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 study which deserve highlighting to readers.

Our first point relates to design. The authors randomly split their dataset into a training and test set. This been long been shown to be an inefficient use of the data [3], reducing the size of the training set (increasing the risk of model overfitting), and creating a test set too small for model evaluation. There are alternative stronger approaches that use the entire data to both develop and internally validate a model based on cross-validation or bootstrapping [3]. This naturally leads us to further elaborate on the sample size. The sample size in a prediction model study is largely influenced by the number of individuals experiencing the event to be predicted (in the study by Wu *et al.* [2], those with severe disease). Using published sample size formulae for developing prediction models [4, 5], based on information reported in the study by Wu *et al.* [2] (75 predictors, outcome prevalence of 0.237), then depending on the anticipated model R-squared, the minimum sample size in the most optimistic scenario (*e.g.* that the model gives the highest R-squared) would be 1285 individuals (306 events). To precisely estimate the intercept alone requires 279 individuals (66 events). After splitting their data, the authors developed their model with a sample size of 239 individuals (57 events): clearly insufficient to estimate even the model intercept, let alone develop a prediction model.

The test set was then used to evaluate the performance of their model comprising 60 individuals of whom ~14 experienced the event. To put this in perspective, current sample size recommendations to evaluate model performance suggest a minimum of 100 events [6]. The performance of the model was also evaluated separately in each of five external validation datasets where the number of events ranged from 7 to 98, none of which meet this minimum requirement.

Other concerns include the handling of missing data; it is hard to believe all patients had complete information on all 75 predictors, and indeed the flow chart reveals 38 individuals with missing data were simply excluded, which can led to bias [7]. Continuous predictors were assumed to be linearly associated with the outcome, which can reduce predictive accuracy. Model overfitting (a clear concern given the small sample size) was not addressed either in adjusting the performance measures for optimism or shrinking the regression coefficients that are likely overestimated (*e.g.* using penalisation techniques [8]). "Synthetic sampling" was used to address imbalanced data, but this is inappropriate since artificially balancing data will produce an incorrect estimation of the model intercept (unless it is re-adjusted post-estimation), leading to incorrect model predictions (miscalibration). Model performance was poorly and inappropriately assessed, including presenting a confusion matrix (inappropriate for evaluating prediction models [8]), reporting sensitivity/specificity (where net benefit would be more informative [9]), and

@ERSpublications
COVID-19 prediction models should adhere to methodological and reporting standards
https://bit.ly/3ebnook

Cite this article as: Collins GS, van Smeden M, Riley RD. COVID-19 prediction models should adhere to methodological and reporting standards. *Eur Respir J* 2020; 56: 2002643 [https://doi.org/10.1183/13993003.02643-2020].

assessing model calibration using weak and again discredited approaches (*e.g.* Hosmer–Lemeshow test, rather than calibration plots with graphical loess curves [6]). We also question the arbitrary choice of risk groupings, and why individuals with a predicted risk of 0.21 are considered the same ("middle risk") as those with a predicted risk of 0.80.

Arguably the most important aspect of a prediction model article is the presentation of the model so that others can use or evaluate it in own their own setting. The authors have presented a nomogram and (prematurely) linked to a web calculator. Whilst both these formats can be used to apply the model to individual patients (though given our concerns we urge against this), for independent validation the prediction model needs to be reported in full; namely, all the regression coefficients and the intercept [10], but these are noticeably absent.

Finally, the authors followed the STARD checklist for reporting their study, but this is not the correct guideline. STARD is for reporting diagnostic test accuracy studies, and not multivariable clinical prediction models. We urge the authors and other investigators developing (COVID-19) prediction models to consult the TRIPOD Statement (www.tripod-statement.org) for key information to report when describing their prediction model study, so that readers have the minimal information required to judge the quality of the study [10]. The accompanying TRIPOD explanation and elaboration paper describes the rationale of the importance of transparent reporting, but also discusses various methodological considerations [6].

Gary S. Collins [1], Maarten van Smeden[2] and Richard D. Riley[3]
[1]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. [2]Julius Center for Health Science and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands. [3]Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Keele, UK.

Correspondence: Gary S. Collins, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. E-mail: gary.collins@csm.ox.ac.uk

Received: 5 July 2020 | Accepted: 6 July 2020

Conflict of interest: None declared.

Support statement: This work was supported by Cancer Research UK (grant C49297/A27294). G.S. Collins was supported by the NIHR Biomedical Research Centre, Oxford. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Wynants L, Van Calster B, Collins GS, *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020; 369: m1328.
2 Wu G, Yang P, Xie Y, *et al.* Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *Eur Respir J* 2020; 56: 2001104.
3 Steyerberg EW, Harrell FE Jr, Borsboom GJ, *et al.* Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; 54: 774–781.
4 Riley RD, Ensor J, Snell KIE, *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; 368: m441.
5 Riley RD, Snell KI, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. *Stat Med* 2019; 38: 1276–1296.
6 Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162: W1–W73.
7 Sterne JAC, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; 338: b2393.
8 Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd Edn. New York, Springer, 2019.
9 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; 352: i6.
10 Collins GS, Reitsma JB, Altman D, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis: the TRIPOD statement. *Ann Intern Med* 2015; 162: 55–63.

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

From the authors:

We would like to thank G.S. Collins, M. van Smeden, and R.D. Riley for their commentary on the design, analysis, and reporting of our article [1]. However, their comments seem to stem from a traditional biostatistics angle rather than from a translational research machine-learning approach and the overwhelming majority of criticisms arise from either misunderstandings or misreading.

The authors inaccurately state that we randomly split datasets. As described in our manuscript we nonrandomly split the data by time and place, making it a stronger design according to the TRIPOD statement. The use of independent cohorts to test model generalisability make it a TRIPOD type 3 study [2]. We agree that splitting reduces the training dataset size, increasing the probability of overfitting. However, as an RNA virus, SARS-CoV-2 may be able to mutate rapidly and develop diverse characteristics. Hence, we split the datasets by time and place rather than using cross-validation or bootstrapping.

The authors used 75 candidate predictors rather than the seven selected ones to perform their sample size calculations for our training dataset [3]. Although we agree that using candidate predictors is a more rigorous approach compared to using only the selected ones, it is too strict in the modern machine-learning and -omics field, and disregards the power of feature dimensionality reduction and selection methods we employed. While we understand that overfitting remains possible, the validation of the model on five datasets from unrelated institutions strengthens the likelihood that the model presented is robust. Test set results are presented separately to improve understanding of robustness, because it is easy to hide possible poor performance in a small test set by combining it with a large test set where the performance is good. More importantly, the selected variables make sense from the clinical point of view [4, 5], making our models explainable, transparent, and therefore acceptable by the end-users.

We agree that excluding missing data may lead to biases, and list this as our first limitation in the Discussion. Given the time-critical nature of this quickly developing pandemic, we decided that excluding 38 patients was preferable to imputation and that the bias introduced by such a selection would be revealed in the five external validations and further validations post-publication. The authors inaccurately state that we assume that continuous predictors are linearly associated with the outcome. We emphasise that neither feature selection nor modelling assume a linear association between predictors and outcomes. The process of randomising the outcomes and re-running of the analysis is a powerful sanity check against overfitting [6].

We must point out that the Adaptive Synthetic (ADASYN) algorithm is a published and validated method for dealing with dataset unbalance. Whilst we agree that this methodology could introduce an error in the model intercept, we believe that this error can be estimated when calculating the model's performance in the five external validation datasets. Everyone has their preferred metrics and often a better metric can be found than those commonly reported. This is especially true in the convergence zone between machine-learning and clinical application, where reporting possibly suboptimal metrics that are easier to understand may have added benefit over more technical metrics used by data scientists. Reporting confusion matrices, a widely used and readily understandable way of evaluating classification performance, can easily be defended. Equally, reporting the universally adopted sensitivity and specificity metrics as well as the results from the calibration plots align well with the readership of this esteemed publication.

@ERSpublications
It is hard to follow a standardised methodology for prediction models, while researchers should adhere to generally accepted reporting standards according to research needs and journal submission requirements https://bit.ly/30zfMIw

Cite this article as: Wu G, Woodruff HC, Chatterjee A, *et al*. Reply to "COVID-19 prediction models should adhere to methodological and reporting standards". *Eur Respir J* 2020; 56: 2002918 [https://doi.org/10.1183/13993003.02918-2020].

The authors call our risk groupings arbitrary. Using three risk groups was a requirement of the clinicians and is common in the clinic, including COVID-19: low-risk (home care), medium-risk (hospital surveillance), and high-risk (ICU admission). The risk probability thresholds were based on the 25th and 75th probability percentiles in the balanced training set. With these thresholds, the low-risk group had <20% incidence of severe outcomes, and the high-risk group had >75% chance of severe outcomes on each test set, which the clinicians deemed clinically useful. The authors reprimand us for not reporting the model parameters explicitly. For us, the main aim of any clinical triage model is the application on individual patients in a clinical setting. We believe both a nomogram and a web calculator satisfy this requirement. In addition, for model evaluation, the model parameters can be fully reconstructed from the nomogram.

There are numerous checklists or guidelines for diagnostic and predictive models [7–10]. In retrospect, we agree that TRIPOD is a more appropriate checklist than STARD for modelling studies due to the details regarding the reporting of methodology and results. We chose a more familiar checklist from the submission guidelines of this journal (guidelines in which TRIPOD was not listed) and will ensure to also include TRIPOD reporting in the future. Given the quickly changing nature of machine learning and the increasing number of guidelines, it is hard to forge standards, while the need for them in the reporting of model studies increases.

Overall, we believe our work is useful and explainable, and have received positive feedback from colleagues, including clinicians, who appreciate that their requirements have been taken into account. We are currently prospectively validating our models out of a conviction that only this approach can truly validate a predefined model.

Guangyao Wu[1], Henry C. Woodruff[1,2], Avishek Chatterjee[1] and Philippe Lambin[1,2]
[1]The D-Lab, Dept of Precision Medicine, GROW - School for Oncology, Maastricht University Medical Center+, Maastricht, The Netherlands. [2]Dept of Radiology and Nuclear Medicine, GROW- School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands.

Correspondence: Guangyao Wu, The D-Lab, Dept of Precision Medicine, GROW - School for Oncology, Maastricht University Medical Center+, 6229 ER, Maastricht, The Netherlands. E-mail: g.wu@maastrichtuniversity.nl

Received: 27 July 2020 | Accepted: 30 July 2020

Conflict of interest: Dr. Wu has nothing to disclose. Dr. Woodruff has (minority) shares in Oncoradiomics, outside the submitted work. Dr. Chatterjee has nothing to disclose. Dr. Lambin has minority shares in The Medical Cloud Company, and reports grants from Varian Medical, Oncoradiomics, ptTheragnostic/DNAmito and Health Innovation Ventures, personal fees from Oncoradiomics, BHV, Varian, Elekta, ptTheragnostic and Convert Pharmaceuticals, outside the submitted work; and has patents PCT/NL2014/050248, PCT/NL2014/050728 and PCT/EP2014/059089 licensed, and patents N2024482, N2024889 and N2024889 pending.

References

1 Wu G, Yang P, Xie Y, *et al.* Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *Eur Respir J* 2020; 56: 2001104.
2 Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162: W1–W73.
3 Riley RD, Snell KI, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. *Stat Med* 2019; 38: 1276–1296.
4 Zhou F, Yu T, Du R, *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020; 395: 1054–1062.
5 Yang X, Yu Y, Xu J, *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med* 2020; 8: 475–481.
6 Chatterjee A, Vallières M, Dohan A, *et al.* An empirical approach for avoiding false discoveries when applying high-dimensional radiomics to small datasets. *IEEE Trans Radiat Plasma Med Sci* 2018; 3: 201–209.
7 Luo W, Phung D, Tran T, *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; 18: e323.
8 Handelman GS, Kok HK, Chandra RV, *et al.* Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *AJR Am J Roentgenol* 2019; 212: 38–43.
9 Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018; 286: 800–809.
10 Bluemke DA, Moy L, Bredella MA, *et al.* Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers from the Radiology editorial board. *Radiology* 2020; 294: 487–489.