Performance of prediction models for COVID-19: the Caudine Forks of the external validation

Glen P. Martin[1], Matthew Sperrin[1] and Giovanni Sotgiu[2]

Affiliations: [1]Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. [2]Clinical Epidemiology and Medical Statistics Unit, Dept of Medical, Surgical and Experimental Sciences, University of Sassari, Sassari, Italy.

Correspondence: Giovanni Sotgiu, Clinical Epidemiology and Medical Statistics Unit, Dept of Medical, Surgical and Experimental Sciences, University of Sassari, via Padre Manzella, 4 Sassari 07100, Italy. E-mail: gsotgiu@uniss.it

@ERSpublications
Existing evidence suggests that none of the COVID-19 prediction models can be supported for clinical use. Here we discuss "what next" in COVID-19 prediction. https://bit.ly/2SMtoLV

Cite this article as: Martin GP, Sperrin M, Sotgiu G. Performance of prediction models for COVID-19: the Caudine Forks of the external validation. *Eur Respir J* 2020; 56: 2003728 [https://doi.org/10.1183/13993003.03728-2020].

Healthcare systems worldwide have observed significant changes to meet demands due to the coronavirus disease 2019 (COVID-19) pandemic. The uncertainty surrounding optimal treatment, the rapid public health urgency and clinical emergencies have caused a chaotic disruption of the cases and their related contacts at inpatient and outpatient settings. Developing more tailored healthcare plans based on the currently available scientific evidence, could help improve clinical efficacy, treatment outcomes, prognosis, and health efficiency.

Development and implementation of risk prediction models to aid risk stratification and resource allocation could improve the current scenario. Clinical prediction models (CPMs) aim to predict an individual's expected outcome value, or an individual's risk of an outcome being present (diagnostic) or happening in the future (prognostic), based on sets of identified predictor variables [1, 2]. A plethora of such models was described during the first wave of the COVID-19 epidemic: a recent "living" systematic review identified (at the time of writing) 145 CPMs focused on COVID-19 patients [3].

Unfortunately, many of the existing COVID-19 CPMs have been identified to be at high risk of bias due to poor reporting, over-estimation of predictive performance, and lack of external validation [3]. External validation, which is an important aspect during the development process of any CPM, can independently evaluate the model focusing on data independent to those data used to derive the model [1, 2]. Crucially, this step assesses the generalisability/transportability of the CPM into new populations before they are recommended for widespread clinical implementation.

To address this gap in the current literature, Gupta *et al.* [4], in this issue of the *European Respiratory Journal*, aimed to externally validate 22 of the CPMs identified in the above-mentioned systematic review [3]. Using data from 411 adults who were admitted to the University College Hospital (London, UK) with clinically diagnosed COVID-19, it showed that all CPMs performed poorly in the new data [4]. Moreover,

Received: 5 Oct 2020 | Accepted: 6 Oct 2020

the authors found that (within their data) baseline oxygen saturation on room air was the most predictive variable for in-hospital worsening, while age was the most predictive variable for in-hospital mortality [4]. Astonishingly, none of the 22 CPMs included in the external validation demonstrated significantly higher clinical utility when compared with these variables alone [4]. Together, these results imply that none of the selected CPMs should be recommended for adoption in daily clinical routine.

Some caution is required in the interpretation of the findings of Gupta *et al.* [4]. First, this is a single site validation. Performance of CPMs may vary widely from site to site [5], so it could be argued that some of the COVID-19 models were simply "unlucky" on this particular site, and may perform better elsewhere. Therefore, wider multi-site external validation is urgently needed, potentially combined with meta-analysis of their predictive performance [6]. Secondly, new CPMs for COVID-19 are being developed all the time, and it is apparent that the quality of those models is now beginning to rise, based on new insights and methodological criticisms raised during this first evolution of the pandemic. For example, the ISARIC 4C prediction model [7] shows promising predictive performance results and has been developed on one of the largest datasets for inpatient COVID-19 admissions to-date, although, there remain some methodological concerns surrounding this model [8]. Thus, it is imperative that further external validation studies are conducted, so that emerging models can be evaluated.

Indeed, despite the widespread interest in developing new CPMs, it is key that the research community takes stock of the available evidence, before adding more models to the mix. For example, the important study by Gupta *et al.* [4] focussed on validating CPMs aimed to predict in-hospital clinical worsening or mortality among COVID-19 patients. Given that there are numerous other outcomes and different clinical settings (*e.g.* outpatient settings) where COVID-19 CPMs have been developed, future external validation studies should be a target for research in the near future.

Nonetheless, the emerging scientific evidence indicates that none of the models can be recommended for clinical use and widespread adoption. Therefore, there remains one key outstanding question: how can we change this situation and develop CPMs for COVID-19 that can be recommended and largely implemented? One potential answer is that incentives need to change [9]. Specifically, having a sufficient and qualitatively representative sample size is a crucial assumption when any CPM is planned to be developed [10–13]; however, this methodological necessity could be challenging in the context of an emerging pandemic where high-quality data is often scarce. In this situation, data sharing becomes paramount, but can be easily hampered by current research incentives [9]. Inevitably, incentives meant that the clinical need for COVID-19 prediction models in different contexts acted as a starting signal for a race towards the "high-impact publication" finish line. This is a multifaceted issue, but it does raise implications if there are future new diseases where the development of CPMs is required, but limited data are available. Even so, there are still opportunities for data sharing in the context of COVID-19 CPMs. Gupta *et al.* [4] nicely articulate this in the context of validating existing COVID-19 CPMs as follows: "future studies may seek to pool data from multiple centres in order to robustly evaluate the performance of existing and newly emerging models across heterogeneous populations". If combined with meta-analytic methods [6, 14], such an approach will arguably be pivotal in addressing the clinical need for robust COVID-19 CPMs.

As the research community looks towards the "what now?" surrounding CPMs for COVID-19, it is important that we carefully utilise the available scientifically sound evidence where possible. Specifically, it is entirely conceivable that emerging models (such as the ISARIC 4C prediction model [7]) will show adequate predictive performance results in data similar to that in which the model was developed, but external validation shows poor transferability of the models to new demographics (*e.g.* new countries) and statistical populations. In this situation, the community should build upon such models, instead of developing *de novo* models in distinct populations. For example, such existing COVID-19 models could be updated and refined using data in other populations and, thereby, model transferability facilitated [15, 16]. A careful assessment and in-depth study of population characteristics can help find out the best approach to adapt a model to a new context, characterised by its own demographic, clinical, and epidemiological covariates. Repeatedly developing new models from scratch in distinct populations wastes prior information and risks overfitting. In contrast, model updating uses the existing models as a foundation, and builds upon this with the new data so that they are tailored to populations of interest.

Relatedly, there have been a multitude of CPMs developed for different aspects and contexts of COVID-19. Indeed, the existing models range from diagnostic CPMs aimed at predicting a virologically confirmed diagnosis of COVID-19, to prognostic CPMs oriented to the prediction of different clinical outcomes in those already diagnosed with COVID-19. However, many of these outcomes are inter-related and, in the context of facilitating decision-making for treatment/management of COVID-19 patients, looking at multiple outcomes simultaneously is often of greater clinical interest and offers more relevant

insights. For example, different models have been developed to predict clinical worsening, longer length of hospital stay or death (at various time-points), whereas others have defined composite outcomes (*e.g.* death and clinical worsening) [3]. For decision-making, clinicians could be more interested in the risk variability of several of those outcomes co-occurring (*e.g.* estimating a probability of clinical worsening and longer hospital stay). Here, joint (rather than marginal) prediction is the primary focus but developing separate models for each outcome individually cannot enable this type of prediction [17]. Future work might wish to consider this approach and type of objectives in the context of COVID-19 CPMs.

Finally, an important consideration for all COVID-19 models developed to-date is that they are derived in the context of current care. However, current care continues to change rapidly as the pandemic unfolds. This means that outputs from COVID-19 CPMs need to be interpreted carefully: when, how and where are question words to be posed to better interpret the potential inference and application of a model. Specifically, the predictions reveal the risk under the applied practices observed within the development dataset. They cannot be used to inform an individual's risk under various competing interventions [18]. A potential solution could be the exploration of counterfactual prediction [19, 20], in which the risk is estimated under fixed-care regimes. This separates the baseline risk and the actions taken to mitigate the risk, thereby enabling users to answer "what if" questions surrounding the impact of different interventions on COVID-19 risk in a particular setting [18]. Incorporating counterfactual prediction into the modelling might also increase the chances of model transferability across populations. Alternatively, embedding COVID-19 models within a dynamic framework [21] would allow the models to adapt rapidly to the changing clinical and temporal context. Such dynamic approaches to model updating help ensure that models maintain predictive performance through time, but it does require appropriate infrastructure to enable the real-time updates as new data are collected. Clearly, any form of model updating should only be undertaken on COVID-19 models that show lowest risk of bias, which are scarce to date [3].

To summarise, the current scientific evidence suggests that none of the existing COVID-19 CPMs can be recommended for clinical use. We urgently recommend additional external validation studies, such as those by Gupta *et al.* [4]. Future work should seek to pool data across different populations and to apply model updating methods, where appropriate, to facilitate refinement of models across population variability.

Conflict of interest: None declared.

References

1 Steyerberg EW. Clinical Prediction Models. New York, Springer, 2009.
2 Riley RD, Windt D, Croft P, *et al.* Prognosis Research in Healthcare: Concepts, Methods, and Impact. Oxford, Oxford University Press, 2019.
3 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.
4 Gupta RK, Marks M, Samuels THA, *et al.* Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. *Eur Respir J* 2020; 56: 2003498.
5 Riley RD, Ensor J, Snell KIE, *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* 2016; 353: 27–30.
6 Snell KIE, Hua H, Debray TPA, *et al.* Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2016; 69: 40–50.
7 Knight SR, Ho A, Pius R, *et al.* 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. *BMJ* 2020; 370: m3339.
8 Riley RD, Collins GS, van Smeden M, *et al.* Is the 4C Mortality Score fit for purpose? Some comments and concerns. *BMJ* 2020; 370: m3339. www.bmj.com/content/370/bmj.m3339/rr-3.
9 Sperrin M, Grant SW, Peek N. Prediction models for diagnosis and prognosis in Covid-19: all models are wrong but data sharing and better reporting could improve this. *BMJ* 2020; 369: m1464. www.bmj.com/content/369/bmj.m1464.
10 Riley RD, Snell KIE, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: part I - continuous outcomes. *Stat Med* 2019; 38: 1262–1275.
11 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.
12 Riley RD, Ensor J, Snell KIE, *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; 368: m441.
13 van Smeden M, Moons KG, de Groot JA, *et al.* Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res* 2019; 28: 2455–2474.
14 Debray TP, Damen JA, Riley RD, *et al.* A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019; 28: 2768–2786.
15 Janssen KJM, Moons KGM, Kalkman CJ, *et al.* Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008; 61: 76–86.
16 Su T-L, Jaki T, Hickey GL, *et al.* A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res* 2018; 27: 185–197.

17 Martin GP, Sperrin M, Snell KIE, *et al.* Clinical Prediction Models to Predict the Risk of Multiple Binary Outcomes: a comparison of approaches. *arXiv* 2020; preprint [https://arxiv.org/abs/2001.07624].
18 Sperrin M, Martin GP, Pate A, *et al.* Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Stat Med* 2018; 37: 4142–4154.
19 Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance* 2019; 32: 42–49.
20 van Geloven N, Swanson SA, Ramspek CL, *et al.* Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol* 2020; 35: 619–630.
21 Jenkins DA, Sperrin M, Martin GP, *et al.* Dynamic models to predict health outcomes: current status and methodological challenges. *Diagn Progn Res* 2018; 2: 23.