Clinical prediction models: a field in crisis?

Gary Collins
Professor of Medical Statistics
Centre for Statistics in Medicine/UK EQUATOR Centre
University of Oxford

email: gary.collins@csm.ox.ac.uk
twitter: @gscollins

03-February-2021

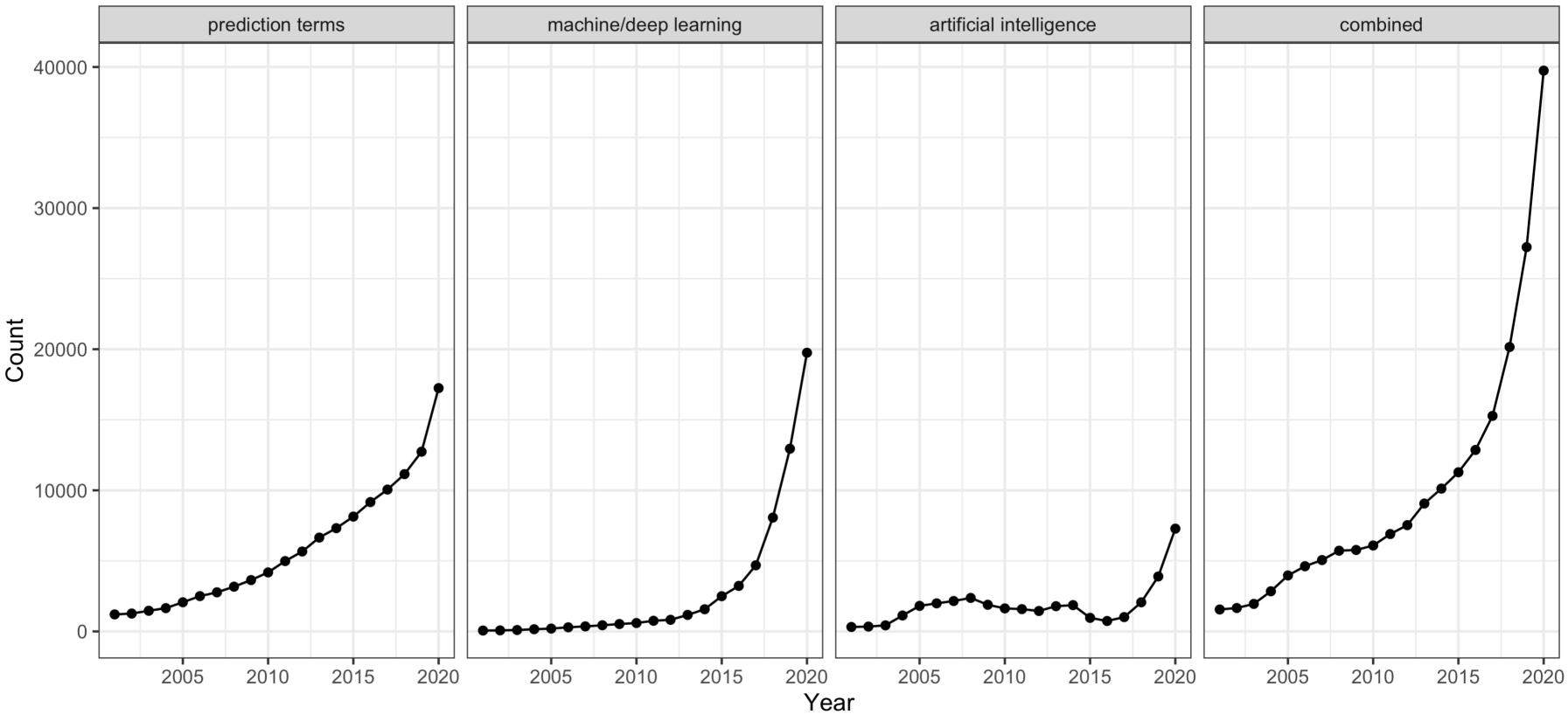
Overview of talk

- **Critical overview of regression-based prediction models in the clinical literature**

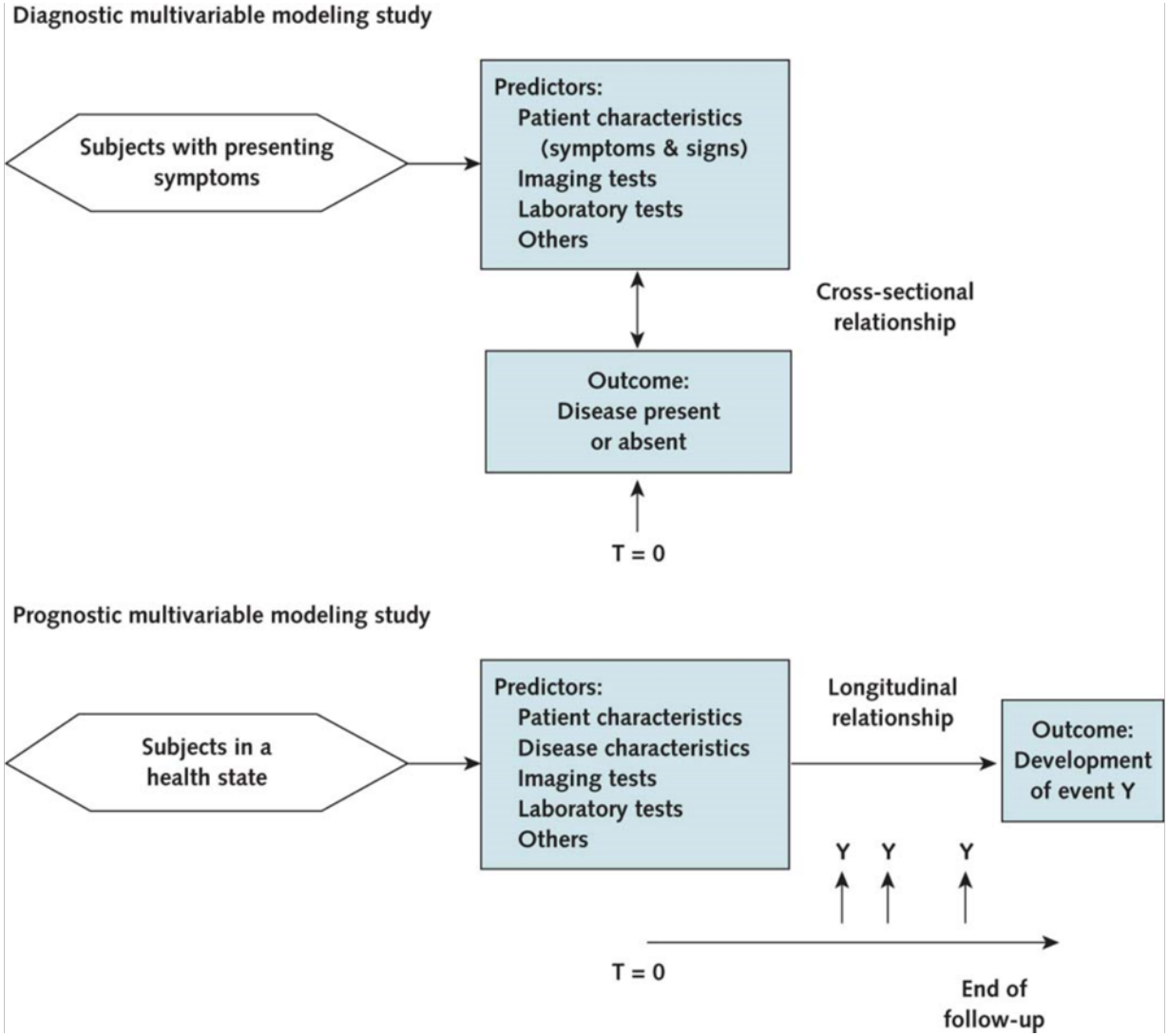
- **Critical overview of machine learning for clinical (risk) prediction**

  - (Some) concerns
  - Comparative studies
  - Reporting

Interest in prediction



Diagnostic / Prognostic models



Collins et al. Ann Intern Med 2015; 162: 55-63

Clinical Prediction Models

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES
UNIVERSITY OF OXFORD

- Aim is to combine multiple patient characteristics to predict the probability of a health outcome
  - Diagnostic
  - Prognostic
- Increasingly recommended in (NICE) Clinical Guidelines
  - E.g. QRISK, ABCD2, FRAX, Blatchford, SAPS, APACHE, NPI
- Most existing models are typically developed using regression based approaches (logistic, Cox)
- Widely available (to both the public and healthcare professionals) on websites, and smartphone apps
  - Little (current) regulation -> slowly seeing movement in this area

iPhone: Framingham Risk Score

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES
UNIVERSITY OF OXFORD

Free

Category: Medical
Updated: Jul 15, 2011
Version: 1.5
Size: 2.6 MB
Language: English
Seller: Austin Physician Productivity, LLC
© STATCODER.COM
Rated 9+ for the following:
Infrequent/Mild Mature/Suggestive Themes

Requirements: Compatible with iPhone, iPod touch, and iPad. Requires iOS 3.0 or later

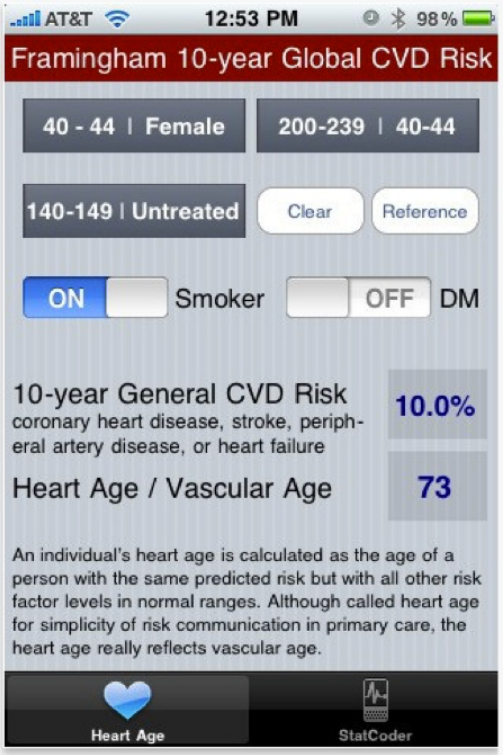
Customer Ratings

Current Version:
★★½ 5 Ratings

All Versions:
★★★ 103 Ratings

More iPhone Apps by Austin Physician Productivity, LLC

iPhone Screenshots



CSM
CENTRE FOR STATISTICS IN MEDICINE

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

predict
breast cancer

NHS

Home | About Predict | Predict Tool | Contact | Legal | Earlier versions

What is Predict?

Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

It is endorsed by the American Joint Committee on Cancer (AJCC).

Start Predict

Did you mean to visit Predict Prostate?

What does Predict do?

Predict asks for some details about the patient and the cancer. It then uses data about the survival of similar women in the past to show the likely proportion of such women expected to

Who is Predict for?

Predict is for clinicians, patients and their families.

Patients should use it in consultation with a medical professional.

Where can I find out more?

To read more go to **About Predict**

CSM
CENTRE FOR STATISTICS IN MEDICINE

NICE Clinical Guidelines

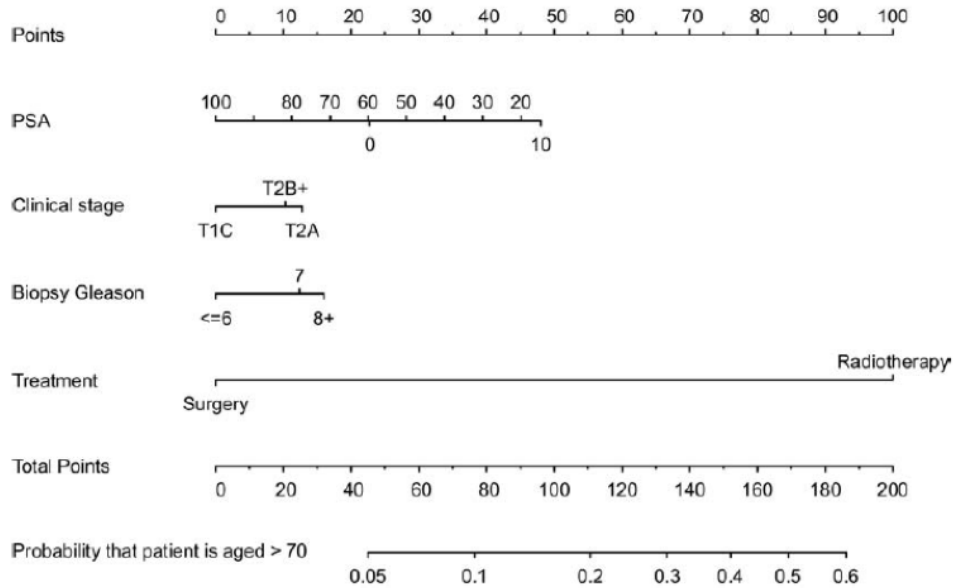
- **QRISK (NICE CG 67)**
  - 10-year risk of developing cardiovascular disease
- **Nottingham prognostic index (NICE CG80)**
  - risk of recurrence and overall survival in breast cancer patients
- **GRACE / PURSUIT / PREDICT / TIMI (NICE CG94)**
  - adverse CVD outcomes (mortality, MI, stroke etc…) for patients with UA/NSTEMI
- **APGAR (NICE CG132/2)**
  - evaluate the prognosis of a newborn baby
- **SAPS / APACHE (NICE CG50)**
  - ICU scoring systems for predicting mortality
- **CRB65/CURB65 (NICE CG191)**
  - Pneumonia
- **FRAX / QFracture (NICE fragility risk short guideline)**
  - 10-year risk of developing osteoporotic & hip fracture

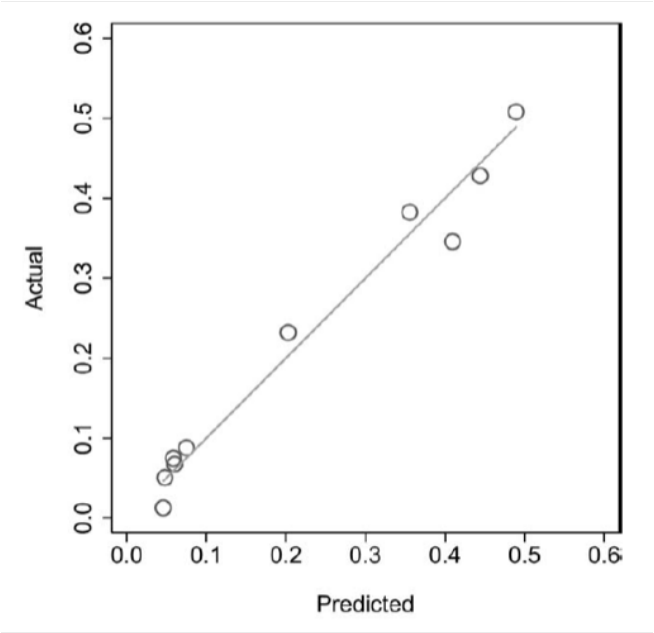
Published reviews

- **POOR QUALITY STUDIES**
- **POORLY REPORTED**
- **MOST HAVE NOT BEEN VALIDATED**
- **MOST ARE NOT BEING USED**
- **RESEARCH WASTE**

- 54 models for breast cancer (Altman 2009)
- 43 models for type 2 diabetes (Collins 2011

Pointless prediction models





Vickers & Cronin. Everything you wanted to know about evaluating prediction models (but were too afraid to ask). J Urol, 2010.

Pointless prediction models

vs radiotherapy). The model has high discrimination (AUC of 0.78) and good calibration (see Fig. 2). In other words, the model is terrific in all ways other than that it is completely useless. So why did we create it? In short, because we could: we have a dataset, and a statistical package, and add the former to the latter, hit a few buttons and *voila*, we have another paper. It is tempting to speculate that the ubiquity of nomograms in the uro-

Pointless prediction models

prediction models

package, and add the former to the latter, hit a few buttons and *voila*, we have another paper. It is tempting to speculate that the ubiquity of nomograms in the ~~urological~~ literature is simply because it is particularly easy research to do: you do not need to collect any data or even think of an interesting scientific question. We would argue that a predictive model should only be published if it is has a compelling clinical use, and that is rarely the case.

Methodological shortcomings (I)

- **Missing data rarely mentioned**
  - often an exclusion criteria (though often not specified)
  - complete-case usually carried out

- **Range of continuous predictors rarely reported**
  - Useful to set-out who the model is intended for

- **Models often not reported in full (nor 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?

Methodological shortcomings (II)

- **Small sample size (number of events); EPV<10 => overfitting**
  - Recent developments in sample size†
- **Large number of candidate predictors**
- **Calibration rarely assessed**
  - not reported in 46% (Bouwmeester: general medical journals) to 85% studies (Altman: 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
  - Meaningless / limited (external) validation (based on convenience data)
- **Lack of comparing competing models (or unfair comparisons)**
- **Unsurprisingly (and fortunately) very few models are used**

† Riley et al 2019/2020; van Smeden et al 2016/2018

Poor reporting

- **Number of events often difficult to identify**
  - candidate predictors (and number) not always easy to find
- **Insufficient information to report EPV (events-per-variable)**
  - 40% of studies (Mallett 2010; Collins 2011)
- **How candidate predictors were selected**
  - unclear in 25% studies (Bouwmeester 2012)
- **How the multivariable model was derived**
  - unclear in 77% of studies in cancer (Mallet 2010)

C

- 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

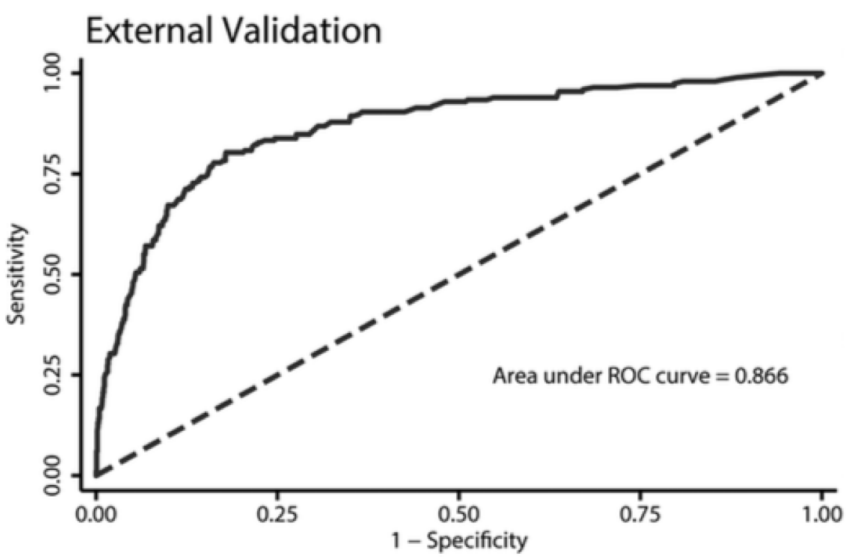
- Tend to be small (few events) (48% < 100 events)
  - 100 events is the current sample size recommendation for validation (Van Calster et al 2016, Collins et al 2016) [for assessing calibration]

- Missing data rarely mentioned (54%)
  - 64% conducted complete-case analyses (not always explicit)
  - 9% used multiple imputation

- Overwhelming focus only on discrimination
  - 73% of external validation studies evaluated discrimination; only 32% assessed calibration; 24% presented 'blank' ROC curves (no cut-points labelled)

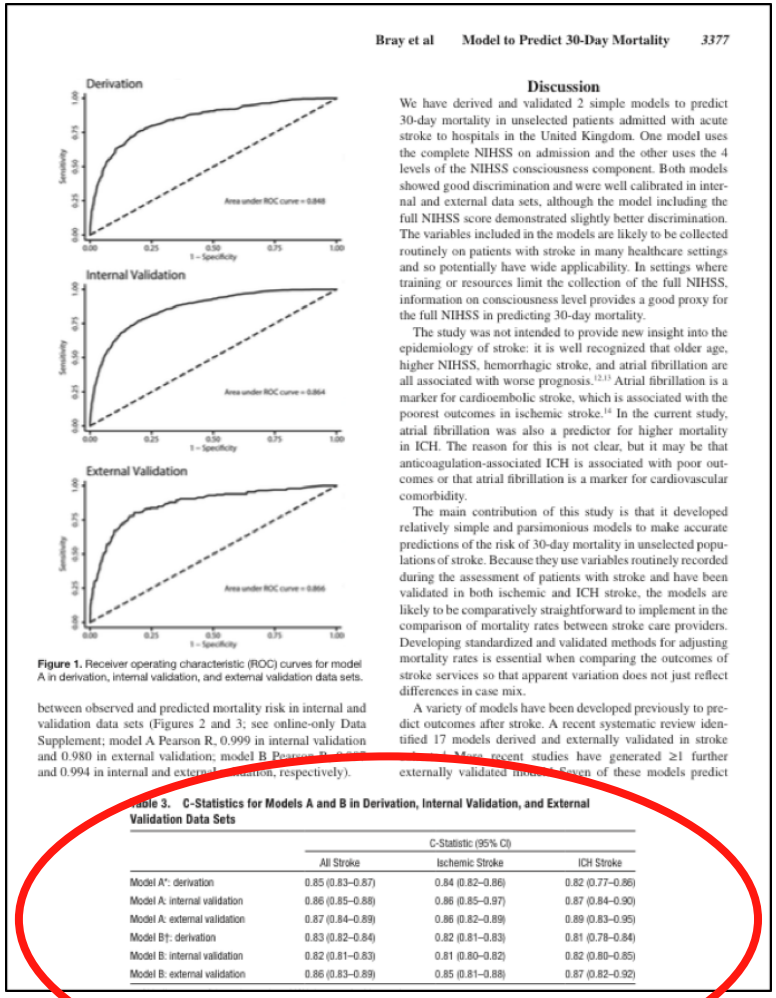
Collins et al, BMC Med Res Methodol 2014

Wasting space



- 3 uninformative ROC curves
- No (informative) calibration curve

=> this is a reporting issue



Bray et al Model to Predict 30-Day Mortality 3377

Derivation

Internal Validation

External Validation

Figure 1. Receiver operating characteristic (ROC) curves for model A in derivation, internal validation, and external validation data sets.

between observed and predicted mortality risk in internal and validation data sets (Figures 2 and 3; see online-only Data Supplement; model A Pearson R, 0.999 in internal validation and 0.980 in external validation; model B Pearson R, [illegible] and 0.994 in internal and external validation, respectively).

Discussion

We have derived and validated 2 simple models to predict 30-day mortality in unselected patients admitted with acute stroke to hospitals in the United Kingdom. One model uses the complete NIHSS on admission and the other uses the 4 levels of the NIHSS consciousness component. Both models showed good discrimination and were well calibrated in internal and external data sets, although the model including the full NIHSS score demonstrated slightly better discrimination. The variables included in the models are likely to be collected routinely on patients with stroke in many healthcare settings and so potentially have wide applicability. In settings where training or resources limit the collection of the full NIHSS, information on consciousness level provides a good proxy for the full NIHSS in predicting 30-day mortality.

The study was not intended to provide new insight into the epidemiology of stroke: it is well recognized that older age, higher NIHSS, hemorrhagic stroke, and atrial fibrillation are all associated with worse prognosis.[12,13] Atrial fibrillation is a marker for cardioembolic stroke, which is associated with the poorest outcomes in ischemic stroke.[14] In the current study, atrial fibrillation was also a predictor for higher mortality in ICH. The reason for this is not clear, but it may be that anticoagulation-associated ICH is associated with poor outcomes or that atrial fibrillation is a marker for cardiovascular comorbidity.

The main contribution of this study is that it developed relatively simple and parsimonious models to make accurate predictions of the risk of 30-day mortality in unselected populations of stroke. Because they use variables routinely recorded during the assessment of patients with stroke and have been validated in both ischemic and ICH stroke, the models are likely to be comparatively straightforward to implement in the comparison of mortality rates between stroke care providers. Developing standardized and validated methods for adjusting mortality rates is essential when comparing the outcomes of stroke services so that apparent variation does not just reflect differences in case mix.

A variety of models have been developed previously to predict outcomes after stroke. A recent systematic review identified 17 models derived and externally validated in stroke [illegible] Many recent studies have generated ≥1 further externally validated models. Seven of these models predict

Table 3. C-Statistics for Models A and B in Derivation, Internal Validation, and External Validation Data Sets

	C-Statistic (95% CI)		
	All Stroke	Ischemic Stroke	ICH Stroke
Model A*: derivation	0.85 (0.83–0.87)	0.84 (0.82–0.86)	0.82 (0.77–0.86)
Model A: internal validation	0.86 (0.85–0.88)	0.86 (0.85–0.97)	0.87 (0.84–0.90)
Model A: external validation	0.87 (0.84–0.89)	0.86 (0.82–0.89)	0.89 (0.83–0.95)
Model B†: derivation	0.83 (0.82–0.84)	0.82 (0.81–0.83)	0.81 (0.78–0.84)
Model B: internal validation	0.82 (0.81–0.83)	0.81 (0.80–0.82)	0.82 (0.80–0.85)
Model B: external validation	0.86 (0.83–0.89)	0.85 (0.81–0.88)	0.87 (0.82–0.92)

ELSEVIER
Check for updates
Journal of Clinical Epidemiology 126 (2020) 207–216
Journal of Clinical Epidemiology

ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES

ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models

Jan Y. Verbakel[a,b], Ewout W. Steyerberg[c], Hajime Uno[d], Bavo De Cock[e], Laure Wynants[e], Gary S. Collins[f,g], Ben Van Calster[c,e,*]

[a]KU Leuven, Department of Public Health and Primary Care, Leuven, Belgium
[b]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[c]Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands
[d]Division of Population Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[e]KU Leuven, Department of Development and Regeneration, Leuven, Belgium
[f]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[g]Oxford University Hospitals NHS Foundation Trust, Oxford, UK

Accepted 20 January 2020; Published online 23 July 2020

ELSEVIER
Check for updates
Journal of Clinical Epidemiology 126 (2020) 217–219
Journal of Clinical Epidemiology

ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES

ROC curves for clinical prediction models part 2. The ROC plot: the picture that could be worth a 1000 words

A. Cecile J.W. Janssens*

Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA

Accepted 24 May 2020; Published online 18 June 2020

ELSEVIER
Check for updates
Journal of Clinical Epidemiology 126 (2020) 220–223
Journal of Clinical Epidemiology

ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES

ROC curves for clinical prediction models part 3. The ROC plot: a picture that needs a 1000 words

Ben Van Calster[a,b,c,*], Laure Wynants[a,d], Gary S. Collins[e,f], Jan Y. Verbakel[c,g,h], Ewout W. Steyerberg[b]

[a]KU Leuven, Department of Development and Regeneration, Leuven, Belgium
[b]Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands
[c]EPI-Centre, KU Leuven, Leuven, Belgium
[d]Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, the Netherlands
[e]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Musculoskeletal Sciences, University of Oxford, Oxford, UK
[f]NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK
[g]Academic Centre for Primary Care, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium
[h]Nuffield Department of Primary Care Health Sciences, University of Oxford, UK

Accepted 24 May 2020; Published online 18 June 2020

ELSEVIER
Check for updates
Journal of Clinical Epidemiology 126 (2020) 224–225
Journal of Clinical Epidemiology

ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES

ROC curves for clinical prediction models part 4. Selection of the risk threshold—once chosen, always the same?

A. Cecile J.W. Janssens*

Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta GA, USA

Accepted 24 May 2020; Published online 18 June 2020

TRIPOD Statement

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES
UNIVERSITY OF OXFORD

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 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 im-

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**.

CSM
CENTRE FOR STATISTICS IN MEDICINE

Heus *et al. BMC Medicine* (2018) 16:120
https://doi.org/10.1186/s12916-018-1099-2

BMC Medicine

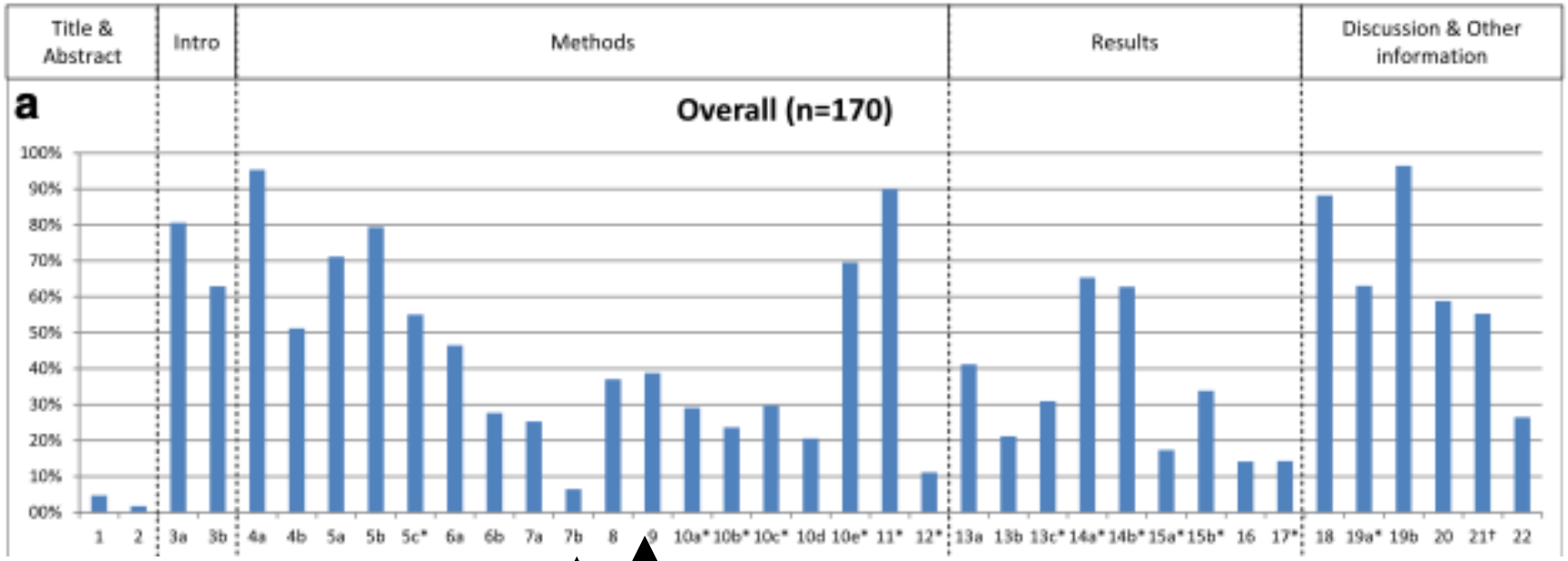
RESEARCH ARTICLE Open Access

Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement

Pauline Heus[1,2*], Johanna A. A. G. Damen[1,2], Romin Pajouheshnia[2], Rob J. P. M. Scholten[1,2], Johannes B. Reitsma[1,2], Gary S. Collins[3], Douglas G. Altman[3], Karel G. M. Moons[1,2] and Lotty Hooft[1,2]

Abstract

Background: As complete reporting is essential to judge the validity and applicability of multivariable prediction models, a guideline for the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or



(7b) Blinding

(8) Sample size

(9) Missing data

(15a) Model presentation

(16) Performance measures with CI

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

Open access Original research

BMJ Open

TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models

Amir H Zamanipoor Najafabadi ,[1] Chava L Ramspek ,[2] Friedo W Dekker,[2] Pauline Heus ,[3] Lotty Hooft,[4] Karel G M Moons,[5] Wilco C Peul,[1,6] Gary S Collins,[7] Ewout W Steyerberg,[8] Merel van Diepen[2]

To cite: Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, *et al*. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open* 2020;**10**:e041537. doi:10.1136/bmjopen-2020-041537

► Prepublication history and additional material for this paper are available online. To view these files, please visit

ABSTRACT

Objectives To assess the difference in completeness of reporting and methodological conduct of published prediction models before and after publication of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.

Methods In the seven general medicine journals with the highest impact factor, we compared the completeness of the reporting and the quality of the methodology of prediction model studies published between 2012 and 2014 (pre-TRIPOD) with studies published between 2016 and 2017 (post-TRIPOD). For articles published in the post-TRIPOD period, we examined whether there was improved

Strengths and limitations of this study

- ► This is the first study to assess the completeness of reporting and methodological conduct of prediction models published before and after publication of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.
- ► A limitation of this study is the short time period evaluated and therefore future studies are needed to assess the long-term effects on completeness of reporting and methodological conduct.

CSM
CENTRE FOR STATISTICS IN MEDICINE

Pre ('12-'14) and post TRIPOD ('16-'17)

- No discernible improvement in reporting
- But improvements in assessment of model performance
  - Calibration (21% vs 87%)
- Handling of missing data, e.g., multiple imputation (12% versus 50%)
- Limitations: Small sample size, short post TRIPOD time frame

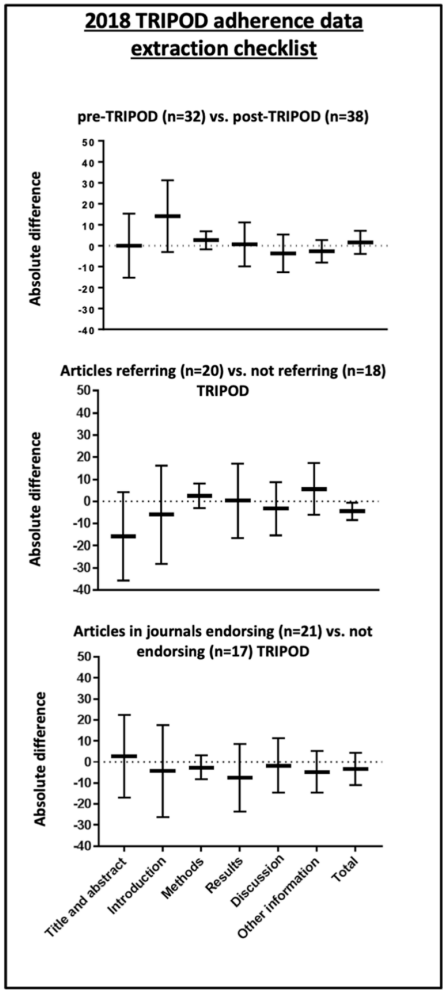


Figure 2 TRIPOD reporting scores. TRIPOD, Transparent Reporting of a multivariable prediction modelfor Individual Prognosis Or Diagnosis.

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]

"Good reporting is not an optional extra; it is an essential component of research"

Altman *et al*. Open Med 2008

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

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

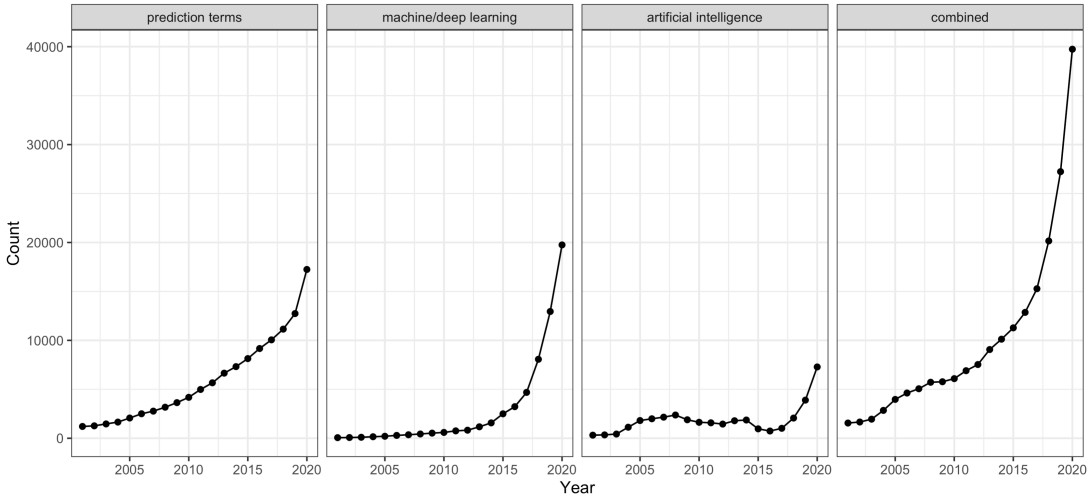
Lancet 2014; 383: 267–76

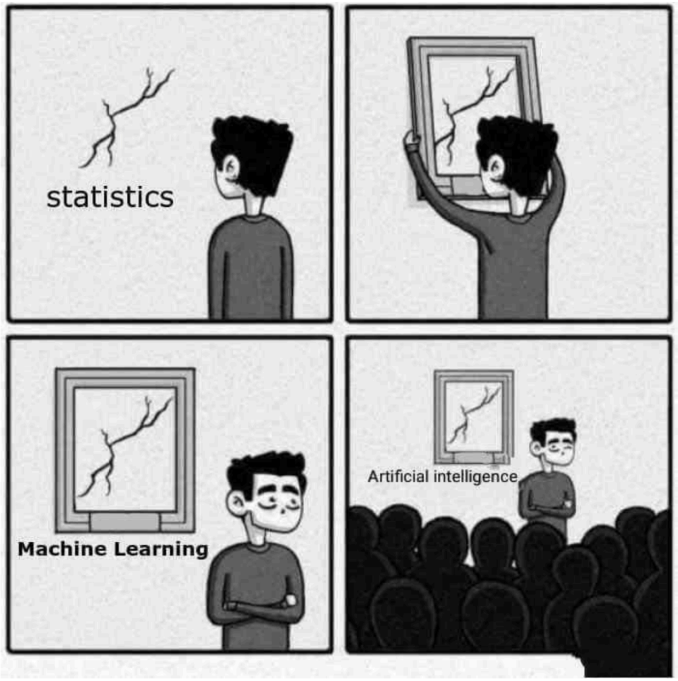
Published Online
January 8, 2014
http://dx.doi.org/10.1016/S0140-6736(13)62228-X

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

CSM
CENTRE FOR STATISTICS IN MEDICINE

Interest in ‘machine learning’





What is machine learning? (Uninteresting question) - Always sparks ‘debate’ between machine learners and statisticians

Interest in machine learning

- **Growing interest/enthusiasm in using machine learning for predicting health outcomes**
  - Google have weighed in by using ML/deep learning to predict outcomes using electronic health records data (Rajkomar et al, NPJ Dig Med 2018)

- **Typical off-the-shelf methods include**
  - Random forests
  - Gradient boosted machines
  - Support vector machines
  - Neural networks
  - (Regression models with/without penalisation)?

- **Claims are that they offer flexibility in**
  - Capturing nonlinearities and higher order interactions
  - Good at handling high-dimensional data
  - Yet frequently used in low-dimensional settings

Classification is not prediction

- We're seeing an overemphasis on classification

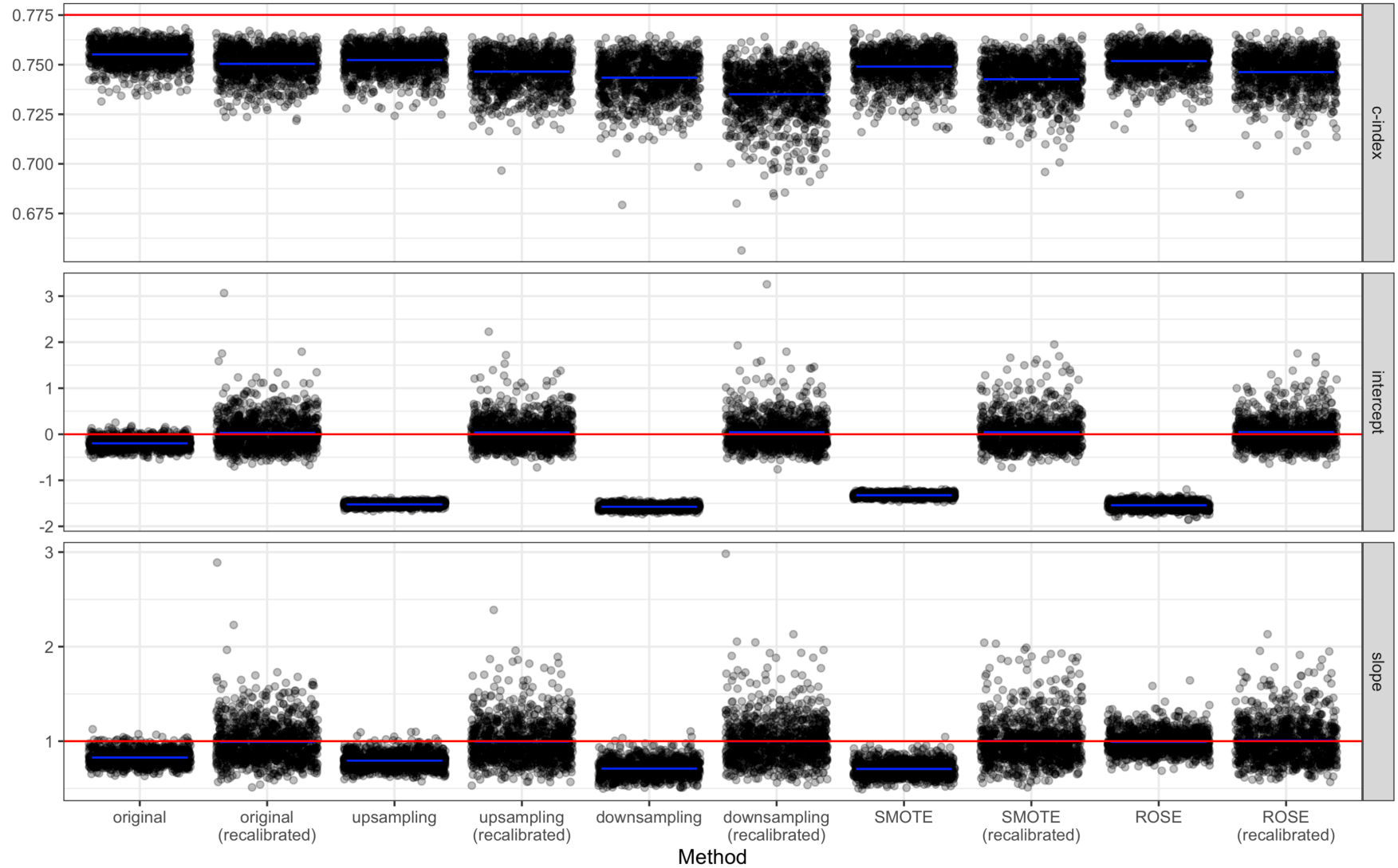
- Prediction (for diagnosis/prognosis) is about getting an individualised probability/risk of the outcome of interest (e.g., what is my risk of developing CVD over the next 10 years)

- Classification is placing an individual in a class/group
  - e.g., dead/alive, disease/no disease
  - (creates unnecessary problems such as 'class imbalance')

- We typically are (or should be) more interested in prediction
  - We can act on an predicted risk
    - e.g., send a patient for further testing or monitor
    - We can intervene to modify that risk (e.g., stop smoking)
    - Communicate this risk to the patient

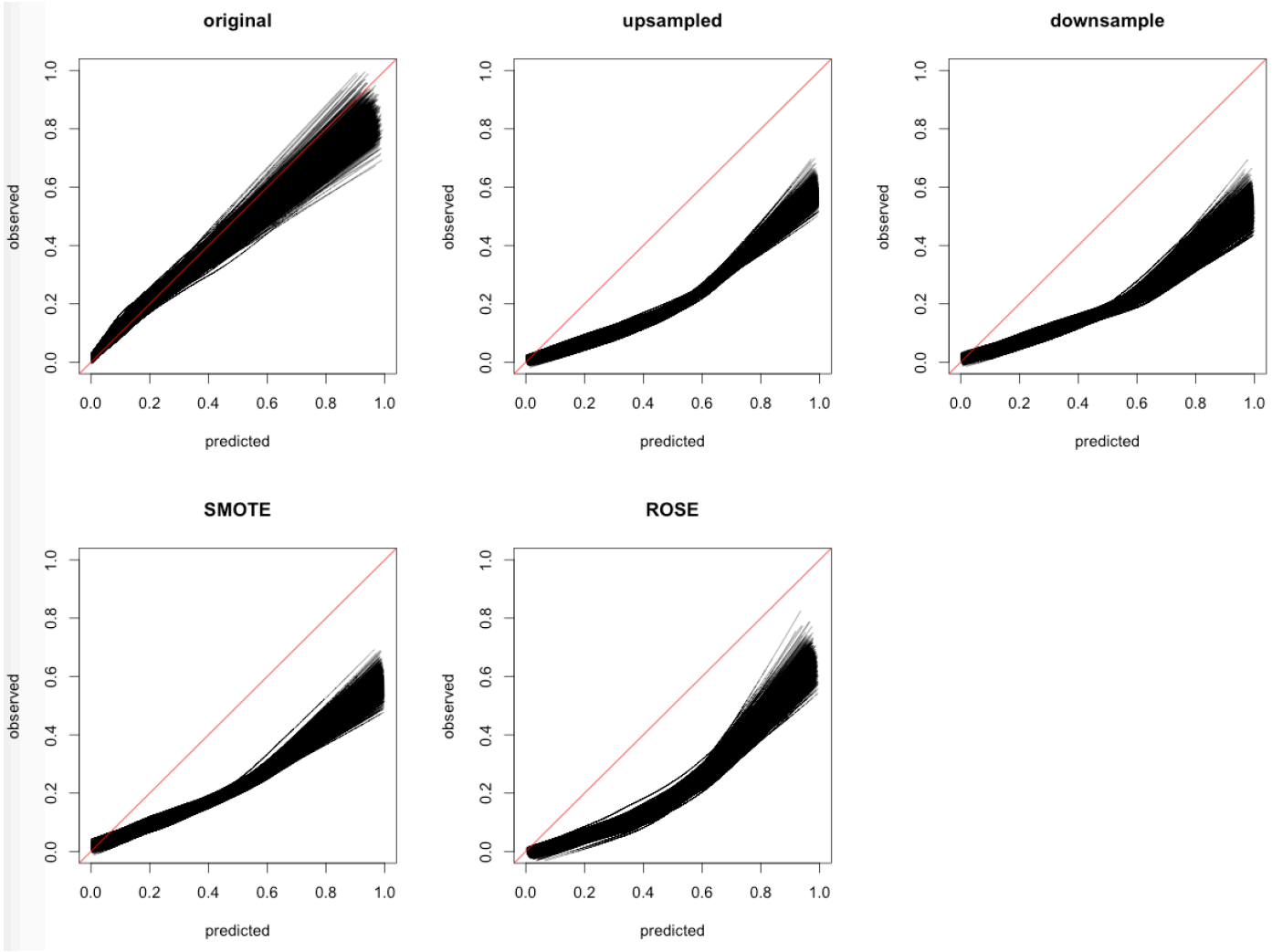
https://www.fharrell.com/post/classification/

Class imbalance



Original (simulated) data set had event fraction of 16%

Class imbalance



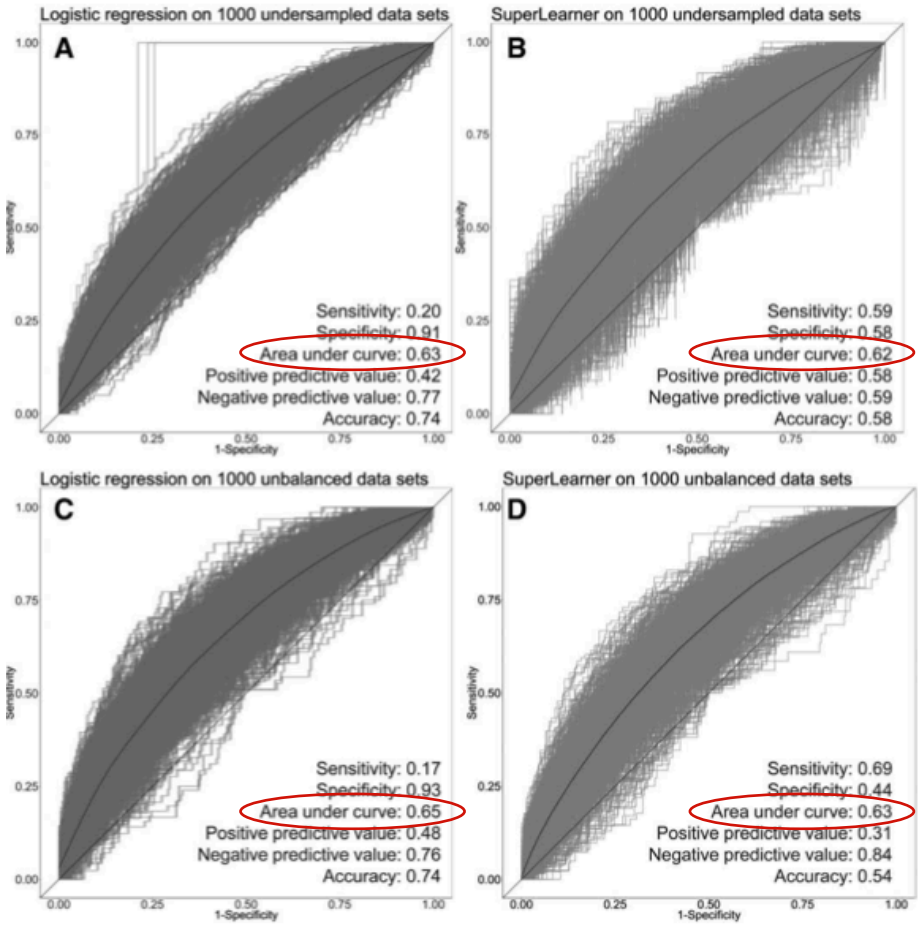
tion. In: Baker FB, Kim S-H, eds. *The Basics of Item Response Theory Using R. Statistics for Social and Behavioral Sciences*. Cham, Switzerland: Springer International Publishing; 2017:89–104.

logistic regression in imbalanced and undersampled data to investigate whether undersampling alters predictive accuracy.

mechanism although logistic regression was intentionally excluded from the SuperLearner library. In our simulations, undersampling did not dramatically improve predictive performance, suggesting that ensemble machine learning can achieve adequate performance in similar settings with moderate class imbalance. These results provide some insight on the optimal use of machine learning for predicting imbalanced outcomes. Example code to reproduce these analyses is available in the eSupplement; http://links.lww.com/EDE/B675.

DOI: 10.1097/EDE.0000000000001198

1,000 unbalanced samples parametrically



Reporting of machine learning models*

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

The completeness of reporting and adherence to the TRIPOD Statement of clinical prediction models using machine learning methods in oncology: a systematic review

Paula Dhiman[1,2], Jie Ma[1], Constanza Andaur Navarro[3], Beni Speich[1,4], Garrett Bullock[5], Shona Kirtley[1], Richard D Riley[6], Ben Van Calster[7], Karel GM Moons[3], Gary S Collins[1,2].

[1] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7LD, UK

[2] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

[3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[4] Department of Clinical Research, Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, University of Basel, Basel, Switzerland

[5] Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[6] Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire, UK, ST5 5BG

*To be submitted shortly

CSM
CENTRE FOR STATISTICS IN MEDICINE

Reporting of machine learning models

- 62 studies (prognostic models) in oncology published in 2019: 48 development, 14 development with external validation
  - Author defined 'machine learning'

- 48 binary outcome; 2 multinomial, 1 continuous and 11 time-to-event

- 36 predict risk 👍; 25 classify patients 👎, 1 unclear! 😕

- Mixture of Neural networks, random forests, CART, SVM, cox/logistic/linear regression(+/-penalisation), GBM, ensemble methods, …

Adherence to TRIPOD

Table 3. Median and range of reporting adherence to TRIPOD

	TRIPOD Adherence Score		
	n	**Median (%)**	**Range (%)**
Overall	62	41.38	10.34 to 66.67
Study design			
Development only	48	37.93	10.34 to 66.67
Development and validation	14	49.20	33.33 to 59.38
Number of models developed in study			
1	26	41.38	17.24 to 66.67
2	13	37.93	31.03 to 59.38
3	6	34.48	10.34 to 44.83
4	6	41.38	31.03 to 51.72
5	8	41.16	17.24 to 58.62
6	3	46.88	13.79 to 54.55

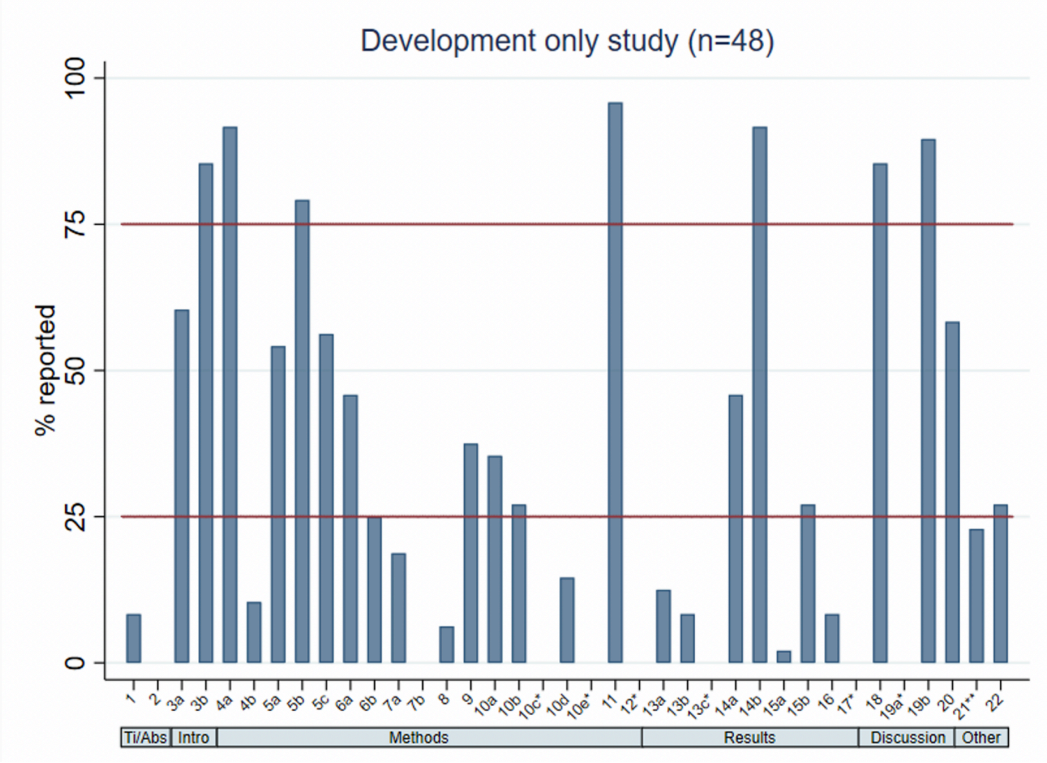
Digression: Are we inadvertently creating an opportunity for scientific fraud?

Consider the following hypothetical scenario…

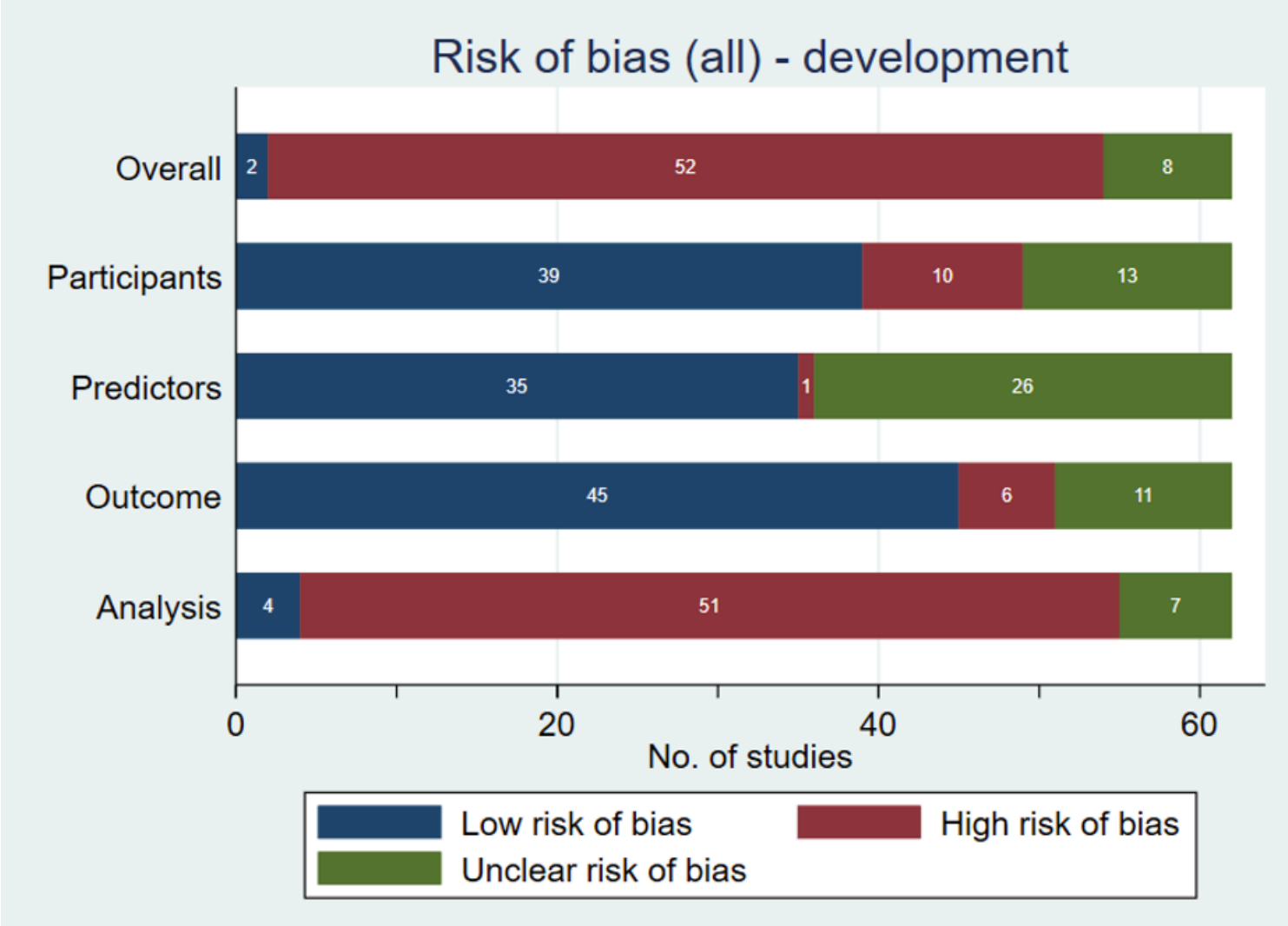
- A model has been developed
  - maybe multiple models for (and an unfair) comparison
- A paper has been prepared describing their development
- None of the models are presented in the paper
- The models are not made available in a software repository (e.g., via Github)
- The paper describes your favourite model as having excellent predictive accuracy
- The paper is published

Reporting deficiencies

- Item 4b - study dates
- Item 8 - Sample size
- Item 10b - model building/internal validation
- Item 13b - characteristics of participants
- Item 15a - model availability
- Item 16 - performance measures with CIs



Risk of bias assessment



NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

PLOS ONE

RESEARCH ARTICLE

A systematic review of machine learning models for predicting outcomes of stroke with structured data

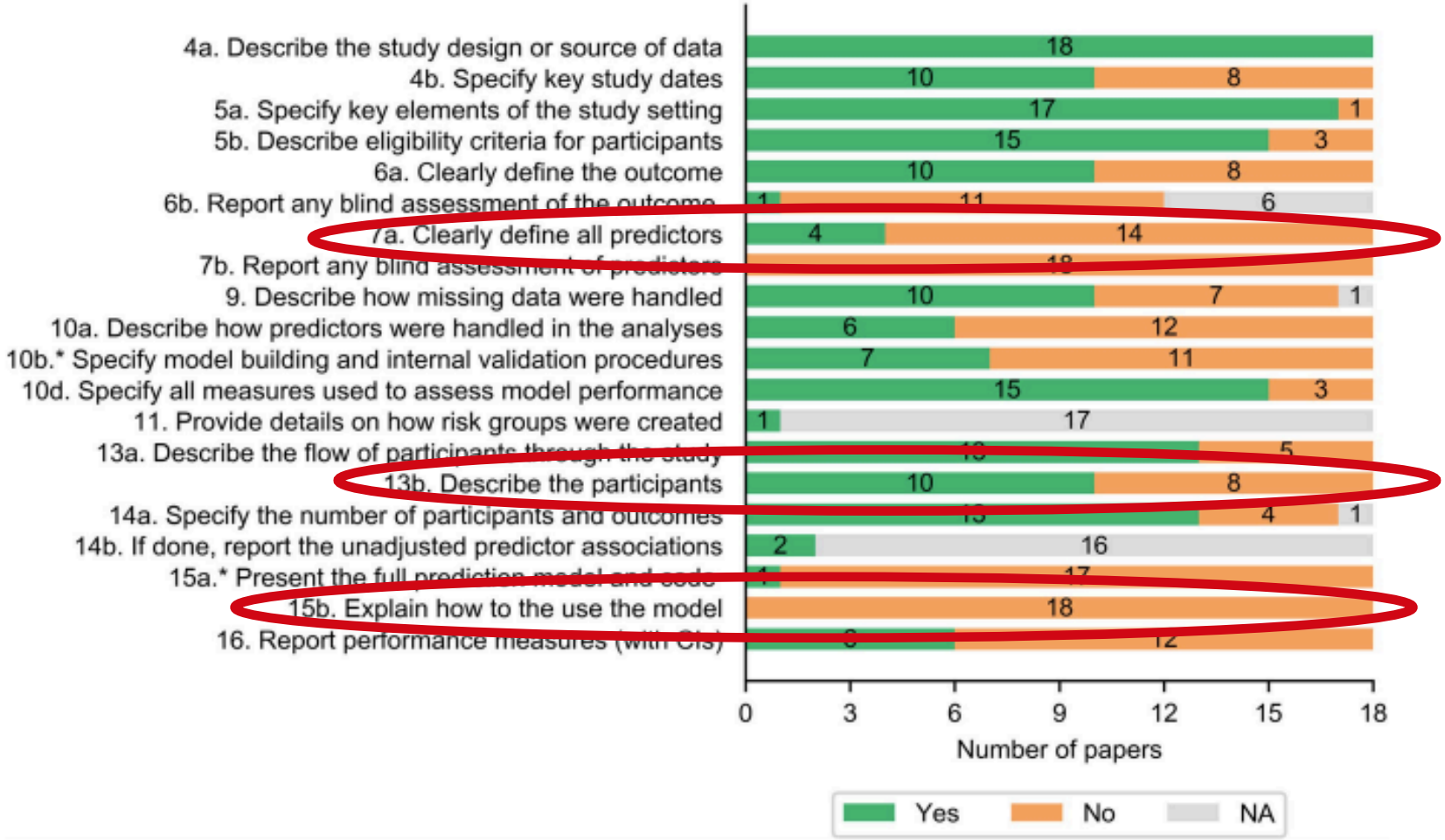
Wenjuan Wang[1]*, Martin Kiik[2], Niels Peek[3,4], Vasa Curcin[1,5,6], Iain J. Marshall[1], Anthony G. Rudd[1], Yanzhong Wang[1,5,6], Abdel Douiri[1,5,6], Charles D. Wolfe[1,5,6], Benjamin Bray[1]

1 School of Population Health & Environmental Sciences, Faculty of Life Science and Medicine, King's College London, London, United Kingdom, **2** School of Medical Education, Faculty of Life Science and Medicine, King's College London, London, United Kingdom, **3** Division of Informatics, Imaging and Data Science, School of Health Sciences, University of Manchester, Manchester, United Kingdom, **4** NIHR Manchester Biomedical Research Centre, Manchester Academic Health Science Centre, University of Manchester, Manchester, United Kingdom, **5** NIHR Biomedical Research Centre, Guy's and St Thomas' NHS Foundation Trust and King's College London, London, United Kingdom, **6** NIHR Applied Research Collaboration (ARC) South London, London, United Kingdom

* wenjuan.wang@kcl.ac.uk

Check for updates

CSM
CENTRE FOR STATISTICS IN MEDICINE


4a. Describe the study design or source of data
4b. Specify key study dates
5a. Specify key elements of the study setting
5b. Describe eligibility criteria for participants
6a. Clearly define the outcome
6b. Report any blind assessment of the outcome
7a. Clearly define all predictors
7b. Report any blind assessment of predictors
9. Describe how missing data were handled
10a. Describe how predictors were handled in the analyses
10b.* Specify model building and internal validation procedures
10d. Specify all measures used to assess model performance
11. Provide details on how risk groups were created
13a. Describe the flow of participants through the study
13b. Describe the participants
14a. Specify the number of participants and outcomes
14b. If done, report the unadjusted predictor associations
15a.* Present the full prediction model and code
15b. Explain how to the use the model
16. Report performance measures (with CIs)
0
3
6
9
12
15
18
Number of papers
Yes
No
NA

Shillan *et al. Critical Care* (2019) 23:284
https://doi.org/10.1186/s13054-019-2564-9

Critical Care

Table 2 Number and proportion of papers according to the type of machine learning used and number of patients analysed (for prediction studies only)

Type of machine learning	Number (%) of papers with this type[a]	Number of patients analysed					
		< 100	100–1000	1000–10,000	10,000–100,000	100,000–1,000,000	Number not reported
Neural network	72 (42.6%)	14 (19.4%)	27 (37.5%)	20 (27.8%)	9 (12.5%)	2 (2.8%)	0 (0.0%)
Support vector machine	40 (23.7%)	12 (30.0%)	15 (37.5%)	8 (20.0%)	4 (10.0%)	1 (2.5%)	0 (0.0%)
Classification/decision trees	35 (20.7%)	6 (17.1%)	11 (31.4%)	10 (28.6%)	5 (14.3%)	1 (2.9%)	2 (5.7%)
Random forest	21 (12.4%)	1 (4.8%)	9 (42.9%)	5 (23.8%)	4 (19.0%)	2 (9.5%)	0 (0.0%)
Naive Bayes/Bayesian networks	19 (11.2%)	4 (21.1%)	5 (26.3%)	6 (31.6%)	2 (10.5%)	1 (5.3%)	1 (5.3%)
Fuzzy logic/rough set	12 (7.1%)	3 (25.0%)	5 (41.7%)	2 (16.7%)	1 (8.3%)	0 (0.0%)	1 (8.3%)
Other techniques[b]	28 (16.7%)	2 (7.1%)	10 (35.7%)	8 (28.6%)	7 (25.0%)	1 (3.6%)	0 (0.0%)
Total (accounting for duplicates)	169	37 (21.9%)	56 (33.1%)	42 (24.9%)	26 (15.4%)	4 (2.37%)	4 (2.37%)

[a]Papers can have more than one approach—percentages may total more than 100
[b]Other techniques (number of studies): causal phenotype discovery (1), elastic net (1), factor analysis (1), Gaussian process (2), genetic algorithm (1), hidden Markov models (1), InSight (4); JITL-ELM (1), k-nearest neighbour (3), Markov decision process (1), particle swarm optimization (1), PhysiScore (1), radial domain folding (1), sequential contrast patterns (1), Superlearner (4), switching linear dynamical system (1), Weibull-Cox proportional hazards model (1), method not described (2)

and MEDLINE databases were searched to identify candidate articles: those on image processing were excluded. The

Key messages

Publication of papers reporting the use of machine learning to analyse routinely collected ICU data is increasing rapidly: around half of the identified studies were published since 2015.

Machine learning methods have changed over time. Neural networks are being replaced by support vector machines and random forests.

The majority of published studies analysed data on fewer than 1000 patients. Predictive accuracy increased with increasing sample size.

Reporting of the validation of predictions was variable and incomplete—few studies validated predictions using independent data.

Methodological and reporting guidelines may increase confidence in reported findings and thereby facilitate the translation of study findings towards routine use in clinical practice.

Caption

Logistic regression vs. Machine learning

ELSEVIER

Check for updates

Journal of Clinical Epidemiology 110 (2019) 12–22

Journal of Clinical Epidemiology

REVIEW

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

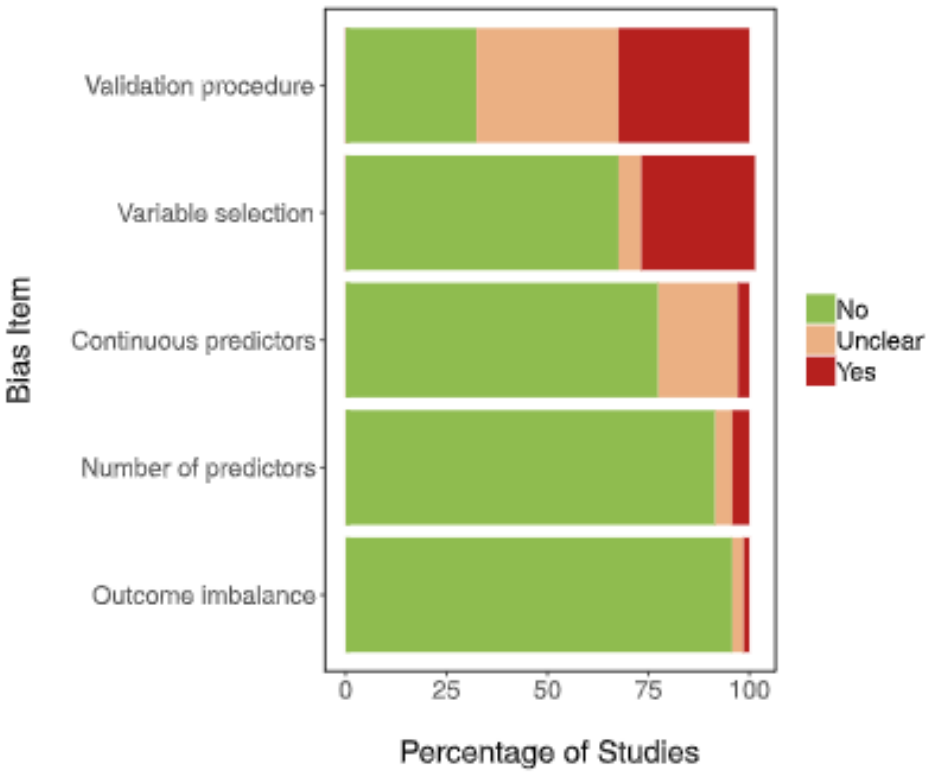
Evangelia Christodoulou[a], Jie Ma[b], Gary S. Collins[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

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

Accepted 5 February 2019; Published online 11 February 2019

What we found

- 71 studies comprising 282 comparisons (low-dimensional settings)
- Median sample size 1250 (range 72 to ~4m)
- Median no. of candidate predictors 19 (range 5-563)
- EPP (0.3 to 6697)



What we found

- Key details often inadequately described, including
  - Handling of continuous variables (for logistic regression)
    - 66% were unclear on how they were handled (including whether nonlinear associations were handled); 23% categorised all continuous variables
  - Interactions
    - 89% of studies did not explicitly mentioned where interactions were considered for the logistic regression models
  - Handling of missing data
    - 45% were unclear on how missing data were handled; 23% performed complete-case
  - Tuning of hyperparameters
    - 50% were unclear on how the tuning parameters were determined
  - Model performance
    - 90% of studies reported an assessment of model discrimination
    - 79% did not mention calibration (and for those that did, it was done based on grouping)

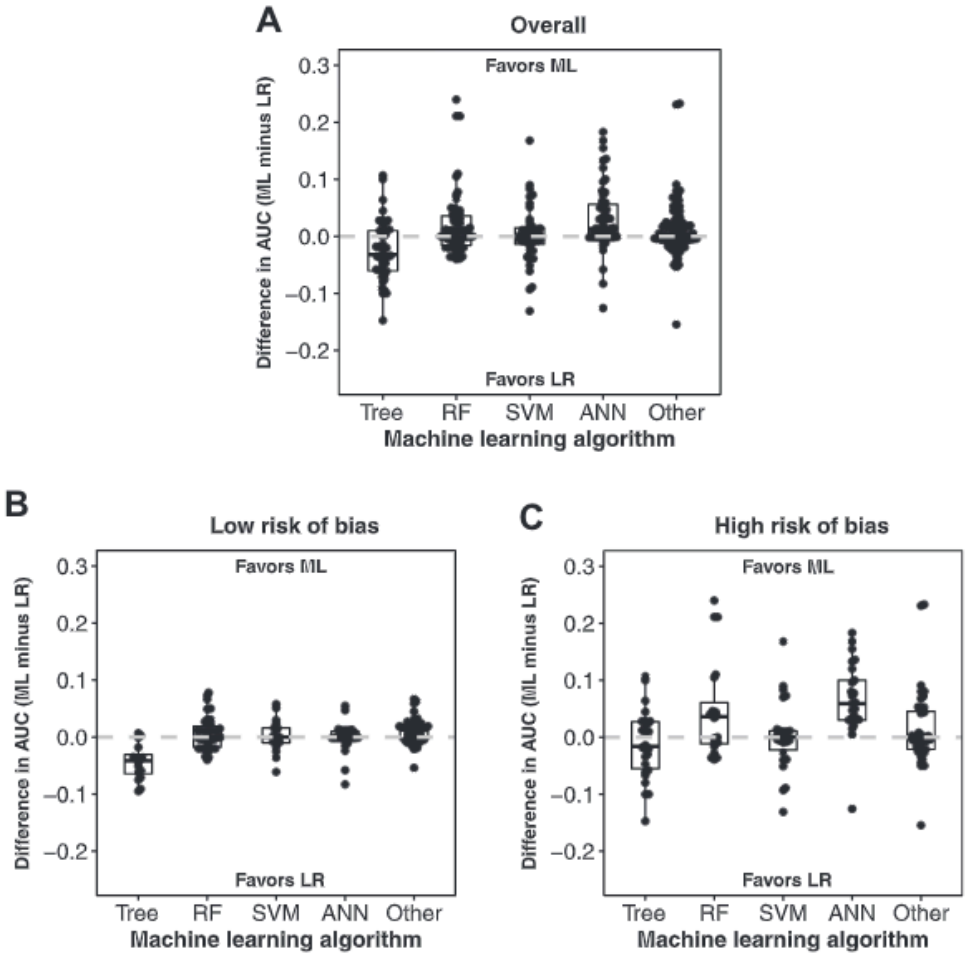


Fig. 3. Beeswarm plots of AUC difference (AUC of ML method minus AUC of LR) for all 282 comparisons by ML category, overall (A) and stratified by risk of bias (B). LR, logistic regression; ML, machine learning; RF, random forest; SVM, support vector machine; ANN, artificial neural network.

Meta-analysis of the AUC

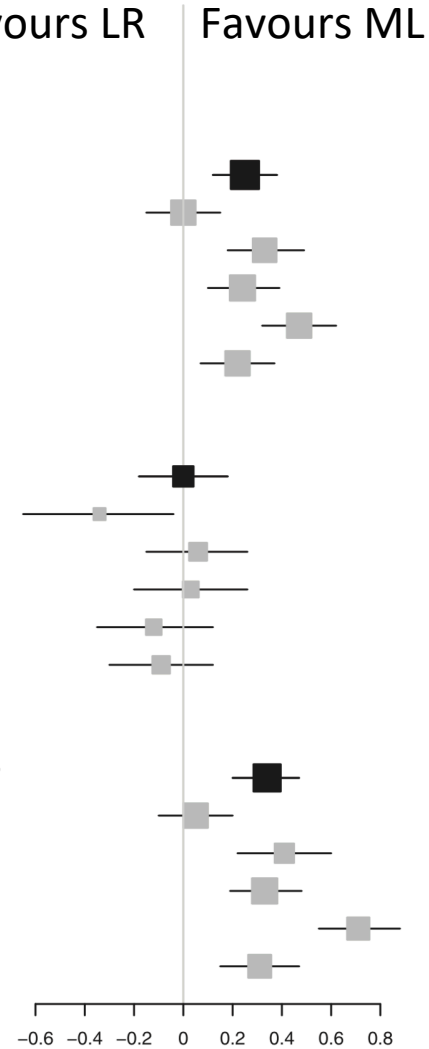
- **282 comparisons between LR and ML models**
  - AUC ranged between 0.52 and 0.97 (logistic regression)
  - AUC ranged between 0.58 and 0.99 (machine learning)

- **145 comparisons (51%) classified as low risk of bias**
  - logit(AUC) difference 0 (95% CI -0.18 to 0.18)

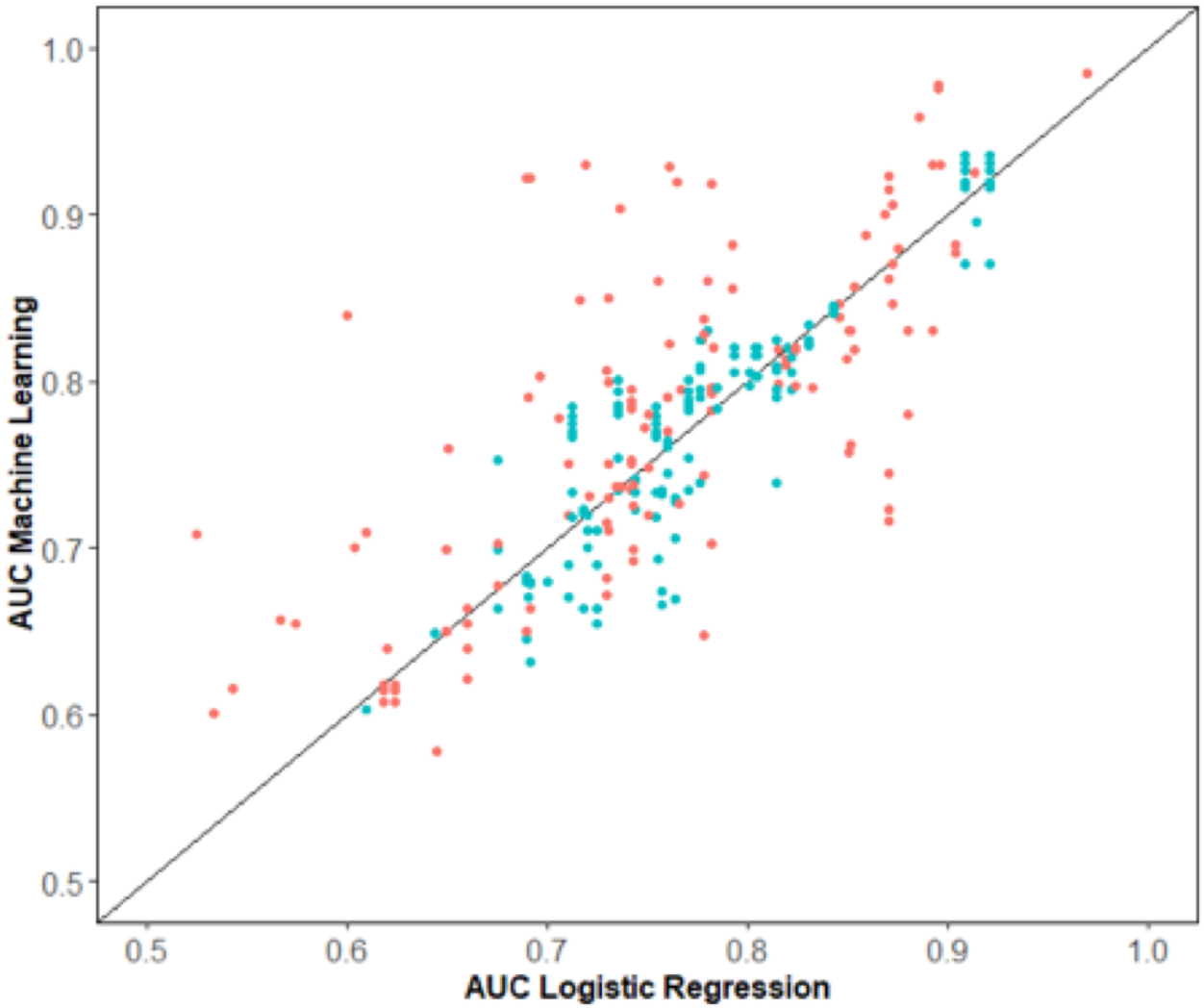
- **137 comparisons (51%) classified as high/unclear risk of bias**
  - logit(AUC) difference 0.34 (95% CI 0.20 to 0.47) [in favour of ML]

NB: logit(AUC) used to circumvent the bounded nature of the AUC

	Diff logit(AUC) (95% CI)	N
Overall		
– Any ML vs LR	0.25 (0.12;0.38)	282
– Tree vs LR	0.00 (–0.15;0.15)	42
– RF vs LR	0.33 (0.18;0.49)	59
– SVM vs LR	0.24 (0.10;0.39)	43
– ANN vs LR	0.47 (0.32;0.62)	52
– Other ML vs LR	0.22 (0.07;0.37)	86
Low risk of bias		
– Any ML vs LR	0.00 (–0.18;0.18)	145
– Tree vs LR	–0.34 (–0.65;–0.04)	16
– RF vs LR	0.06 (–0.15;0.26)	39
– SVM vs LR	0.03 (–0.20;0.26)	17
– ANN vs LR	–0.12 (–0.35;0.12)	27
– Other ML vs LR	–0.09 (–0.30;0.12)	46
High risk of bias		
– Any ML vs LR	0.34 (0.20;0.47)	137
– Tree vs LR	0.05 (–0.10;0.20)	26
– RF vs LR	0.41 (0.22;0.60)	20
– SVM vs LR	0.33 (0.19;0.48)	26
– ANN vs LR	0.71 (0.55;0.88)	25
– Other ML vs LR	0.31 (0.15;0.47)	40



Low RoB (cyan), High RoB (red)



Miles *et al. Diagnostic and Prognostic Research* (2020) 4:16
https://doi.org/10.1186/s41512-020-00084-1

Diagnostic and Prognostic Research

RESEARCH Open Access

Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review

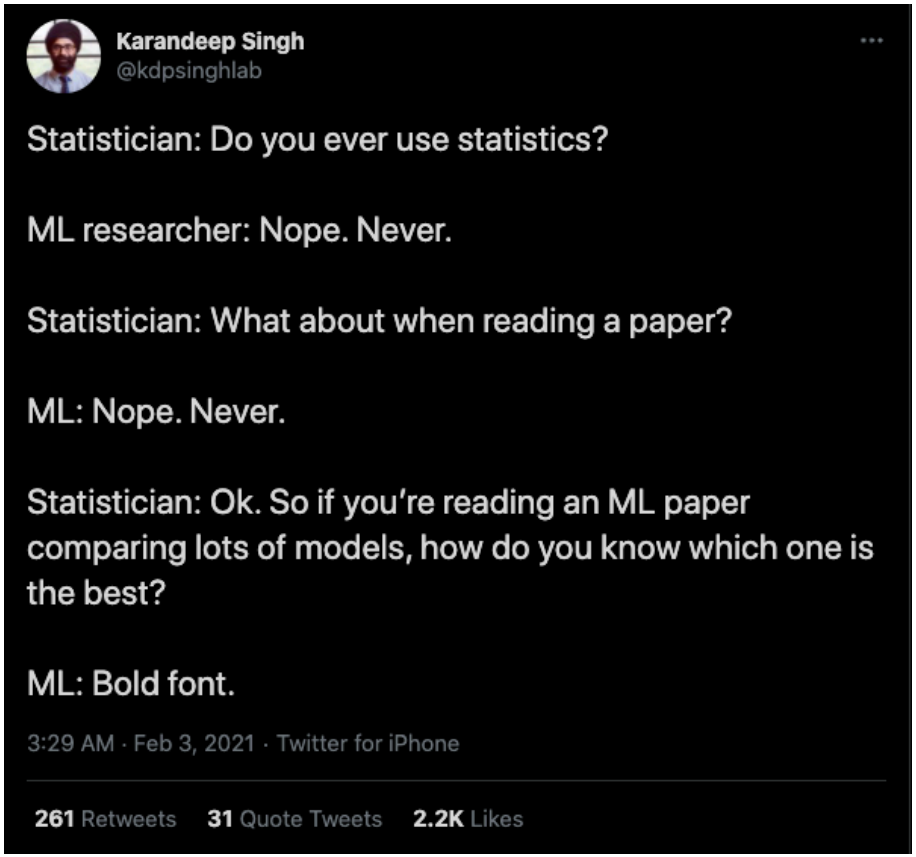
Jamie Miles[1*], Janette Turner[2], Richard Jacques[2], Julia Williams[3] and Suzanne Mason[2]

Conclusions: Machine-learning methods appear accurate in triaging undifferentiated patients entering the Emergency Care System. There was no clear benefit of using one technique over another; however, models derived by logistic regression were more transparent in reporting model performance. Future studies should adhere to reporting guidelines and use these at the protocol design stage.

Machine learning: comparative studies

- Characterised (often) by unfair comparisons
  - Expertise bias (domain knowledge)
    - Nothing new: Duin (1996), Salzberg (1997), Hand (2006)
    - Researcher typically favours one approach over another
  - Software default values used?
  - (often incorrect) focus on classification
    - Do we want to classify individuals as to whether they experience a CVD event within 10 years, or are we interested in the probability of experiencing a CVD event within 10 years?
  - Inadequate assessment of model performance

=>High risk of bias



Importance of fair comparisons

Received: 15 August 2017 | Revised: 20 October 2017 | Accepted: 22 October 2017
DOI: 10.1002/bimj.201700129

LETTER TO THE EDITOR

Biometrical Journal

On the necessity and design of studies comparing statistical methods

In data analysis sciences in general and in biometrical research particularly, there are strong incentives for presenting work that entails new methods. Many journals require authors to propose new methods as a prerequisite for publication, as this is the most straightforward way to claim the necessary novelty. The development of new methods is also factually often a sine qua non condition to be recruited as a faculty member or to obtain personnel funding from a methods-oriented research agency, not least because it noticeably increases the chance to get published as outlined above. Thus, in statistical research and related methodology-oriented fields such as machine learning or bioinformatics, the well-known adage "publish or perish" could be translated into "propose new methods or perish."

Such a research paradigm is not favorable for studies that aim at meaningfully comparing alternative existing methods or, more generally, studies assessing the behavior and properties of existing methods. Yet, given the exponential increase in the number and complexity of new statistical methods being published every year, the end users are often at a loss regarding what are the "optimal" or even "appropriate" methods to answer the research question of interest given a particular data structure. It becomes more and more difficult to get an overview of existing methods, not to mention the overview of their respective performances in different settings (Sauerbrei, Abrahamowicz, Altman, Le Cessie, & Carpenter, 2014).

Moreover, it is well known that studies comparing a suggested new method to existing methods may be (strongly) biased in favor of the new method. This is a consequence of various factors starting with the authors' better expertise on the new method compared to the competing methods. Another factor is the combination of publication pressure (publish or perish) and *publication bias*—in the sense that a new method performing worse than existing ones has (severe) difficulties to get published (Boulesteix, Stierle, & Hapfelmeier, 2015). This may lead to simulation designs that might be—intentionally or unintentionally—biased. Note that not only empirical evaluations but also theoretical properties suggesting the superiority of a method under particular assumptions may be in principle potentially affected by this kind of bias. Deriving theoretical results for statistical approaches relevant in practice is extremely difficult and possible only under strong assumptions (Picard & Cook, 1984). We speculate that authors assessing the theoretical properties of their new method tend to make assumptions that are rather favorable for the new method—also a form of bias.

In contrast, *neutral* comparison studies, as defined by Boulesteix, Wilson, and Hapfelmeier (2017a), are dedicated to the comparison itself: they do not aim to demonstrate the superiority of a particular method and are thus not designed in a way that may increase the probability to observe incorrectly this superiority. Furthermore, they involve authors who are, as a collective, approximately equally competent on all considered methods. Neutral comparison studies can be thus considered as unbiased.

Boulesteix et al, Biom J, 2017

Received: 29 November 2017 | Revised: 23 August 2018 | Accepted: 2 November 2018
DOI: 10.1002/sim.8086

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

Using simulation studies to evaluate statistical methods

Tim P. Morris[1] | Ian R. White[1] | Michael J. Crowther[2]

[1] London Hub for Trials Methodology Research, MRC Clinical Trials Unit at UCL, London, United Kingdom
[2] Biostatistics Research Group, Department of Health Sciences, University of Leicester, Leicester, United Kingdom

Correspondence
Tim P. Morris, MRC Clinical Trials Unit at UCL, London, United Kingdom.
Email: tim.morris@ucl.ac.uk

Present Address
Tim P. Morris, 90 High Holborn, London WC1V 6LJ, United Kingdom.

Funding information
Medical Research Council, Grant/Award Number: MC_UU_12023/21, MC_UU_12023/29, and MR/P015433/1

Simulation studies are computer experiments that involve creating data by pseudo-random sampling. A key strength of simulation studies is the ability to understand the behavior of statistical methods because some "truth" (usually some parameter/s of interest) is known from the process of generating the data. This allows us to consider properties of methods, such as bias. While widely used, simulation studies are often poorly designed, analyzed, and reported. This tutorial outlines the rationale for using simulation studies and offers guidance for design, execution, analysis, reporting, and presentation. In particular, this tutorial provides a structured approach for planning and reporting simulation studies, which involves defining aims, data-generating mechanisms, estimands, methods, and performance measures ("ADEMP"); coherent terminology for simulation studies; guidance on coding simulation studies; a critical discussion of key performance measures and their estimation; guidance on structuring tabular and graphical presentation of results; and new graphical presentations. With a view to describing recent practice, we review 100 articles taken from Volume 34 of *Statistics in Medicine*, which included at least one simulation study and identify areas for improvement.

KEYWORDS
graphics for simulation, Monte Carlo, simulation design, simulation reporting, simulation studies

Morris et al, Stat Med, 2018

Evaluating performance of machine learning

- **Traditional prediction model literature is relatively clear on key aspects to assess model performance, namely**
  - Discrimination (separation of individuals with/out event)
  - Calibration (accuracy of predictions)
  - Clinical utility (decision curve analysis)
- **Calibration often ignored in the 'traditional prediction model' literature (and often evaluated poorly)**
- **Calibration rarely assessed in the ML prediction literature (and often evaluated poorly)**
  - Often as a consequence of focusing on classification
  - Calibration often has a different meaning in ML prediction literature
- **In ML: recall (sensitivity)/precision (PPV), F-scores**
  - Requiring some dichotomisation of the predicted outcome (often at the 0.5 probability threshold) => creates the so-called class-imbalance problem
  - Be very sceptical when you see very high AUCs - particularly those that are substantially higher for one method compared to another

Machine learning

- **No one approach is likely to be universally 'best'**
- **Need to think about setting, context and moment of implementation**
  - A machine learning model with many predictors (the situations where it is claimed to have usefulness) unlikely to be useful in many settings
- **Need to think about mechanisms for independent evaluation and expect this as routine practice**
- **'Validations' should be meaningful**

TRIPOD for machine learning/AI

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

THE LANCET

COMMENT | VOLUME 393, ISSUE 10181, P1577-1579, APRIL 20, 2019

Reporting of artificial intelligence prediction models

Gary S Collins • Karel G M Moons

Published: April 20, 2019 • DOI: https://doi.org/10.1016/S0140-6736(19)30037-6 • Check for updates

References

Article Info

Figures

Data-driven technologies that form the basis of the digital health-care revolution provide potentially important opportunities to deliver improvements in individual care and to advance innovation in medical research. Digital health technologies include mobile devices and health apps (m-health), e-health technology, and intelligent monitoring. Behind the digital health revolution are also methodological advancements using artificial intelligence and machine learning techniques. Artificial intelligence, which encompasses machine learning, is the scientific discipline that uses computer algorithms to learn from data, to help identify patterns in data, and make predictions. A key feature underpinning the excitement behind artificial intelligence and machine learning is their potential to analyse large and complex data structures to create prediction models that personalise and improve diagnosis, prognosis, monitoring, and administration of treatments, with the aim of improving individual health outcomes. Prediction models to support clinical decision making have existed for decades, and these include well known tools such as the Framingham Risk Score,[1] QRISK3,[2] Model for End-stage Liver Disease,[3] ABCD[2] score,[4] and the Nottingham Prognostic Index.[5] Health-care professionals, medical researchers, policy makers, guideline developers, patients, and members of the general public are all

CSM
CENTRE FOR STATISTICS IN MEDICINE

TRIPOD challenge: availability

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES
UNIVERSITY OF OXFORD

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, Em
Atlanta, GA 30322, USA; cecile.janssens@emory.edu; Tel.: +1-404-

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

Abstract: Direct-to-consumer genetic testing companies ain
using proprietary algorithms. Companies keep algorithms a
but a market that thrives on the premise that customers can
testing should respect customer autonomy and informed deci

rate (to their

- Commercial exploitation, p

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 interoperability. 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).

CSM
CENTRE FOR STATISTICS IN MEDICINE

Model availability + independent evaluation

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

Check for updates

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[1,2], Massive Analysis Quality Control (MAQC) Society Board of Directors*, Levi Waldron[8], Bo Wang[2,3,9,10], 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/

Journal of the American Medical Informatics Association, 26(12), 2019, 1651–1654
doi: 10.1093/jamia/ocz130
Advance Access Publication Date: 2 August 2019
Perspective

Perspective

Predictive analytics in health care: how can we know it works?

Ben Van Calster,[1,2] Laure Wynants,[1] Dirk Timmerman,[1,3] Ewout W Steyerberg,[2] and Gary S Collins[4,5]

Table 1. Summary of arguments in favor of making predictive algorithms fully available, hurdles for doing so, and reasons why developers choose to hide and sell algorithms

Why should predictive algorithms be fully and publicly available?	• Facilitate external validation and assessment of heterogeneity in performance • Facilitate uptake of algorithm by researchers and clinicians, avoid research waste • Facilitate updating for specific settings • For publicly funded research, this makes research results available to the community
Recommendations to maximize algorithm availability	• Report the full equation of a predictive algorithm, where possible (eg, regression-based models); this includes reporting of the intercept, or baseline hazard information for time-to-event regression models • When making an algorithm available online or via a mobile app, provide relevant and complete background information • For complex algorithms (eg, black-box machine learning), provide software to facilitate implementation and large-scale validation studies
Potential reasons why developers might choose to hide and sell algorithms	• Generate income for further research • More control over how people use an algorithm • Facilitate FDA approval or CE certification, because a commercial entity can be identified • To install a profitable business model

COMPUTER SCIENCE

Artificial intelligence faces repr is

Unpubli
make m

The most basic problem is that research-
ers often don't
the AAAI meeti
computer scien
versity of Scienc
heim, reported
400 algorithms
top AI conferen
found that only
the algorithm's c
data they tested
half shared "ps
mary of an algo
is also absent fi
journals, includi

Researchers say there are many reasons for the missing details: The code might be a work in progress, owned by a company, or held tightly by a researcher eager to stay ahead of the competition. It might be dependent on other code, itself unpublished. Or it might be that the code is simply lost, on a crashed disk or stolen laptop—what Rougier calls the "my dog ate my program" problem.

WHO and ITU establish benchmarking process for artificial intelligence in health

Growing populations, demographic changes, a
shortage of health practitioners have placed pres
on the health-care sector. In parallel, increasing am
of digital health data and information have be
available. Artificial intelligence (AI) models that
from these large datasets are in development and
the potential to assist with pattern recognitior
classification problems in medicine—for example,

requirements are met, AI models can be submitted via an online platform to be evaluated with the test data. Established in this way, the benchmarking process will not only provide a reliable, robust, and independent evaluation system that can demonstrate the quality of AI models, but will also provide an independent test dataset for model validation consistent with best-practice recommendations for reporting multivariable prediction models in health.[4]

Harmonisation of two languages

Statistics	Machine learning
Covariates	Features
Outcome variable	Target
Model	Network, graphs
Parameters	Weights
Model for discrete var.	Classifier
Model for continuous var.	Regression
Log-likelihood	Loss
Multinomial regression	Softmax
Measurement error	Noise
Subject/observation	Sample/instance
Dummy coding	One-hot encoding
Measurement invariance	Concept drift

Statistics	Machine learning
Prediction	Supervised learning
Latent variable modeling	Unsupervised learning
Fitting	Learning
Prediction error	Error
Sensitivity	Recall
Positive predictive value	Precision
Contingency table	Confusion matrix
Measurement error model	Noise-aware ML
Structural equation model	Gaussian Bayesian network
Gold standard	Ground truth
Derivation–validation	Training–test
Experiment	A/B test

Slide taken from Maarten van Smeden, Biometrischen Kolloquium, 2020

Consensus statement

- Delphi about to be launched
  - Interested in participating in the Delphi then contact me
- Anticipate TRIPOD-AI to be not too dissimilar to the original TRIPOD
- Biggest difference will be in the terminology, examples, and methods guidance

Last few slides…a missed opportunity?

An opportunity to take centre(ish) stage, but…

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

RESEARCH

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)

CSM
CENTRE FOR STATISTICS IN MEDICINE

Update 3 (1-July-2020)

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES
UNIVERSITY OF OXFORD

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

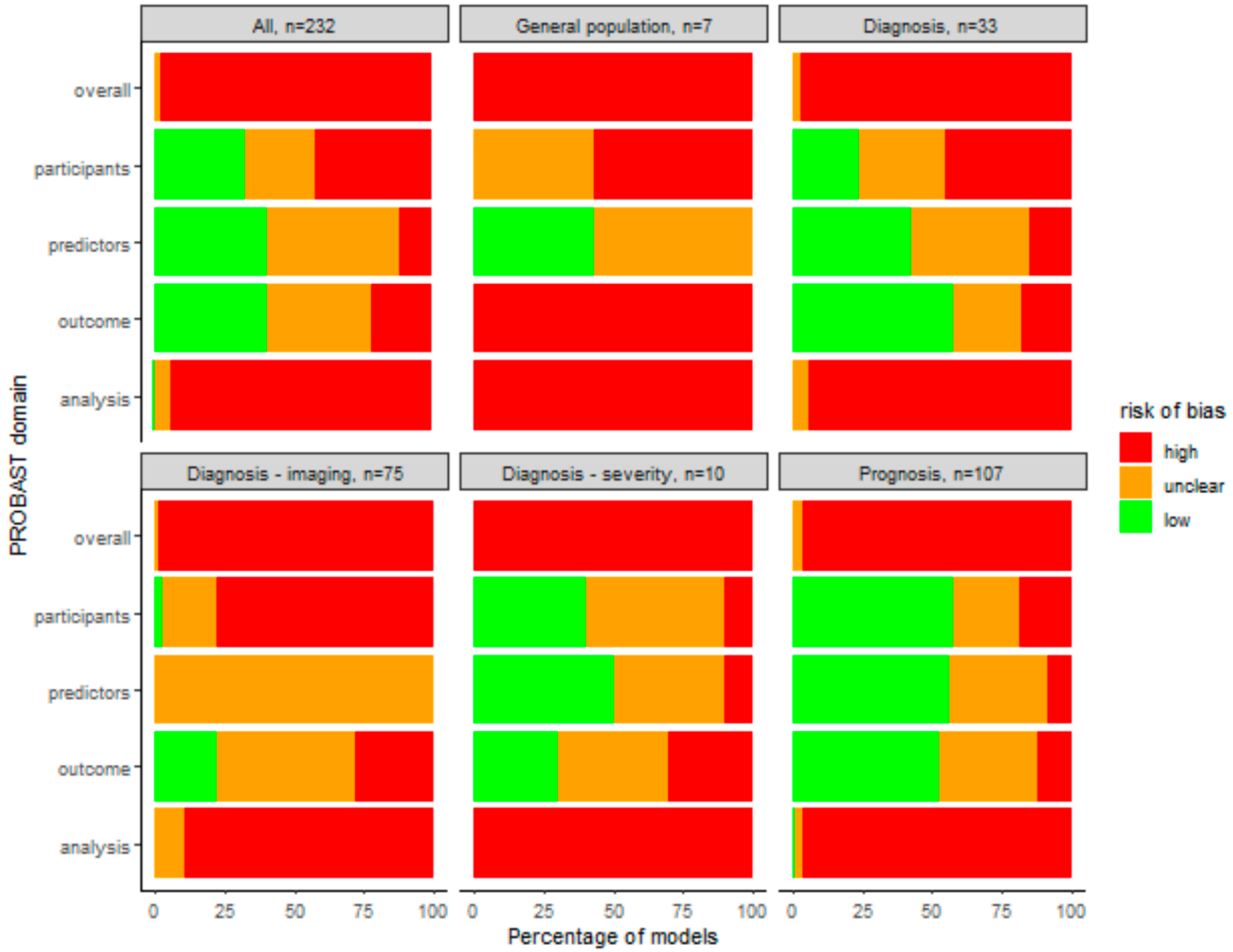
- Reported c-index values ranged from
  - 0.71 to 0.99 (risk scores)
  - 0.65 to 0.99 (diagnostic models)
  - 0.54 to 0.99 (prognostic models)

- Calibration rarely assessed (and often incorrectly)

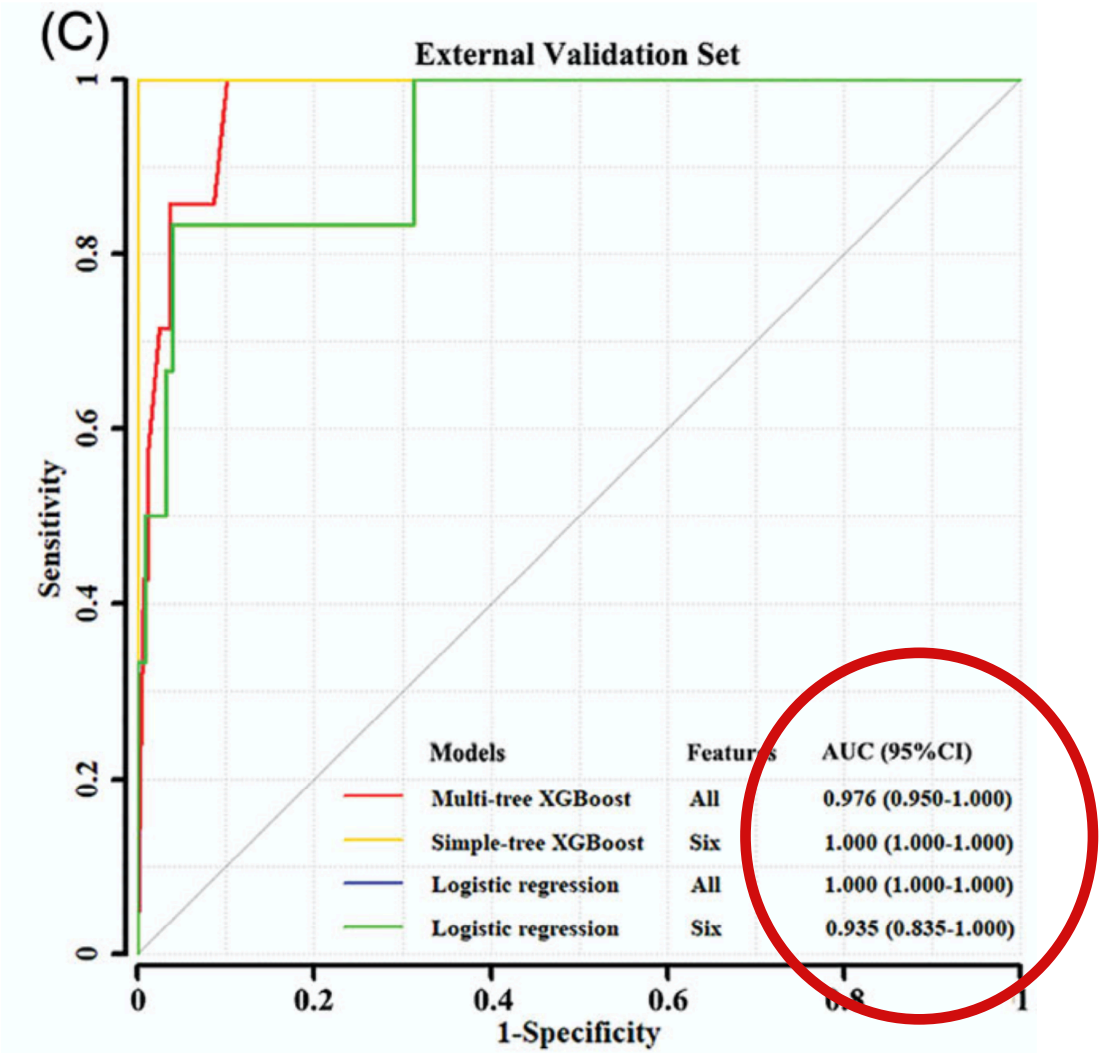
- Bottom line: 226 at high risk of bias; 6 at unclear risk of bias

Some concerns

- Inappropriate or unclear study design
- Selection of controls often unclear
- Proxy outcomes (e.g., hospital admission due to severe respiratory disease - absence of covid-19 patients)
- Dichotomisation of continuous predictors
- Inappropriate in- or exclusion of study participants
  - Participants excluded because they did not experience the outcome by the end of the study
- Predictor measurements also part of the outcome
- Lack of internal or external validation; small to modest sample size; overfitting
- Other issues (not part of the RoB assessment) include changing populations (case-mix)


All, n=232
General population, n=7
Diagnosis, n=33
Diagnosis - imaging, n=75
Diagnosis - severity, n=10
Prognosis, n=107
PROBAST domain
overall
participants
predictors
outcome
analysis
0
25
50
75
100
Percentage of models
risk of bias
high
unclear
low

Predicting covid mortality



Sample size
- n=279
- #events= 7

No calibration

Guan et al, Ann Med 2021

Letters to the editor

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES
UNIVERSITY OF OXFORD

Clinical Infectious Diseases
CORRESPONDENCE
IDSA Infectious Diseases Society of America · hivma hiv medicine association · OXFORD

Flaws in the Development and Validation of a Coronavirus Disease 2019 Prediction Model

To the Editor—The coronavirus disease 2019 (COVID-19) pandemic has seen the development of a number of clinical prediction models to support assessing disease severity or aiding prognosis. A recent systematic review identified 145 models and concluded that all were at high risk of bias, citing concerns with data quality, statistical analysis, and reporting, leading to the conclusion (which will likely be overestimated [6]). Other major analysis concerns include categorization of continuous predictors (which results in loss of information [7]), no mention of missing data, use of lasso followed by "multivariate" [sic] Cox regression to screen predictors for inclusion, incorrect (ie, does not reflect the actual model building process) and confusing implementation of cross-validation on the validation data, weak assessment of model calibration by binning observations, and assessment of both the area under the curve and the

submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

Gary S. Collins,[1] Richard D. Riley,[2] and Maarten van Smeden[3]

[1]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom, [2]Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, United Kingdom, and [3]Julius Center for Health Science and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands

References

1. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. BMJ **2020**; 369:m1328.

DOI: 10.1002/jmv.26390
LETTER TO THE EDITOR
JOURNAL OF MEDICAL VIROLOGY · WILEY

Statistical issues in the development of COVID-19 prediction models

To the Editor,

Clinical prediction models to aid diagnosis, assess disease severity, or prognosis have enormous potential to aid clinical decision making during the coronavirus disease 2019 (COVID-19) pandemic. A living systematic review has, so far, identified 145 COVID-19 prediction models published (or preprinted) between 3 January and 5 May 2020. Despite the considerable interest in developing COVID-19 prediction models, the review concluded that all models to date, with no exception, are at high risk of bias with concerns related to data quality, flaws in the statistical analysis, and poor reporting, and none

Another concern is the actual model. The final model contains seven predictors and the authors have fully reported this permitting individualized prediction. However, an obvious and major concern is the regression coefficient reported for procalcitonin, with a value of 48.8309 and accompanying odds ratio with a confidence interval of ">999.999 (>999.999, >999.999)" (sic). This is clearly nonsensical, and to put it bluntly, makes the model unusable. The reason for the large regression value (standard error and confidence interval) is due to an issue called *separation*.[5] This occurred because there was little or no overlap in the procalcitonin values between individuals with mild and severe dis

AGORA
CORRESPONDENCE

Received: 27 July 2020 | Revised: 2 August 2020 | Accepted: 4 August 2020
DOI: 10.1111/tbed.13828
LETTER TO EDITOR
Transboundary and Emerging Diseases · WILEY

There are no shortcuts in the development and validation of a COVID-19 prediction model

A recent living systematic review has identified 145 COVID-19 prediction models published up until May 2020, to support clinical decision-making during the global COVID-19 pandemic (Wynants et al., 2020). Despite this surge in developing prediction models, the systematic review concluded that all these models are at high risk of bias citing concerns regarding poor data quality, flaws in the statistical analysis and incomplete or poor reporting. As a consequence, validation, and no external validation (i.e. evaluating the model in a separate data set), is a major limitation.

Other concerns include the data quality, namely the presence and handling of missing data. Missing values are largely unavoidable, and the study by Dong included 30 predictors—in the absence of any mention of missing data, one can only assume that individuals with missing were excluded from the analysis—such an approach

COVID-19 prediction models should adhere to methodological and reporting standards

To the Editor:

The coronavirus disease 2019 (COVID-19) pandemic has led to a proliferation of clinical prediction models to aid diagnosis, disease severity assessment and prognosis. A systematic review has identified 66 COVID-19 prediction models: concluding that all, with no exception, are at high risk of bias due to concerns surrounding the data quality, statistical analysis and reporting, and none are recommended for use [1]. Therefore, we read with interest the recent paper by Wu *et al.* [2] describing the development of a model to identify COVID-19 patients with severe disease on admission to facilitate triage. However, our enthusiasm was dampened by a number of concerns surrounding the design, analysis and reporting of the

CSM
CENTRE FOR STATISTICS IN MEDICINE

Noteworthy covid-19 prediction models

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

RESEARCH

OPEN ACCESS

Check for updates

FAST TRACK

Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score

Stephen R Knight,[1] Antonia Ho,[2,3] Riinu Pius,[1] Iain Buchan,[4] Gail Carson,[5] Thomas M Drake,[1] Jake Dunning,[6,7] Cameron J Fairfield,[1] Carrol Gamble,[8] Christopher A Green,[9] Rishi Gupta,[10] Sophie Halpin,[8] Hayley E Hardwick,[11] Karl A Holden,[11] Peter W Horby,[5] Clare Jackson,[8] Kenneth A Mclean,[1] Laura Merson,[5] Jonathan S Nguyen-Van-Tam,[12] Lisa Norman,[1] Mahdad Noursadeghi,[13] Piero L Olliaro,[14] Mark G Pritchard,[14] Clark D Russell,[15] Catherine A Shaw,[1] Aziz Sheikh,[1] Tom Solomon,[11,16] Cathie Sudlow,[17] Olivia V Swann,[18] Lance CW Turtle,[11,19] Peter JM Openshaw,[7] J Kenneth Baillie,[20,21] Malcolm G Semple,[11,22] Annemarie B Docherty,[1,21] Ewen M Harrison,[1,23] on behalf of the ISARIC4C investigators

RESEARCH

OPEN ACCESS

Check for updates

Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study

Ash K Clift,[1] Carol A C Coupland,[2] Ruth H Keogh,[3] Karla Diaz-Ordaz,[3] Elizabeth Williamson,[3] Ewen M Harrison,[4] Andrew Hayward,[5] Harry Hemingway,[6] Peter Horby,[7] Nisha Mehta,[8] Jonathan Benger,[9] Kamlesh Khunti,[10] David Spiegelhalter,[11] Aziz Sheikh,[4] Jonathan Valabhji,[12] Ronan A Lyons,[13] John Robson,[14] Malcolm G Semple,[15] Frank Kee,[16] Peter Johnson,[12] Susan Jebb,[1] Tony Williams,[17] Julia Hippisley-Cox[1]

CSM
CENTRE FOR STATISTICS IN MEDICINE

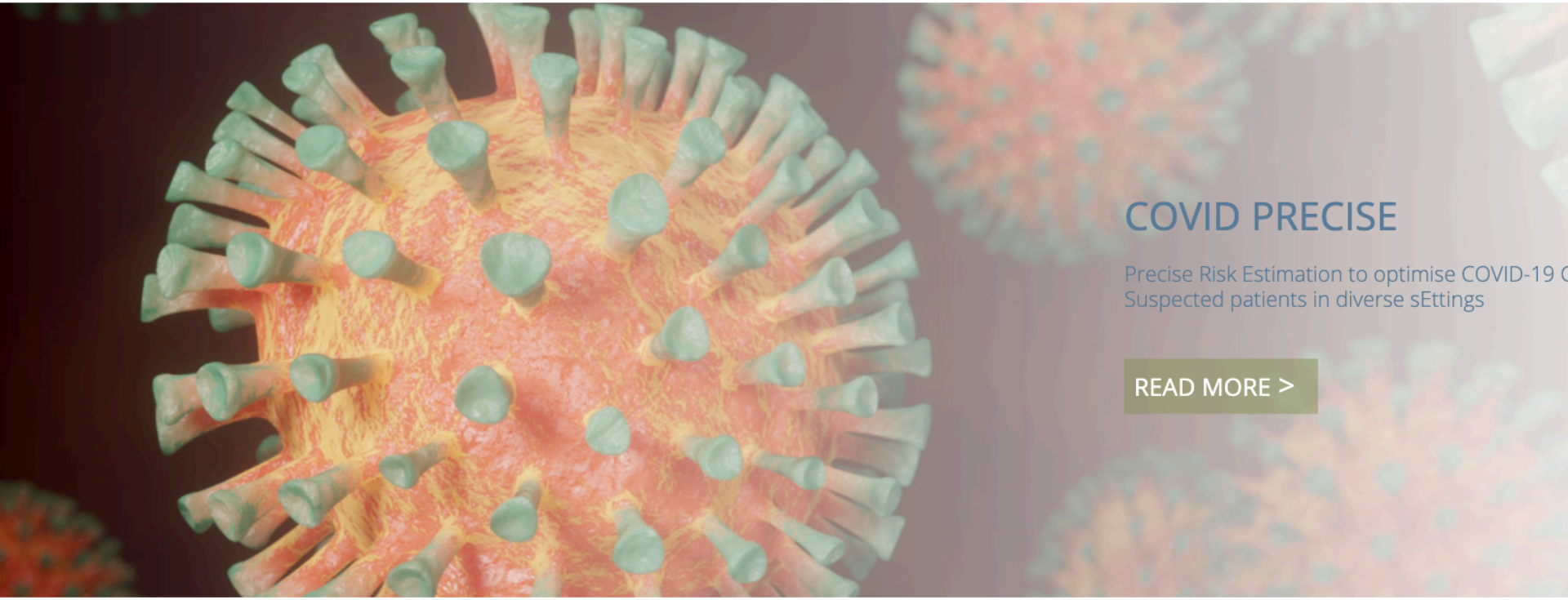
covidprecise.org

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

COVID-19 PRECISE

HOME PROJECT LIVING REVIEW IPDMA



LATEST NEWS

The most comprehensive systematic review of all COVID-

KEY DOCUMENTS

PROBAST Tool

ALTMETRIC

CENTRE FOR STATISTICS IN MEDICINE

Summary - crisis?

- **Glass half empty**
  - Deluge of low quality, poorly reported prediction models shows no sign of abating -> research waste
  - Not learning (enough) from earlier mistakes / concerns
  - Potential to cause harm

- **Glass half full**
  - Few models actually being used -> patients not being potentially harmed
  - Interest in prediction at an all time high -> we will get it right more often (I hope) -> improve patient outcomes

Machine learning and AI for patient benefit

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

RESEARCH METHODS AND REPORTING

Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness

Sebastian Vollmer,[1,2] Bilal A Mateen,[1,3,4] Gergo Bohner,[1,2] Franz J Király,[1,5] Rayid Ghani,[6] Pall Jonsson,[7] Sarah Cumbers,[8] Adrian Jonas,[9] Katherine S L McAllister,[9] Puja Myles,[10] David Grainger,[11] Mark Birse,[11] Richard Branson,[11] Karel G M Moons,[12] Gary S Collins,[13] John P A Ioannidis,[14] Chris Holmes,[1,15] Harry Hemingway[16,17,18]

For numbered affiliations see end of the article.

Correspondence to: C Holmes cholmes@stats.ox.ac.uk (ORCID 0000-0002-6667-4943)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;368:l6927
http://dx.doi.org/10.1136/bmj.l6927

Accepted: 22 October 2019

Machine learning, artificial intelligence, and other modern statistical methods are providing new opportunities to operationalise previously untapped and rapidly growing sources of data for patient benefit. Despite much promising research currently being undertaken, particularly in imaging, the literature as a whole lacks preliminary solution here) is the current lack of best practice guidance specific to machine learning and artificial intelligence. However, we believe that interdisciplinary groups pursuing research and impact projects involving machine learning and artificial intelligence for health would benefit from explicitly addressing a series of

CSM
CENTRE FOR STATISTICS IN MEDICINE

Scandal of Poor Medical Research

NDORMS
NUFFIELD DEPARTMENT OF ORTHOPAEDICS, RHEUMATOLOGY AND MUSCULOSKELETAL SCIENCES

UNIVERSITY OF OXFORD

unacceptable.

What, then, should we think about researchers who use the wrong techniques (either wilfully or in ignorance), use the right techniques wrongly, misinterpret their results, report their results selectively, cite the literature selectively, and draw unjustified conclusions? We should be appalled. Yet numerous studies of the medical literature, in both general and specialist journals, have shown that all of the above phenomena are common.[1-7] This is surely a scandal.

When I tell friends outside medicine that many papers

phenomena are common.[1-7] This is surely a scandal.

When I tell friends outside medicine that many papers published in medical journals are misleading because of methodological weaknesses they are rightly shocked. Huge sums of money are spent annually on research that is seriously flawed through the use of inappropriate designs, unrepresentative samples, small samples, incorrect methods of analysis, and faulty interpretation. Errors are so varied that a whole book on the topic,[7] valuable as it is, is not comprehensive; in any case, many of those who make the errors are unlikely to read it.

Altman BMJ 1994

CSM
CENTRE FOR STATISTICS IN MEDICINE