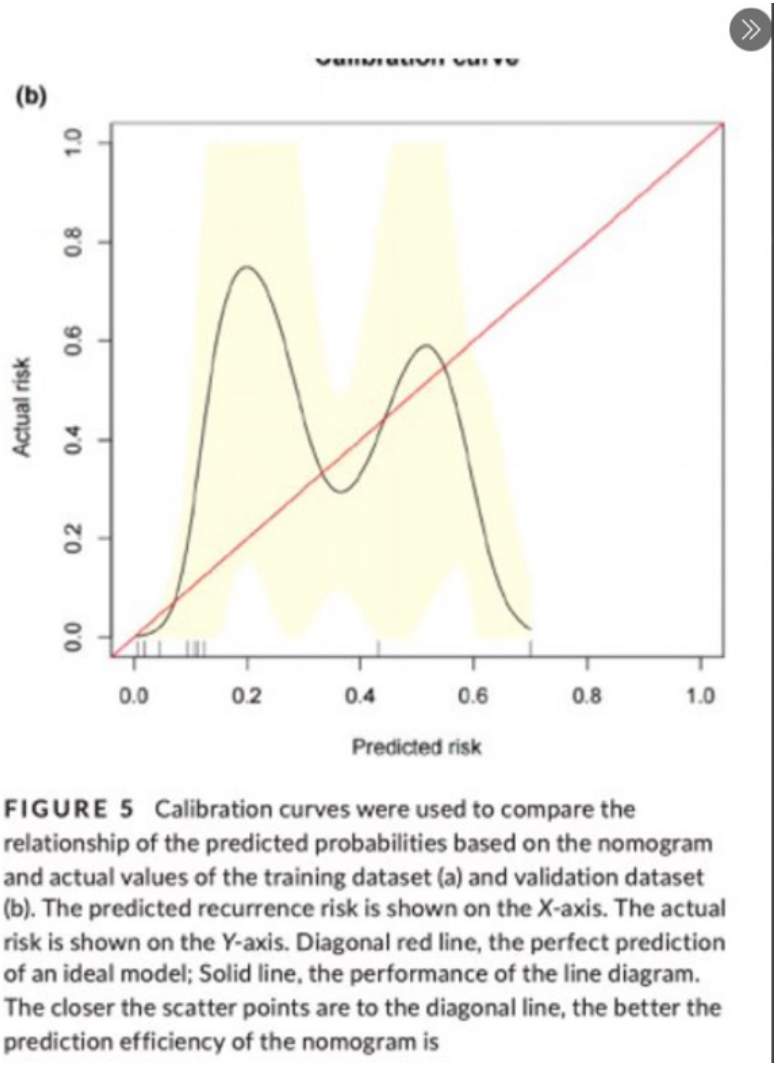


FIGURE 5 Calibration curves were used to compare the relationship of the predicted probabilities based on the nomogram and actual values of the training dataset (a) and validation dataset (b). The predicted recurrence risk is shown on the *X*-axis. The actual risk is shown on the *Y*-axis. Diagonal red line, the perfect prediction of an ideal model; Solid line, the performance of the line diagram. The closer the scatter points are to the diagonal line, the better the prediction efficiency of the nomogram is

"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 excl
malignancies. All eligible indivi
tumour markers (AFP, CEA, C.
markers (AFP, CEA, CA19-9, C
CEA, CA19-9, SCC, PSA, CA1

Subsequently, a ratio of 2:1 (training to validation) was used to randomly allocate individuals to the training or validation set. All randomisations were performed using Matlab (MathWorks, Natick, MA, USA). For the men, 67 cases of newly diagnosed cancer and 6128 noncancer cases were randomised to the training set. Moreover, for the training set, random undersampling 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

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

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

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

Prediction models are developed to aid health care providers in estimating the probability or risk that a specific disease or condition is present (diagnostic models) or that a specific event will occur in the future (prognostic models), to inform their decision making. However, the overwhelming evidence shows that the quality of reporting of prediction model studies is poor. Only with full and clear reporting of information on all aspects of a prediction model can risk of bias and potential usefulness of prediction models be adequately assessed. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative developed a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. This article describes how the TRIPOD Statement was developed. An extensive list of items based on a review of the literature was created, which was reduced after a Web-based survey and revised during a 3-day meeting in June 2011 with methodologists, health care professionals, and journal editors. The list was refined during several meetings of the steering group and in e-mail discussions with the wider group of TRIPOD contributors. The resulting TRIPOD Statement is a checklist of 22 items, deemed essential for transparent reporting of a prediction model study. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. The TRIPOD Statement is best used in conjunction with the TRIPOD explanation and elaboration document. To aid the editorial process and readers of prediction model studies, it is recommended that authors include a completed checklist in their submission (also available at www.tripod-statement.org).

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

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 Cancer, British Journal of Surgery, BMC Medicine,

nostic studies). Prediction is therefore inherently multivariable. Prediction models (also commonly called "prognostic models," "risk scores," or "prediction rules" [6]) are tools that combine multiple predictors by assigning relative weights to each predictor to obtain a

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

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

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

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

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

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

Predictors are also referred to as *covariates*, *risk indicators*, *prognostic factors*, *determinants*, *test results*, or—more statistically—*independent variables*. They may range from demographic characteristics (for example, age and sex), medical history-taking, and physical examination results to results from imaging, electrophys-

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

equator network

Section / Topic			Checklist item
Title and abstract			
Title	1	D,V	Identify the study as developing and/or validating a multivariable predic population, and the outcome to be predicted.
Abstract	2	D,V	Provide a summary of objectives, study design, setting, participants, sa statistical analysis, results, and conclusions.
Introduction			
Background & Objectives	3a	D,V	Explain the medical context (including whether diagnostic or prognostic or validating the multivariable prediction model, including references to
	3b	D,V	Specify the objectives, including whether the study describes the develo model or both.
Methods			
Source of data	4a	D,V	Describe the study design or source of data (e.g. randomised trial, coho for the development and validation datasets, if applicable.
	4b	D,V	Specify the key study dates, including start of accrual, end of accrual, a up.
Participants	5a	D,V	Specify key elements of the study setting (e.g. primary care, secondary including number and location of centres.
	5b	D,V	Describe eligibility criteria for participants.
	5c	D,V	Give details of treatments received, if relevant.
Outcome	6a	D,V	Clearly define the outcome that is predicted by the prediction model, inc assessed.
	6b	D,V	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	D,V	Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured.
	7b	D,V	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	D,V	Explain how the study size was arrived at.
Missing data	9	D,V	Describe how missing data were handled (e.g. complete-case analysis, imputation) with details of any imputation method.
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.
	10b	D	Specify type of model, all model-building procedures (including any pred for internal validation.
	10c	V	For validation, describe how the predictions were calculated.
	10d	D,V	Specify all measures used to assess model performance and, if relevant
	10e	V	Describe any model-updating (e.g. recalibration) arising from the valida
Risk groups	11	D,V	Provide details on how risk groups were created, if done.
Development versus validation	12	V	For validation, identify any differences from the development data in se outcome and predictors.
Results			
Participants	13a	D,V	Describe the flow of participants through the study, including the numbe without the outcome and, if applicable, a summary of the follow-up ti
	13b	D,V	Describe the characteristics of the participants (basic demographics, clin predictors), including the number of participants with missing data for p
	13c	V	For validation, show a comparison with the development data of the dis (demographics, predictors and outcome).
Model development	14a	D	Specify the number of participants and outcome events in each analysis
	14b	D	If done, report the unadjusted association between each candidate pred
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e. model intercept or baseline survival at a given time point).
	15b	D	Explain how to use the prediction model.
Model performance	16	D,V	Report performance measures (with confidence intervals) for the predic
Model-updating	17	V	If done, report the results from any model-updating (i.e. Model specifica
Discussion			
Limitations	18	D,V	Discuss any limitations of the study (such as non-representative sample missing data).
Interpretation	19a	V	For validation, discuss the results with reference to performance in the other validation data.
	19b	D,V	Give an overall interpretation of the results considering objectives, limita studies and other relevant evidence.
Implications	20	D,V	Discuss the potential clinical use of the model and implications for futur
Other information			
Supplementary information	21	D, V	Provide information about the availability of supplementary resources, s calculator, and datasets.
Funding	22	D, V	Give the source of funding and the role of the funders for the present study.

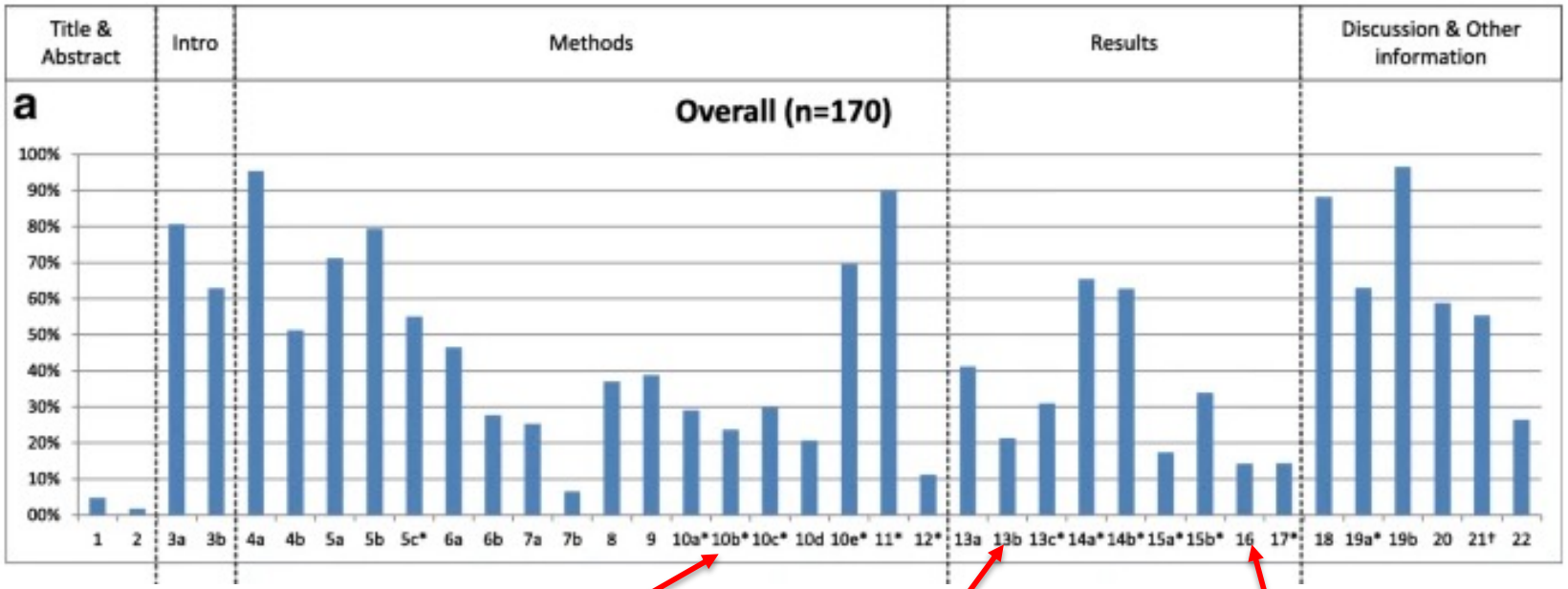
37 items covering
22 'topics' that should be included on an articles describing the development or validation of a prediction model

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

Participant characteristics

Report performance measures with CIs

* 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**

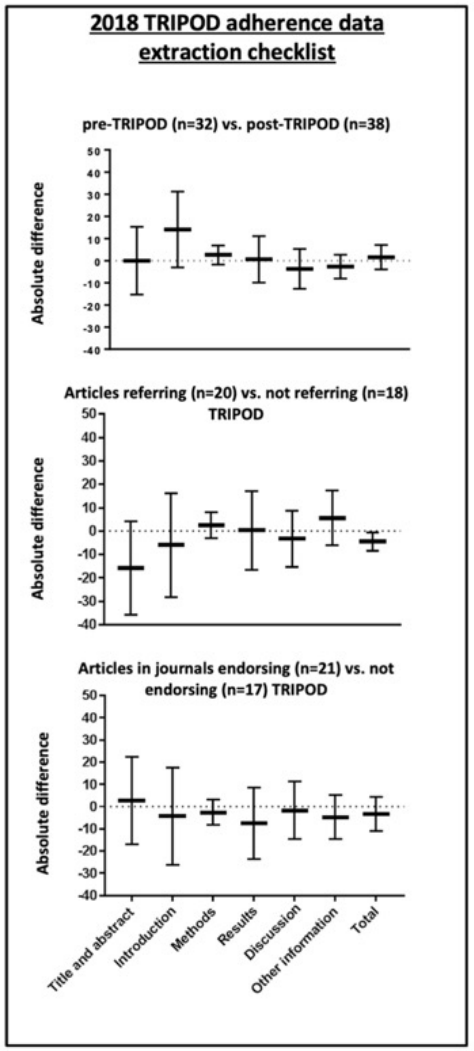


Figure 2 TRIPOD reporting scores. TRIPOD, Transparent Reporting of a multivariable prediction modelfor Individual Prognosis Or Diagnosis.

* 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

* Related initiatives; STARD-AI (in write-up), DECIDE-AI (to appear shortly), SPIRIT-AI (Nat Med/BMJ), CONSORT-AI (Nat Med/BMJ)

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/diag QUIPS (Hayden et al 2013) op 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; quar PROBAST* performance; internal validation

*Mod**el 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 Cochrane Risk of Bias tool es – relative to not using the model -> comparative studies.

*Wolff et al Ann Intern Med 2019

Models for organ transplantation*

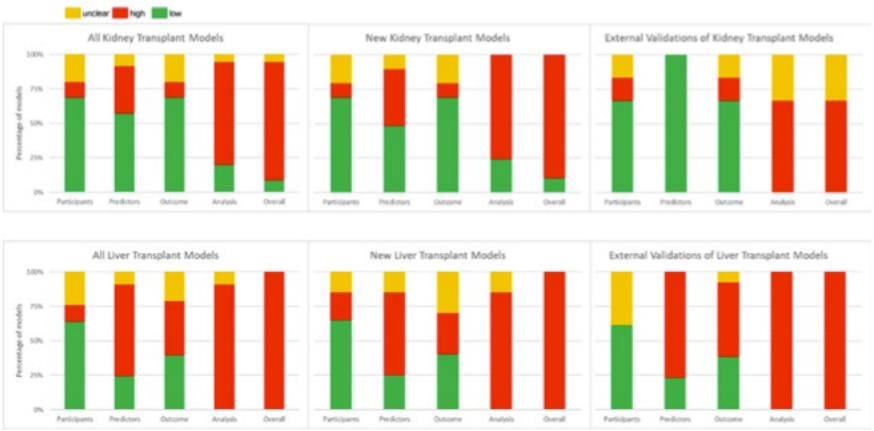


Fig. 3. Prediction model risk of bias assessment tool (PROBAST) risk of bias for included models ($n = 35$ for kidney transplant models, $n = 29$ for new kidney transplant models, $n = 6$ for external validations of kidney transplant models; $n = 33$ for liver transplant models, $n = 20$ for new liver transplant models, $n = 13$ for external validations of liver transplant models).

"We advise against applying poorly developed, reported, or validated prediction models "

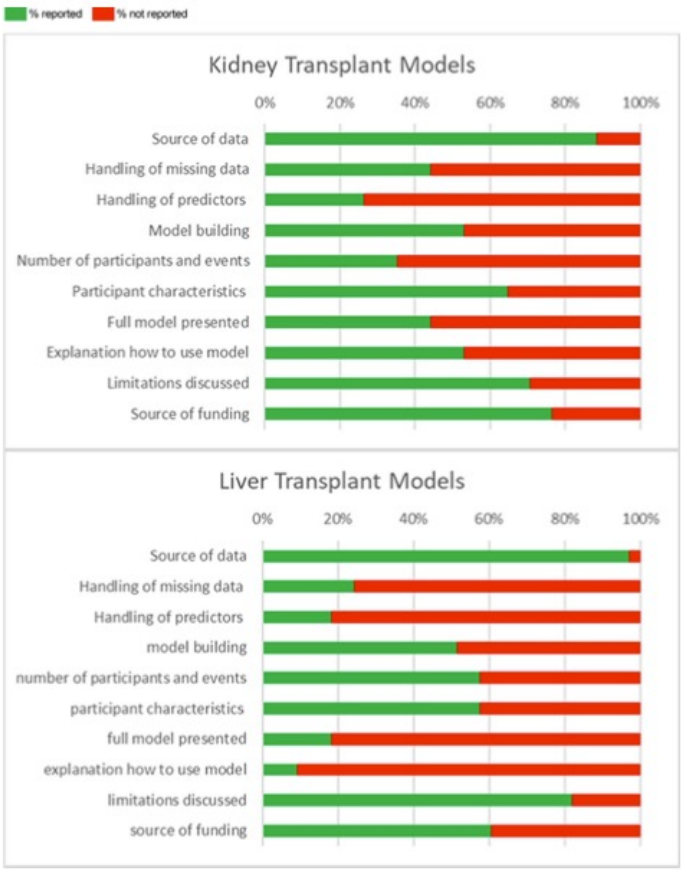
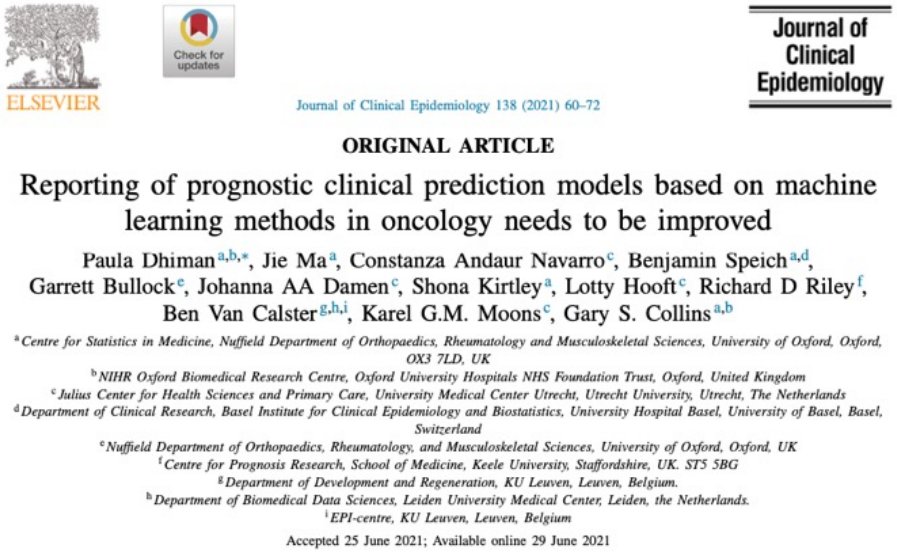


Fig. 4. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis checklist (TRIPOD) quality of reporting for selected items for included models ($n = 35$ kidney transplant models, $n = 33$ for liver transplant models).

* Haller et al J Clin Epidemiol 2022

Reporting of machine learning models



ELSEVIER

Check for updates

Journal of Clinical Epidemiology 138 (2021) 60–72

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

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

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

[a] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7LD, UK
[b] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom
[c] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
[d] Department of Clinical Research, Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, University of Basel, Basel, Switzerland
[e] Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[f] Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, UK. ST5 5BG
[g] Department of Development and Regeneration, KU Leuven, Leuven, Belgium.
[h] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands.
[i] EPI-centre, KU Leuven, Leuven, Belgium

Accepted 25 June 2021; Available online 29 June 2021

Andaur Navarro et al.
BMC Medical Research Methodology (2022) 22:12
https://doi.org/10.1186/s12874-021-01469-6

BMC Medical Research Methodology

RESEARCH Open Access

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

Constanza L. Andaur Navarro[1,2*], Johanna A. A. Damen[1,2], Toshihiko Takada[1], Steven W. J. Nijman[1], Paula Dhiman[3,4], Jie Ma[3], Gary S. Collins[3,4], Ram Bajpai[5], Richard D. Riley[5], Karel G. M. Moons[1,2] and Lotty Hooft[1,2]

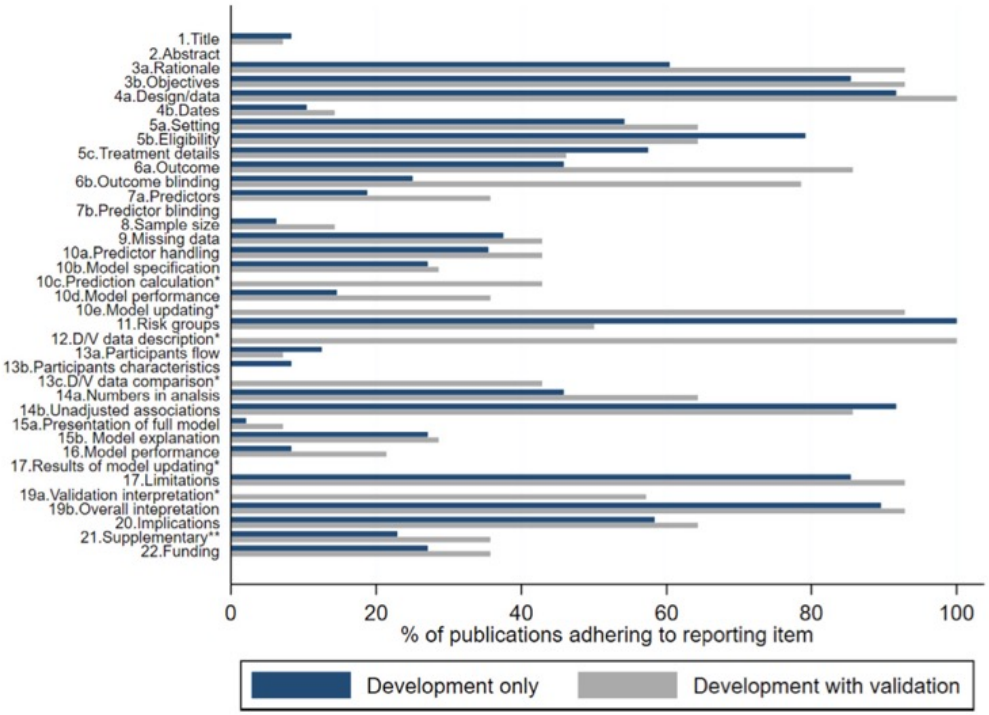
Abstract

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.

Dhiman et al J Clin Epidemiol 2021

Andaur Navarro et al BMC MRM 2022

Adherence to TRIPOD



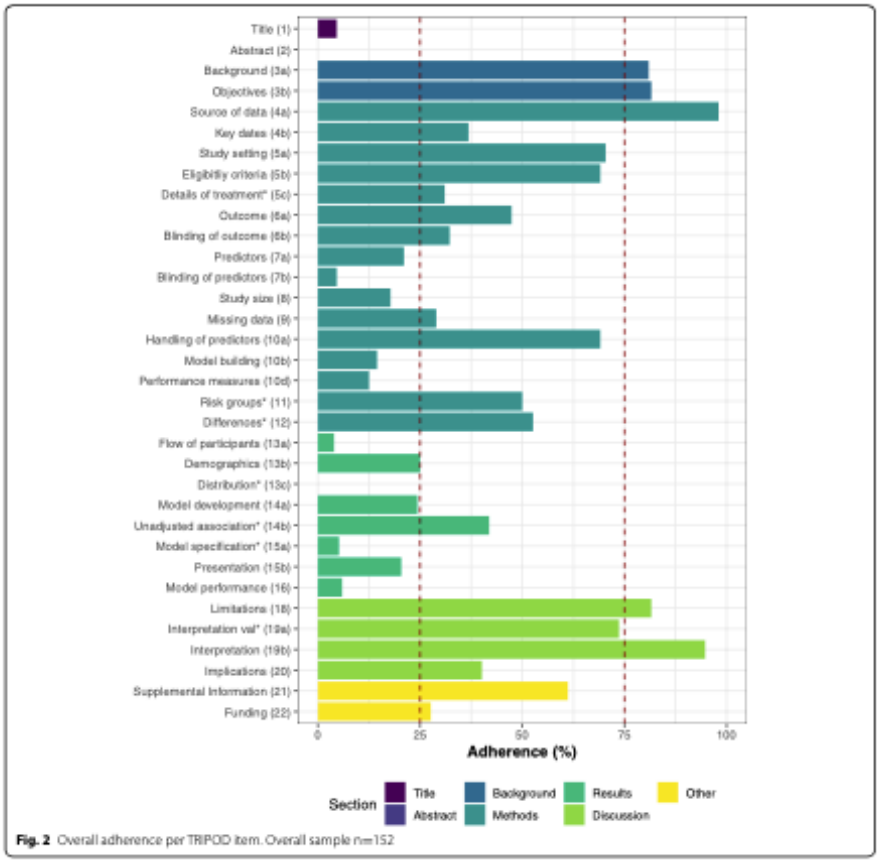
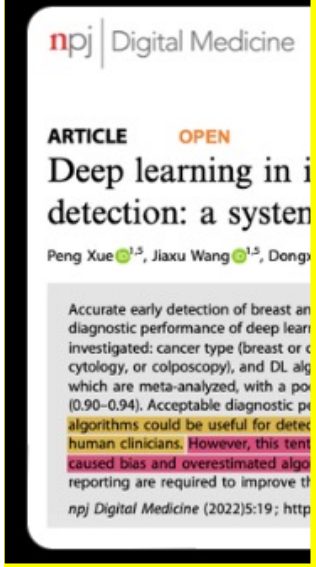


Fig. 2 Overall adherence per TRIPOD item. Overall sample n=152

"Most of the issues discussed here could be avoided through more robust designs & high-quality reporting, although several hurdles must be overcome before <deep learning> breast & cervical

--> tinyurl.com/3


npj | Digital Medicine

ARTICLE OPEN

Deep learning in i
detection: a systen

Peng Xue[1,5], Jiaxu Wang[1,5], Dongx

Accurate early detection of breast an
diagnostic performance of deep lear
investigated: cancer type (breast or c
cytology, or colposcopy), and DL alg
which are meta-analyzed, with a po
(0.90–0.94). Acceptable diagnostic pe
algorithms could be useful for detec
human clinicians. However, this tent
caused bias and overestimated algo
reporting are required to improve th

npj Digital Medicine (2022)5:19 ; http

👇"We recommend [...] greater emphasis on standardized ad
@TRIPODStaten
reproducibility"

TRIPOD -> tinyu
tinyurl.com/p8

#mltwitter #ml4


CNS Podcasts Journal Club

Jonathan Huang, BS
Nathan A. Shlobin, BA
Michael DeCuypere, MD, PhD
Sandi K. Lam, MD, MBA

Ann and Robert H. Lurie Children's Hospital, Division of Pediatric Neurosurgery, Department of Neurological Surgery, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

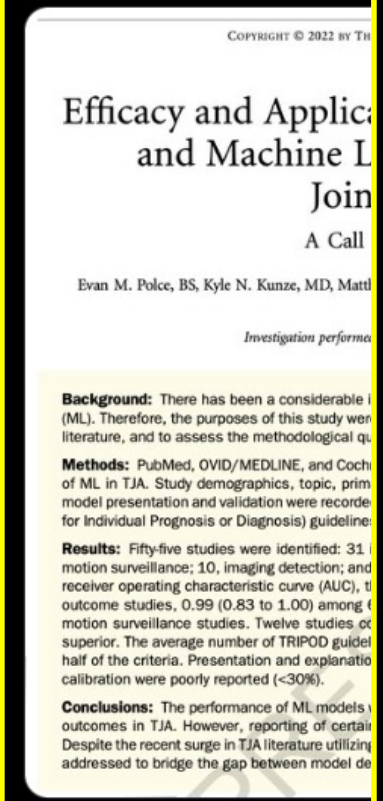
This work has been accepted for publication in abstract form for the 2021 American Association of Neurological Surgeons Annual Scientific Meeting held between August 21 and August 25, 2021, in Orlando, FL. This work has not been previously published in any other form.

Correspondence:
Sandi K. Lam, MD, MBA,
Ann and Robert H. Lurie Children's Hospital,
Division of Pediatric Neurosurgery,
Department of Neurological Surgery,
Feinberg School of Medicine,
Northwestern University,

"reporting of certain key methodological & model presentation criteria was inadequate. [...] lack of consistent adherence to reporting guidelines"
tinyurl.com/bdft4w9r

Adhere to @TRIPODSt
tinyurl.com/3bva4rna

#DataScience #machi


COPYRIGHT © 2022 BY TH

Efficacy and Applic
and Machine L
Join
A Call

Evan M. Polce, BS, Kyle N. Kunze, MD, Mattl

Investigation performa

Background: There has been a considerable
(ML). Therefore, the purposes of this study wer
literature, and to assess the methodological qu

Methods: PubMed, OVID/MEDLINE, and Coch
of ML in TJA. Study demographics, topic, prim
model presentation and validation were recorde
for Individual Prognosis or Diagnosis) guideline

Results: Fifty-five studies were identified: 31
motion surveillance; 10, imaging detection; and
receiver operating characteristic curve (AUC), t
outcome studies, 0.99 (0.83 to 1.00) among
motion surveillance studies. Twelve studies co
superior. The average number of TRIPOD guidel
half of the criteria. Presentation and explanatio
calibration were poorly reported (<30%).

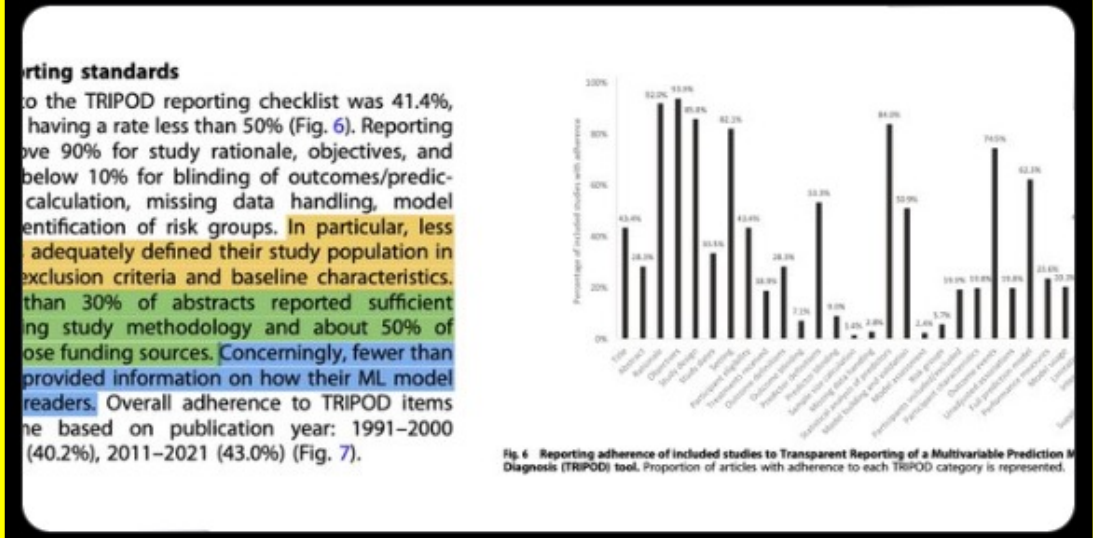
Conclusions: The performance of ML models
outcomes in TJA. However, reporting of certai
Despite the recent surge in TJA literature utilizin
addressed to bridge the gap between model de

👇#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


rting standards

o the TRIPOD reporting checklist was 41.4%, having a rate less than 50% (Fig. 6). Reporting ve 90% for study rationale, objectives, and below 10% for blinding of outcomes/predic- calculation, missing data handling, model entification of risk groups. In particular, less adequately defined their study population in exclusion criteria and baseline characteristics. than 30% of abstracts reported sufficient ng study methodology and about 50% of ose funding sources. Concerningly, fewer than provided information on how their ML model readers. Overall adherence to TRIPOD items he based on publication year: 1991–2000 (40.2%), 2011–2021 (43.0%) (Fig. 7).

Fig. 6 **Reporting adherence of included studies to Transparent Reporting of a Multivariable Prediction M Diagnosis (TRIPOD) tool.** Proportion of articles with adherence to each TRIPOD category is represented.

Twitter: @GSCollins

UNIVERSITY OF OXFORD

Impact of risk of bias

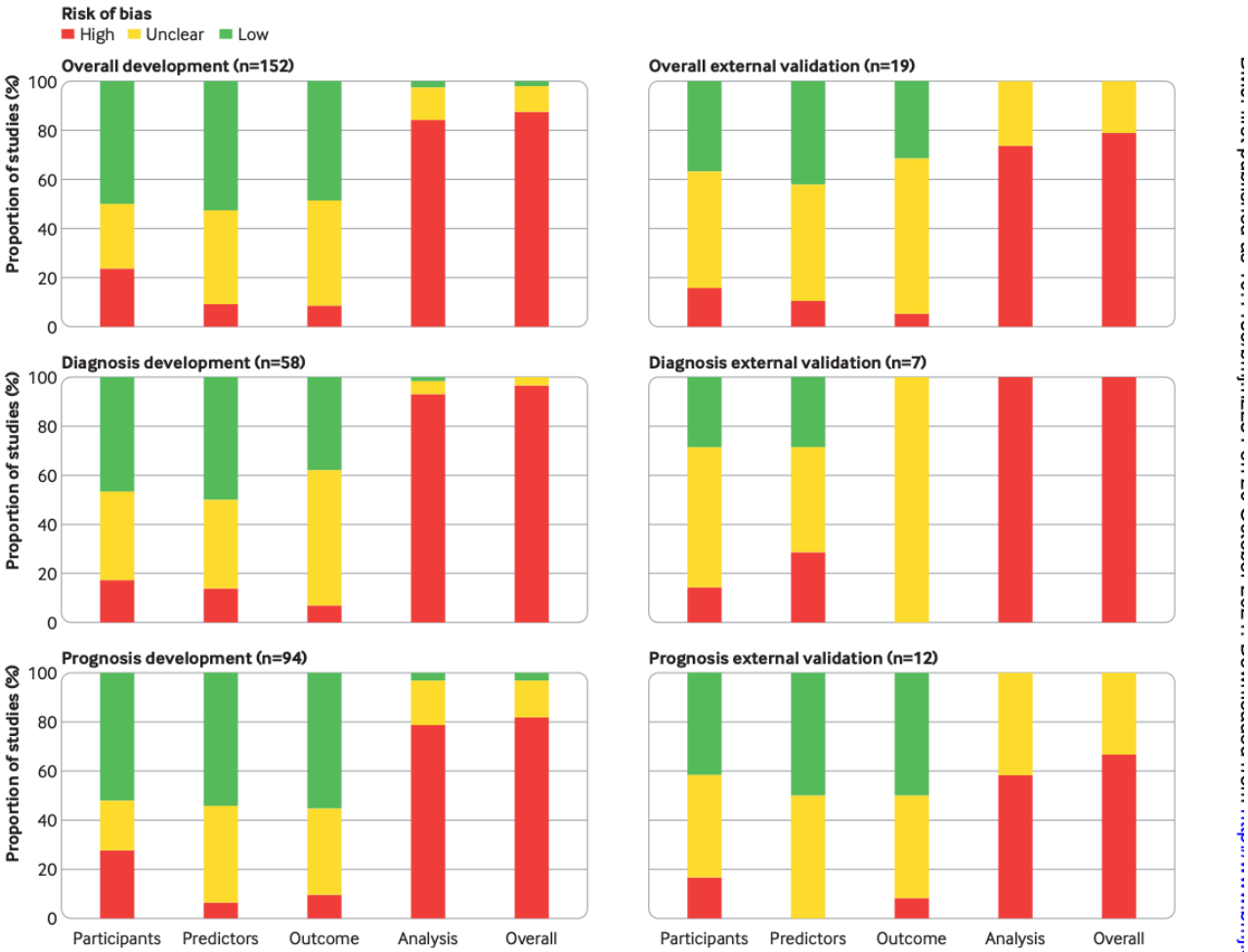


Fig 2 | Risk of bias of included studies (n=152) and stratified by study type

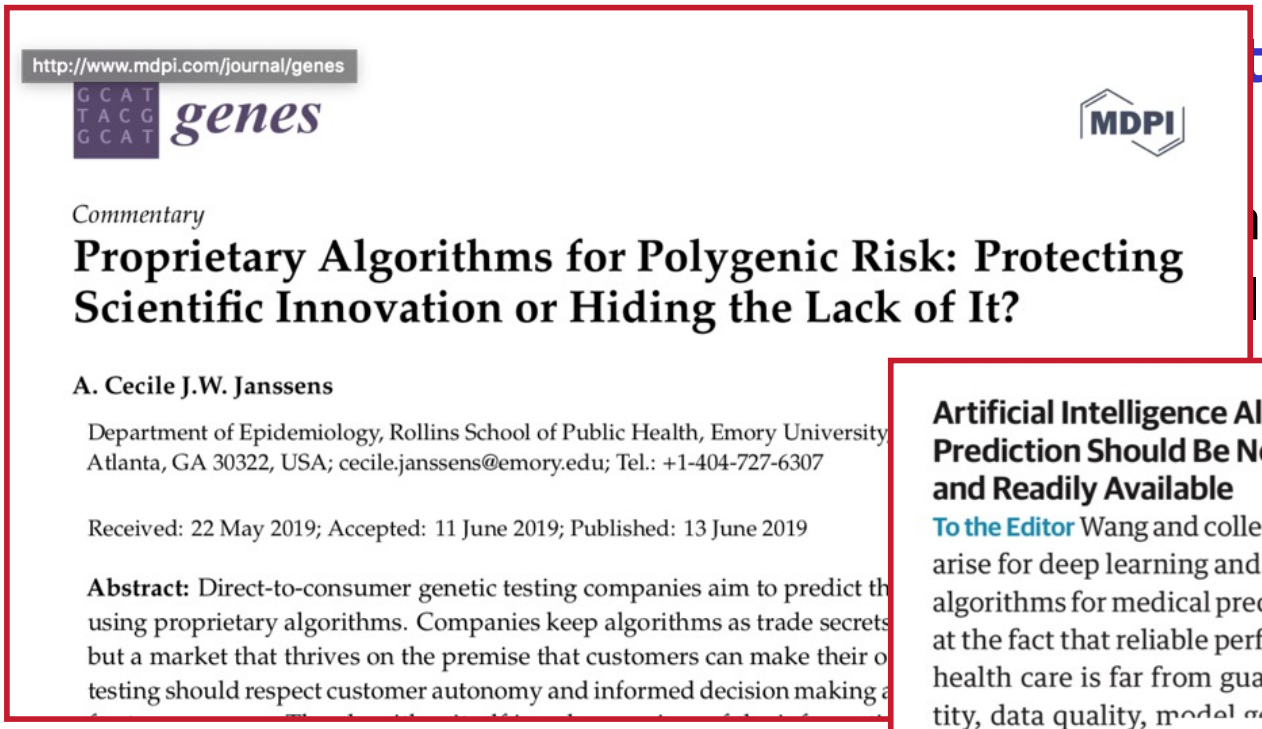
"Most studies on prediction models developed using machine learning show poor methodological quality and are at high risk of bias"

Andaur Navarro et al BMJ 2021

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



http://www.mdpi.com/journal/genes

genes

MDPI

Commentary

Proprietary Algorithms for Polygenic Risk: Protecting Scientific Innovation or Hiding the Lack of It?

A. Cecile J.W. Janssens

Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA; cecile.janssens@emory.edu; Tel.: +1-404-727-6307

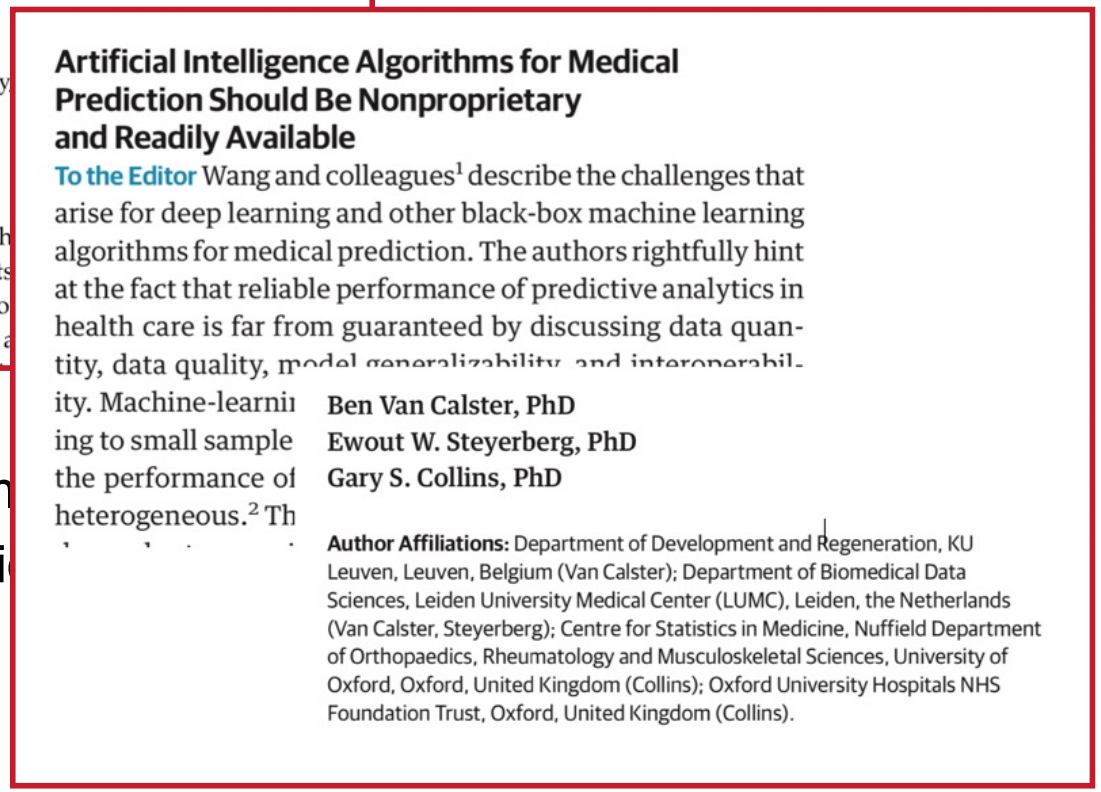
Received: 22 May 2019; Accepted: 11 June 2019; Published: 13 June 2019

Abstract: Direct-to-consumer genetic testing companies aim to predict th
using proprietary algorithms. Companies keep algorithms as trade secrets
but a market that thrives on the premise that customers can make their o
testing should respect customer autonomy and informed decision making a

typically be written

seline survival

lidate and recalibrate (to



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

To the Editor Wang and colleagues[1] 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 interoperabil-

ity. Machine-learni
ing to small sample
the performance of
heterogeneous.[2] Th

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).

- Issues of proprietary
  - Protecting scientific in
  - Commercial exploitati

Model availability

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.

Beau Norgeot, Giorgio Quer, Brett K. Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S. Kohane, Suchi Saria, Eric Topol, Ziad Obermeyer, Bin Yu and Atul J. Butte

The application of artificial intelligence (AI) in medicine is an old idea[1–3], but methods for this in the past involved programming computers with patterns or rules ascertained from human experts, which resulted in deterministic, rules-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[4–6]. The use of data labeled by clinical experts to train machine, probabilistic, and statistical models is called 'supervised machine learning'. Successful uses of these new machine-learning approaches include targeted real-time early-warning systems for adverse events[7], the detection of diabetic retinopathy[8], the classification of pathology and other images, the prediction of the near-term future state of patients with rheumatoid arthritis[9], patient discharge disposition[10], and more.

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

Matters arising

Transparency and reproducibility in artificial intelligence

https://doi.org/10.1038/s41586-020-2766-y
Received: 1 February 2020
Accepted: 10 August 2020
Published online: 14 October 2020
Check for updates

Benjamin Haibe-Kains[1,2,3,4,5✉], George Alexandru Adam[3,5], Ahmed Hosny[6,7], Farnoosh Khodakarami[1,3], Massive Analysis Quality Control (MAQC) Society Board of Directors*, Levi Waldron[8], Bo Wang[2,3,5,9,10], Chris McIntosh[3,5,9], Anna Goldenberg[3,5,11,12], Anshul Kundaje[13,14], Casey S. Greene[15,16], Tamara Broderick[17], Michael M. Hoffman[1,2,3,5], Jeffrey T. Leek[18], Keegan Korthauer[19,20], Wolfgang Huber[21], Alvis Brazma[22], Joelle Pineau[23,24], Robert Tibshirani[25,26], Trevor Hastie[25,26], John P. A. Ioannidis[25,26,27,28,29], John Quackenbush[30,31,32] & Hugo J. W. L. Aerts[6,7,33,34]

ARISING FROM S. M. McKinney et al. *Nature* https://doi.org/10.1038/s41586-019-1799-6 (2020)

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

Resource	URL
Code	
BitBucket	https://bitbucket.org
GitHub	https://github.com
GitLab	https://about.gitlab.com
Software dependencies	
Conda	https://conda.io
Code Ocean	https://codeocean.com
Gigantum	https://gigantum.com
Colaboratory	https://colab.research.google.com
Deep-learning models	
TensorFlow Hub	https://www.tensorflow.org/hub
ModelHub	http://modelhub.ai
ModelDepot	https://modeldepot.io
Model Zoo	https://modelzoo.co
Deep-learning frameworks	
TensorFlow	https://www.tensorflow.org/
Caffe	https://caffe.berkeleyvision.org/
PyTorch	https://pytorch.org/

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…

RESEARCH

OPEN ACCESS

Check for updates

FAST TRACK

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

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] 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,[5] 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 J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[25] René Spijker,[8,9,26] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[27,28] Sander M J van Kuijk,[29] Florien S van Royen,[8] Jan Y Verbakel,[30,31] Christine Wallisch,[7,32,33] Jack Wilkinson,[22] Robert Wolff,[34] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

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.

Cite this as: *BMJ* 2020;369:m1328 http://dx.doi.org/10.1136/bmj.m1328

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)

Red flag – should've been picked up during editorial process / peer review of primary studies

- **Calibration rarely assessed/reported (and often incorrectly)**

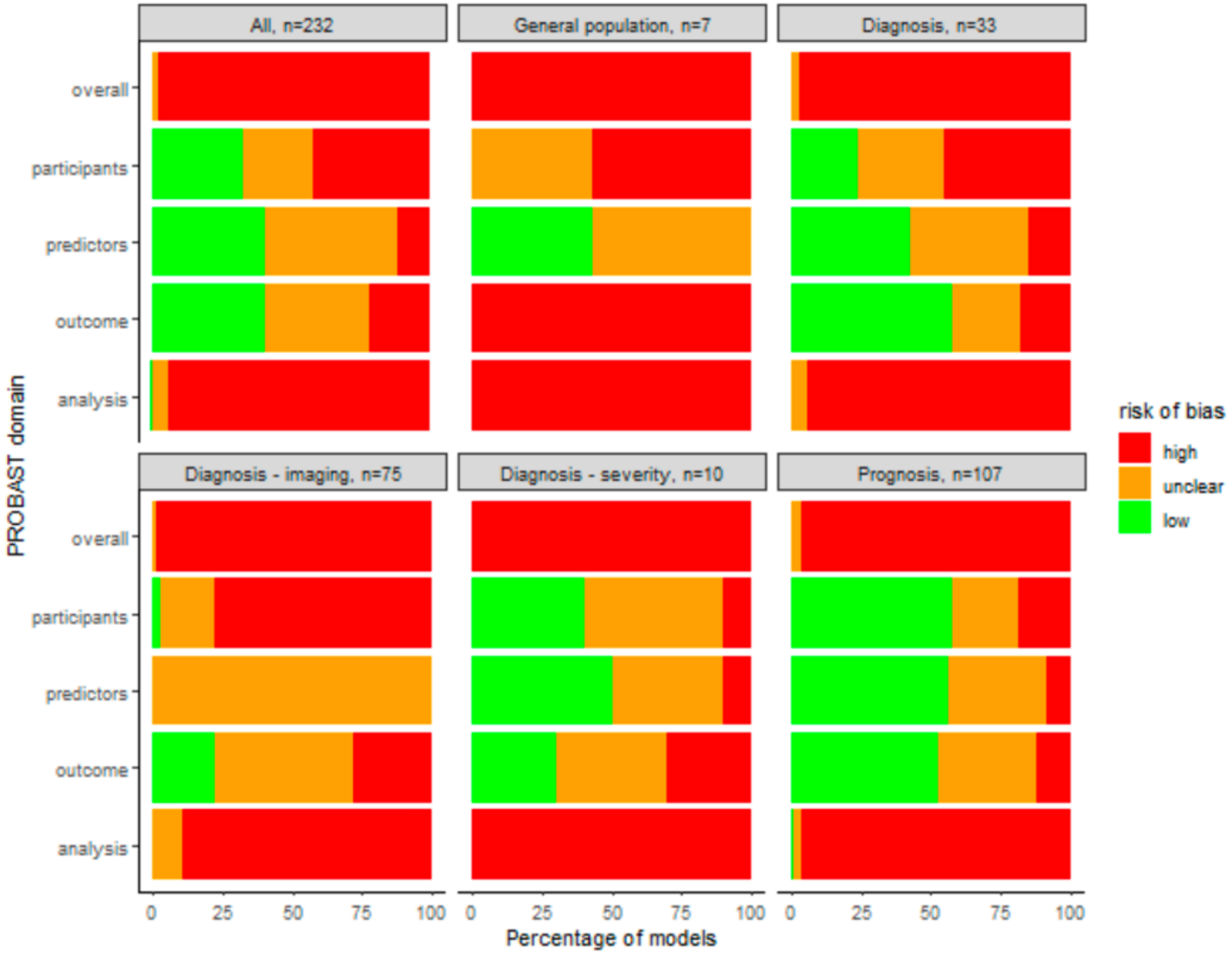
- **Table of participant characteristics sometimes missing**

- **"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



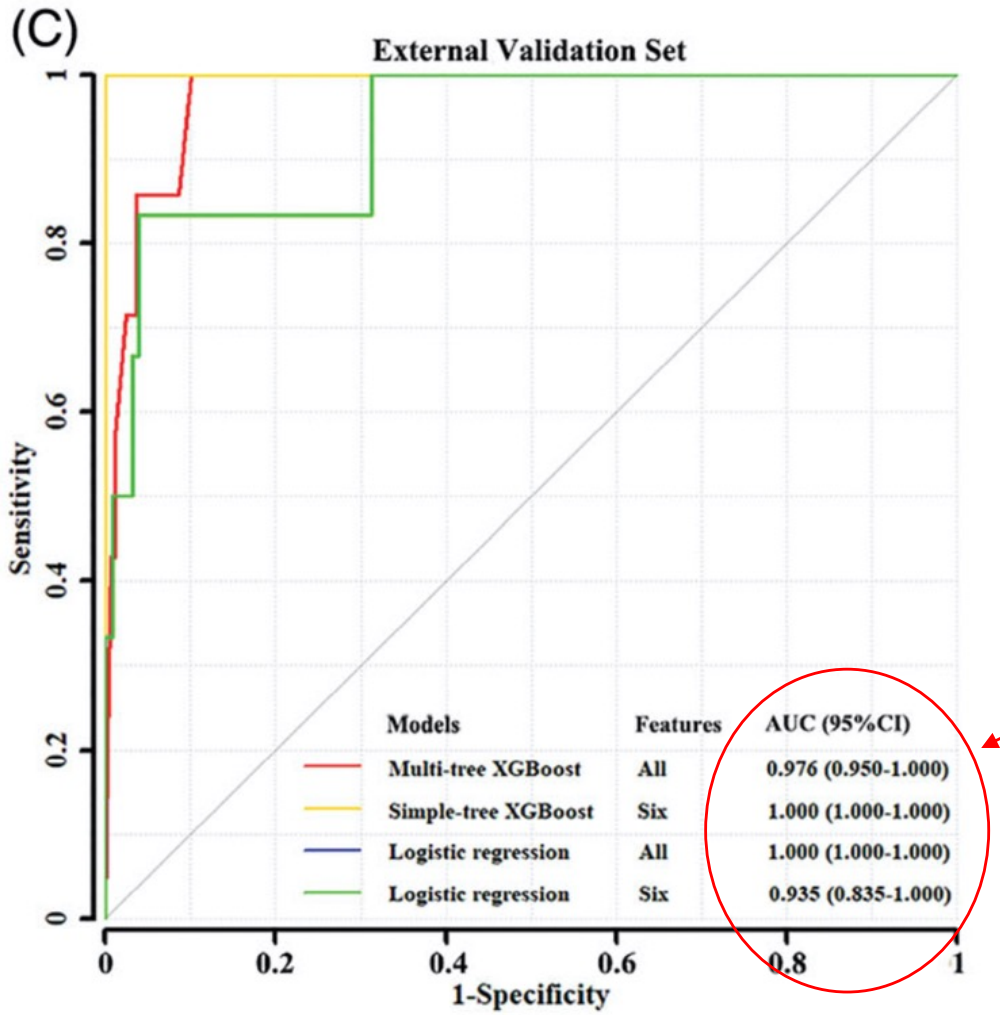
COVID Example 1* (generally poor)

- **Sample size for development (after splitting data in to 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"

* Wu et al, Eur Respir J 2020; LTE Collins et al, Eur Respir J 2020 + author response

Small validation sample size → misleading conclusions



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

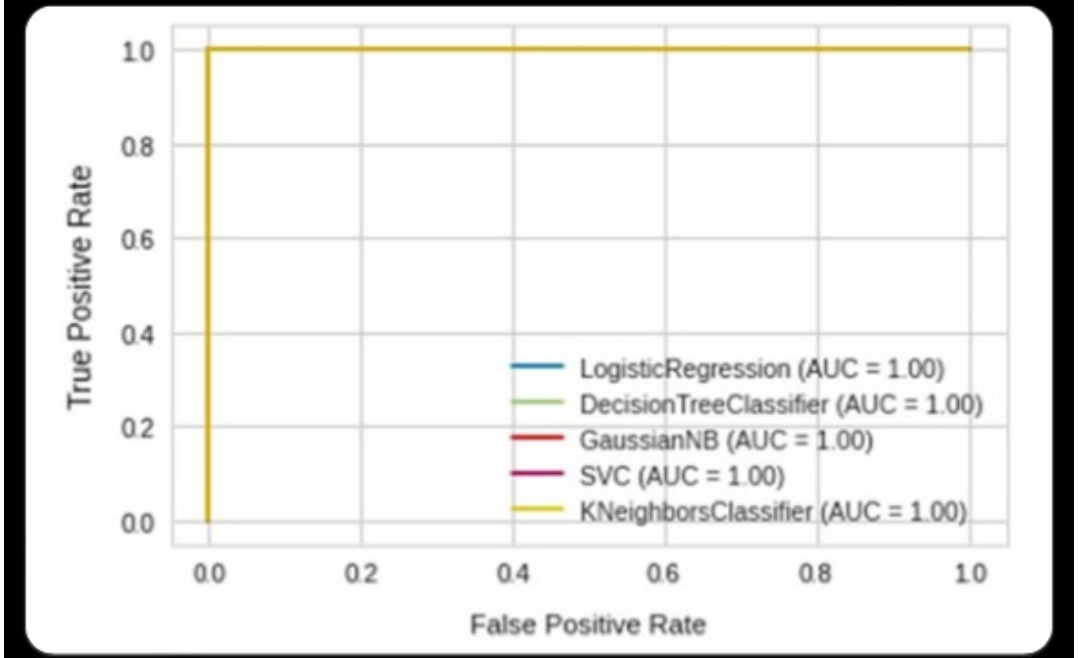
No calibration

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/upyxmf6s)

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)**
  - Guidance for Abstracts (TRIPOD for Abstracts) [Heus et al, Ann Intern Med 2020]
  - New reporting guidelines for machine learning (TRIPOD-AI), systematic reviews (TRIPOD-SRMA) and protocols (TRIPOD-P) in preparation

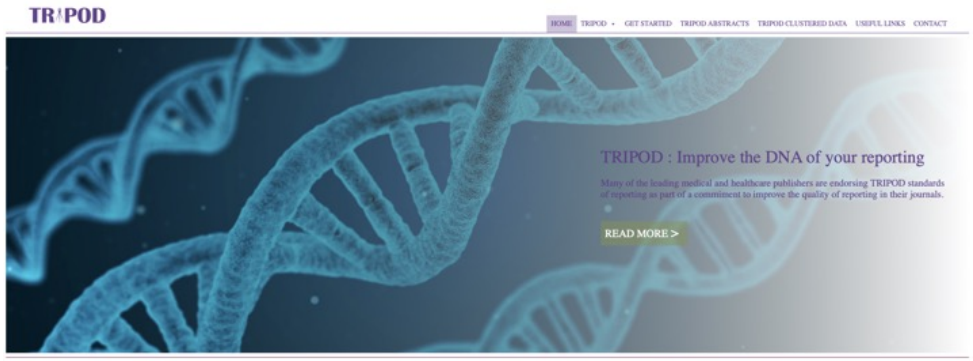
* Wolff et al. Ann Intern Med 2019; Moons et al. Ann Intern Med 2019

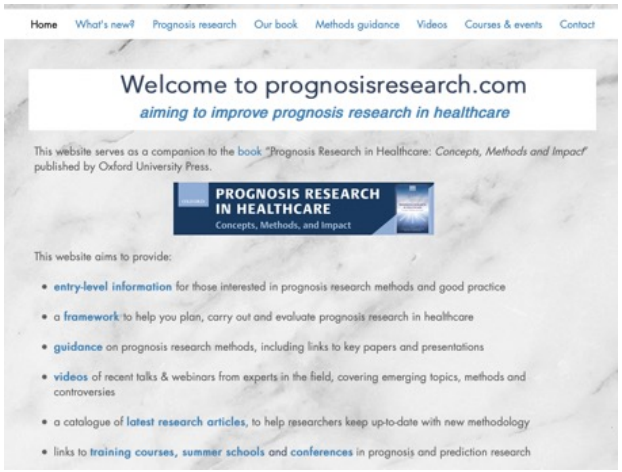
Thank you for listening

Email: gary.collins@csm.ox.ac.uk
Twitter: @GSCollins

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

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





PROBAST

Risk of bias: www.probast.org

www.prognosisresearch.com

STRATOS INITIATIVE

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

equator network

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

UNIVERSITY OF OXFORD