EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Correspondence

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

Gary S. Collins, Maarten van Smeden, Richard D. Riley

Please 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; in press (https://doi.org/10.1183/13993003.02643-2020).

This manuscript has recently been accepted for publication in the *European Respiratory Journal*. It is published here in its accepted form prior to copyediting and typesetting by our production team. After these production processes are complete and the authors have approved the resulting proofs, the article will move to the latest issue of the ERJ online.

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

Gary S. Collins, PhD

Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Windmill Road, Oxford OX3 7LD, United Kingdom. Email: gary.collins@csm.ox.ac.uk

Maarten van Smeden, PhD

Julius Center for Health Science and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, the Netherlands.

Richard D. Riley, PhD

Centre for Prognosis Research, School of Primary, Community and Social Care,

Keele University, Staffordshire, ST5 5BG, United Kingdom.

Word count: 804

The authors report no conflict of interest.

Letter Re: Development of a clinical decision support system for severity risk prediction and triage of COVD-19 patients at hospital admission: an international multicentre study.

The 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 sixty-six covid-19 prediction models – concluding 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 and colleagues describing the development of a model to identify covid-19 patients with severe disease on admission to facilitate triage [2]. 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 in 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 Wu's study, those with severe disease). Using published sample size formulae for developing prediction models [4, 5], based on information reported in the Wu study (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, all not meeting 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 penalization 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 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].

REFERENCES

1. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Damen JAA, Debray TPA, De Vos M, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Kreuzberger N, Lohmann A, Luijken K, Ma J, Andaur Navarro CL, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM, Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, van Kuijk SMJ, 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, Woodruff HC, Rao X, Guiot J, Frix A-N, Louis R, Moutschen M, Li J, Li J, Yan C, Du D, Zhao S, Ding Y, Liu B, Sun W, Albarello F, D'Abramo A, Schininà V, Nicastri E, Occhipinti M, Barisione G, Barisione E, Halilaj I, Lovinfosse P, Wang X, Wu J, Lambin P. Development of a Clinical Decision Support System for Severity Risk Prediction and Triage of COVID-19 Patients at Hospital Admission: an International Multicenter Study. *Eur Respir J* 2020; : 2001104.

3. Steyerberg EW, Harrell Jr FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. 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, Harrell FE, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; 368: m441.

5. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine* 2019; 38: 1276–1296.

6. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. 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, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. 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 edition. 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, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis: The TRIPOD statement. *Ann Intern Med* 2015; 162: 55–63.