

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 are recommended for use.¹ Disappointingly, the recent study by Yang et al² describing the development of a prediction model to identify COVID-19 patients with severe disease is no different. The study has failed to report important information needed to judge the study findings, but numerous methodological problems are apparent.²

Our first point relates to the sample size. The sample size requirements in a prediction model study are largely influenced by the number of individuals experiencing the event to be predicted (in Yang's study, those with mild COVID-19 disease, as this is the smaller of the two outcome categories). Using published sample size formulae for developing prediction models,³ based on information reported in the Yang study (40 predictors, outcome prevalence of 0.489), the minimum sample size in the most optimistic scenario would be 538 individuals (264 events). To precisely estimate the intercept alone requires 384 individuals (188 events). The study by Yang included 133 individuals, where 65 had the outcome of mild disease, substantially lower than required.

Developing a prediction model with a small sample size and a large number of predictors will result in a model that is overfit, including unimportant or spurious predictors, and overestimating the regression coefficients. This means that the model will appear to fit the data (used in its development) too well—leading to a model that has poor predictive accuracy in new data. An important step in all model development studies is to carry out an internal validation of the model building process (using either bootstrapping or cross-validation), whereby the overestimation in regression coefficients can be determined and shrunk as well as estimating the optimism in model performance.⁴ This important step is absent in the study of Yang, who reported an area under the curve of 0.8842 in the same data used to develop their model—this will almost certainly be substantially overestimated.


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 disease. The statistical software used by the authors, SAS, will report odds ratios as greater than 999 when this occurs. Instead of retaining this in the model as is, one preferred approach would be to use Firth's correction, available in both SAS and R5. The authors used the model to develop an early warning score—this score has not been presented by the authors—and we caution against such an approach with a preference for alternative formats that permit estimation of absolute risk.⁶

Other concerns include the handling of missing data. While the authors mention discarding observed values with more than 20% missing—it is unclear whether individuals were omitted, or whether entire predictors were omitted. Regardless, one can only assume a complete-case analysis was conducted in preference for more suitable approaches using multiple imputations.⁷ Finally, we note the use of univariate screening, whereby predictors are omitted based on the lack of statistical association. This approach is largely discredited, as predictors can be spuriously retained or omitted.⁸

We urge the authors and other investigators developing (COVID-19) prediction models to read the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement (www.tripod-statement.org) for key information to report when describing their study so that readers have the minimal information required to judge the quality of the study.⁹ The accompanying TRIPOD explanation and elaboration paper describes the rationale of the importance of transparent reporting, examples of good reporting, but also discusses methodological considerations.¹⁰ Until improved methodological standards are adopted, we should not expect prediction models to benefit patients, and should consider the possibility that they might do more harm than good.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

Gary S. Collins PhD¹ 
Jack Wilkinson PhD²

¹Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, University of Oxford, Oxford, UK

²Division of Population Health, Health Services Research and Primary Care, Centre for Biostatistics, University of Manchester, Manchester, UK

Correspondence

Gary S. Collins, PhD, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, University of Oxford, Windmill Rd, Oxford OX3 7LD, UK.

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

Funding Information

Cancer Research UK, Grant/Award Number: C49297/A27294; NIHR Biomedical Research Centre, Oxford

ORCID

Gary S. Collins  <http://orcid.org/0000-0002-2772-2316>

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. <https://doi.org/10.1136/bmj.m1328>
2. Yang P, Wang P, Song Y, Zhang A, Yuan G, Cui Y. A retrospective study on the epidemiological characteristics and establishment of an early warning system of severe COVID-19 patients [published online ahead of print June 2, 2020]. *J Med Virol*. 2020. <https://doi.org/10.1002/jmv.26022>
3. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. <https://doi.org/10.1136/bmj.m441>
4. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Switzerland: Springer; 2019.
5. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. *Am J Epidemiol*. 2017;187:864-870. <https://doi.org/10.1093/aje/kwx299>
6. Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. *BMJ*. 2019;365:l737. <https://doi.org/10.1136/bmj.l737>
7. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
8. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol*. 1996;49(8):907-916.
9. 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.
10. Moons KG, 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(1):W1-W73.