PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Does prognostic model development and internal validation design matter with big data?
AUTHORS	Reps, Jenna; Ryan, Patrick; Rijnbeek, P

VERSION 1 – REVIEW

REVIEWER	Liu, Vincent Kaiser Permanente Division of Research
REVIEW RETURNED	26-Apr-2021

GENERAL COMMENTS	The authors are attempting to address an important question in predictive model development specifically related to the identification of optimal methods for reducing bias and improving model transportability or generalizability. In general, this is an interesting contribution for the field of predictive model development and the methods they used were relevant. They test various iterations of test data and cross validation and compare a standard metric of performance (AUROC) and find that cross-validation with or without internal test set appeared to reduce 'bias' when the models were replicated externally. While this work is important and innovative, I do have several concerns. First, the overall manuscript is written from a very technical perspective and may be difficult to access by a more general, much less clinical, audience. Thus, I was somewhat surprised to see this manuscript in this journal versus another informatics or statistical methods oriented journal. If the editors decide that this is suitable for this journal's audience, I would suggest that the overall manuscript be written so that it is more accessible to an audience including those without significant familiarity with predictive model development. They may not grasp at all why this question is even important to begin with. Second, I suggest the authors improve their presentation of the concepts of the study. For example, rather than using nomenclature like 'test_0_cv_0', which in some ways is descriptive, they use more descriptive nomenclature that readers will find accessible whether they're looking at the tables/figures or reading the text. Third, I suggest the authors replicate these analyses but also choose other metrics of model performance. While many use the AUROC as the single most common metric, it does not adequately address performance, particularly when a dataset is imbalanced and an outcome is rare. I would suggest that the authors also show panels that demonstrate AUPRC (precision-recall curves) and calibration metrics. These are essential to the utility and

validity of predictive models, and they should be relatively easily analyzed with the existing dataset. If there is a need to simplify to fit in these additional analyses, I suggest the authors remove the 5-fold CV (since the 3 and 10 fold are quite similar already). Also, Figure 3 is not particularly informative with the results largely displayed already in Figure 2.

Fourth, it's not clear to me why there is such a stark difference in the performance when there is no cross validation (row 1 of Figure 1) in using no test (red) versus test (blue) set. Given that the 'test set' performance is simply a static 20% sample, is the discrepancy related to the fact that the train set simply overfits on a larger dataset? Isn't the test set sample defining a model on 80% of the data while the no test set sample defining a model on 100% of the data? This doesn't seem like it would cause as much difference as is seen?

Figure 1 needs a much clearer legend because it is very difficult to understand without reading the entire description. Please also include colors in the figure itself to identify what all fo the various displays are.

The Discussion should describe as a limitation that these are very different data samples: one being a commercially representative population under 65 and the other being a population over 65 from Medicare. If anything, I was surprised that the AUCs were as close as they were, given my expectation about their considerable differences.

REVIEWER	Yoon, Hyung-Jin Seoul Natl Univ, College of Medicine, Department of Biomedical Engineering
REVIEW RETURNED	11-May-2021

GENERAL COMMENTS

To determine the impact that the choice of internal validation design has on the internal discrimination estimate bias and model transportability in big data analysis, the authors assessed the AUROC of the LASSO logistic regression models with 8 different designs of internal validation. Although there are many designs of internal validation, which is a practical proxy of external validation for testing generalizability, the choice of the internal validation method is typically based on personal preference or experience of the researchers, rather than scientific background. Regarding this reality, the authors' manuscript could substantially contribute to the scientific community. Despite these merits, I have several concerns that need to be clarified before it gets published.

1. First of all, the authors chose a Medicare dataset for the external validation, of which age distribution did not overlap those of the datasets for the internal validation. Clinical manifestations of most diseases are very much age-dependent. Is there any explanation for choosing an external validation dataset of different age distribution?

2. The second point is an extension of the first point. According to Table 1, the rare outcomes frequently have event counts 2 to 3 times larger in the validation data which has a smaller sample size, than in the development data. Therefore, the lower generalizability of the model in the cases of rare outcomes might

	be due to the difference of the age distribution between the development data and the validation data, not due to the number of the outcome events. 3. The role of a test set in the internal validation is to test, not to improve, the generalizability. As the authors pointed out, a test set decreased the sample size for the performance of a developing model. Therefore, the decreased performance in the cases with a test set seems to be quite natural. 4. Twenty-one outcomes were predicted in this study. How were these outcomes chosen? Was it based on the frequency of events? 5. The authors need to explain why they randomly sampled 500,000 patients from the 1,964,494 treated for depression. 6. The number of available predictors was approximately 86,000. The authors should briefly explain how they treated these large numbers of predictors. 7. In line 35 & 36 of page 13, the authors described that it might be preferable to use a 3-fold CV when data are large to decrease the development time. Were the time consumptions substantially different between the numbers of the fold of CV? 8. The introduction is too long and some of the contents are described in duplicate. I think that the introduction needs to be refined more concisely to deliver the contents to the readers effectively.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. AbdulAzeez Anjorin, Lagos State University
Comments to the Author:
1. Good protocol paper. Please work on the suggestions as attached, including the references.

Authors' Response: Thank you for the positive feedback and the proposed edits. The changes requested, including to the references, have been implemented throughout the text.

2. Methods section: What informed the choice of each site, and 3 provinces out of 10 provinces and 3 territories?

Author's response: These sites were selected since they are 4 of the largest HIV clinics in Canada and have established research infrastructures to support the recruitment, enrolment and follow-up of a high volume of diverse study participants. These sites also have strong track records for rapid enrollment of participants in CTN studies. This text has been added to the Methods, Study design section (Page 10).

3. Sample size: The sample proportion here is better otherwise a standard formula like Kish Leslie sample formula or others may be used to better justify the sample population.

Author's response: We are aware that Kish published a book, "Survey Sampling" (not accessible to us), which discusses sample size calculation (referenced in another article by Israel in 1992). We are not clear about the reason why the reviewer has suggested this. However, we now reference that we used the UCSF sample size calculator (https://data.ucsf.edu/research/sample-size), which uses the typical normal distribution assumption with the continuity correction as an approximation to the binomial distribution. (Page 14).

4. References: Reference 5 was published in 2007. Please correct and verify other references.

Author's response: References 5 and 22-24 have been corrected, and the other references have been verified. Reference 55 has been removed since it is no longer relevant.

5. CITF CDE Baseline Questionnaire: Was the questionnaire pretested? It will interesting to include the Cronbach's alpha value as part of the statistical analyses?

To our knowledge, the CITF questionnaire has not been pretested in the general population or PLWH. Cronbach's alpha measures the internal consistency of responses. This is useful when one is asking questions which one believes all relate to a particular attribute, e.g., anxiety. In this instance, all of the questions should be correlated, and Cronbach's alpha would measure this.
Cronbach's alpha is also commonly used when there are multiple Likert questions in a survey/questionnaire that form a scale and one wishes to determine if the scale is reliable.
We can see applying Cronbach's alpha to, for example, questionnaire items 34 a-e, where all the questions purportedly measure the extent to which people adopted preventive measures (p.19).

6. What is the purpose of Question 6: "Are you an Indigenous person originating from North America?"

Author's response: It is important to determine the percentage of Indigenous persons participating in our study since they represent a vulnerable but often underrepresented group of PLWH. Due to systemic inequities and socio-economic marginalization, Indigenous peoples are at disproportionate risk of both HIV and COVID-19

Reviewer: 2
Dr. Claire Deleage, Frederick National Laboratory for Cancer Research
Comments to the Author:
In this manuscript, the authors are detailing the protocol they will follow to set up an observational cohort study to assess the immunogenicity outcomes in HIV positive participants following COVID-19 vaccination.
As the authors mentioned this cohort comes a little late looking at the current situation of the pandemic and its impact on immunosuppressed patients. However, this type of cohort is still essential to understand how HIV infected people built up specific immune response to SARS-CoV-2 post vaccination and if any additional immune boost would be needed to keep this population safe.
The main goals, general methodology, recruitment, Subpopulation of interest and timeline for participants are clearly presented and were specifically designed to answer key questions. The different parameters listed to assess the humoral and cellular immunity is clearly based on recent literature on SARS-CoV-2 and should be able to answer a lot of questions regarding the magnitude, specificity and durability of the immune response of the participants.

1. My only question is did the authors considered enrolling PLWH who didn't receive the vaccines but did get infected to compare the specificity, durability and type of immune response in patients who have been in direct contact replicating the virus compare to people who have been vaccinated.

Adding such a subgroup to this cohort can also provide key information regarding the impact of HIV and SASRS-CoV-2 infection.

Authors' Response: Although we did discuss the possibility of enrolling unvaccinated PLWH infected with SARS-CoV-2 in the cohort, the availability of such participants would be insufficient. From our clinical experience, the number of PLWH refusing vaccination is relatively low given the tremendous efforts from advertising campaigns and health care providers encouraging vaccine uptake in Canada. Therefore, our power would likely be too low to draw conclusions between persons with immunity via natural infection vs vaccination. Furthermore, in our experience, individuals refusing vaccination may be less likely to participate in a longitudinal research study due to medical mistrust, further reducing the power.

VERSION 2 – REVIEW

REVIEWER	Liu, Vincent Kaiser Permanente Division of Research
REVIEW RETURNED	14-Sep-2021

GENERAL COMMENTS	The authors have addressed my concerns.

REVIEWER	Yoon, Hyung-Jin Seoul Natl Univ, College of Medicine, Department of Biomedical Engineering
REVIEW RETURNED	02-Sep-2021

GENERAL COMMENTS	This study is considered to have sufficient scientific value as an empirical study on the impact of model development/validation design on the generalizability of the developed models using big data. All my concerns have been addressed.