Developing objective metrics to diagnose PatientLevelPrediction model designs

Jenna M. Reps and Yauheniya Cherkas

Background

Prognostic models aim to predict a patient's future health and may be able to guide medical decision making. Unfortunately, most prediction models fail to make any clinical impact. The main reasons that prognostic models do not get used are: 1) models are developed without a clear intended use, 2) models are developed using poor methodology and 3) models are often not shared and impossible to reproduce. Numerous reviews have shown how widespread the issue of poor methodology is in the field of healthcare prognostic models[1,2]. To address this issue, guidelines have been developed to improve prediction/prognostic model development, such as PROGRESS[3] and PROBAST[4]. PROBAST is used when reviewing publications that present new prognostic models and convers four main areas:

- Participants: investigates whether there are any issues in the design that may limit the generalizability of the model. That is, are the database and inclusions/exclusion criteria suitable for the desired target population?
- Predictors: investigates whether the predictors (aka covariates/independent variables) are suitable. That is, do they only include data that would be available when the model is intended to be used and are there any potential issues such as the outcome being used somehow as a predictor, or the predictors being measured differently across those with and without the outcome?
- Outcomes: investigates whether the outcome definition is correct and is a commonly used definition. Is the database suitable for the outcome (i.e., if you are predicting death does the database have good capture of death)? Was the outcome determined independently of the predictors? Is the outcome definition being applied consistently across participants?
- Design: investigates whether the analysis is suitable. Is there sufficient data to learn a model? Were any of the participants excluded? How are missing predictors addressed? Was an appropriate internal validation design used? Are standard evaluation metrics used?

These considerations help assess the likelihood of the model being problematic. However, some of the considerations may be subjective. The PatientLevelPrediction R package enables users to go end-to-end from observational data in the OMOP common data model to a fully validated prognostic models[5]. The PatientLevelPrediction package requires that users specify a model design. A model design is composed of seven components: 1) the target definition, 2) the outcome definition, 3) additional inclusion criteria, 4) the time-at-risk, 5) selected candidate covariates, 6) the test/validation/train design and 7) the supervised learning method. In this study we propose a new set of standard checks/metrics/plots that can be calculated for a PatientLevelPrediction model design across multiple databases prior to developing any models to ensure the design is PROBAST compliant.

Methods

For each of the PROBAST criteria we explain the check/metric/plot that is proposed.

Table 1- The checks/metrics/plots proposed for each PROBAST consideration. Include indicates whether our proposed diagnostic checks address this PROBAST consideration. Default true indicates whether this PROBAST consideration is automatically passed due to the PatientLevelPrediction framework. Proposed Check Rule provides the logical check that can be

applied to the model design settings. Proposed empirical metric/plot is a proposed output that the user can check for a given model design applied to a database to see whether the PROBAST criterium fails or passes.

1A - participants		Include	Default True	Proposed Check Rule	Proposed Empirical Metric/Plot
1.1	Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	Yes	Yes	By default as PatientLevelPrediction uses a cohort design	
1.2	Were all inclusions and exclusions of participants appropriate?	Yes			Include table showing participant size, outcome rate, min/median/mean/max age, male % when no participants are excluded vs when the design settings { *min prior observation, date restriction, min time-at-risk, first exposure only, prior observation* } exclude some participants. This table can be inspected to see whether the design settings impacted the participant age/gender/outcome distributions.
2A predictors					
2.1	Were predictors defined and assessed in a similar way for all participants?	Yes	Yes	FeatureExtraction R package used to create standard predictors and code is applied consistently across participants.	
2.2	Were predictor assessments (construction) made without knowledge of outcome data?	Yes		Check: *If predictor end date < time-at-risk start date then Pass*	
2.3	Are all predictors available at the time the model is intended to be used?	Yes		Check: *If predictor end date <=0 then Pass*	
3A outcome					
3.1	Was the outcome determined appropriately?	Yes			Include outcome rate plots across age/index month/gender/year and check whether the observed rates match known rates consistently across datasets.
3.2	Was a pre-specified or standard outcome definition used?	Yes		Check: *If the outcome cohort is from the OHDSI phenotype library the Pass*	

3.3	Were predictors excluded from the outcome definition?	Yes			Include Kaplan Meier plot for the outcome to inspect whether the outcome occurs close to index or further away. Close to include may indicate issue with predictors and outcome occurring close in time. After model development inspect predictors to ensure no predictors semantically the same as the outcome.
3.4	Was the outcome defined and determined in a similar way for all participants?	Yes	Yes	By default as we use the same outcome definition for all participants	
3.5	Was the outcome determined without knowledge of predictor information?	Yes	Yes	By default as we create the outcome definition independently of the prediction design and ensure the time period to create predictors is before the time period to identify the outcome.	
3.6	Was the time interval between predictor assessment and outcome determination appropriate?	Yes		Check: *If predictor end date < time-at-risk start date then Pass*	
4A design					
4.1	Were there a reasonable number of participants with the outcome?	Yes		Check: *If the number of outcomes < 200 then Fail else Unknown*	Include table containing the number of outcomes – this can be viewed to see whether the number of outcomes appears to be sufficient.
4.2	Were continuous and categorical predictors handled appropriately?	Yes	Yes	By default as the recommended predictors are binary indicators	
4.3	Were all enrolled participants included in the analysis?	NA		PatientLevelPrediction uses all patients in the target cohort unless this is very large and then a random sample may be taken.	
4.4	Were participants with missing data handled appropriately?	No		The user is required to ensure any custom features with the potential to have missing values are handled.	
4.5	Was selection of predictors based on univariable analysis avoided?	Yes	Yes	By default as this is not an option in the standard PatientLevelPrediction framework	
4.6	Were complexities in the data (e.g. censoring, competing risks,	NA		This requires the user to pick a suitable prediction task.	

	sampling of controls) accounted for appropriately?				
4.7	Were relevant model performance measures evaluated appropriately?	Yes	Yes	By default as the PatientLevelPrediction framework contains the TRIPOD recommended metrics	
4.8	Were model overfitting and optimism in model performance accounted for?	NA		This must be assessed after model development	
4.9	Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?	NA		This must be assessed after model development	

To include the proposed prediction diagnostic we added the diagnosePlp() function into the PatientLevelPrediction R package. In addition, the interactive PatientLevelPrediction shiny app has been updated to display the diagnostic results.

We demonstrate the new PatientLevelPrediction diagnostics for the prediction task: In patients with new atrial fibrillation, predict the risk of stroke within 1 to 365 days. Patients were included if they had >365 days prior observation and no history of prior stroke. We used a LASSO logistic regression with a 75% train and 25% test split and 3-fold cross validation on the train data to pick the optimal hyper-parameter. The predictors used were demographics (age and gender) plus all drugs/conditions recorded within 1-year prior to index (not including index).

We ran the diagnostics for the model design on the CCAE database. This is an insurance claims database for people who are commercially insured and their dependents. Patients are generally less that 65 years old.

Results

The shiny app has the option to view the diagnostics for a model design, see Figure 1. After clicking on the 'View Diagnostics' a summary table will present the diagnostic results, see Figure 2. In Figure 2 we see that the prediction task investigated passes the PROBAST checks 1.1, 1.2, 2.1, 2.2, 3,4 and 3,6 (detailed

in Table 1) but does not pass 2.3 and 4.1 which were classed as 'Unknown' and require the user to explore the additional tables/plots.

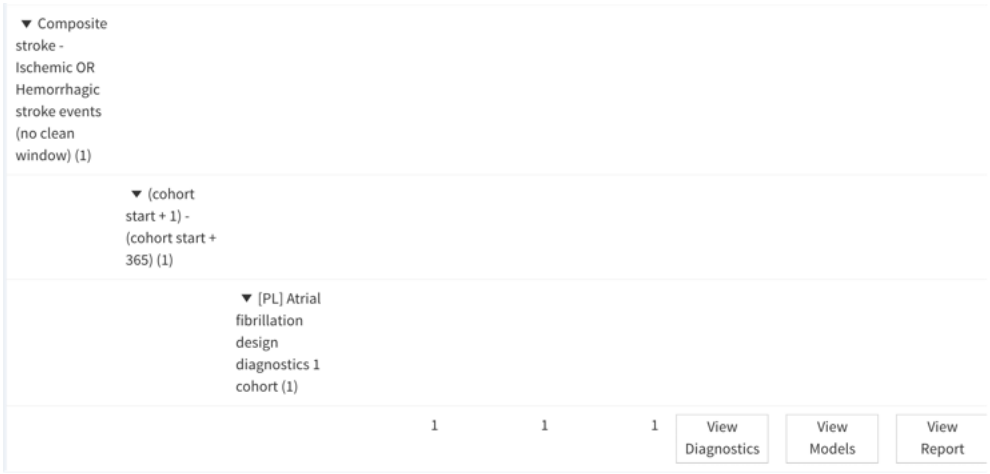

Figure 1 - The shiny viewer for exploring model designs. The user can view the diagnostic results for the model design by clicking on the 'View Diagnostics' button.

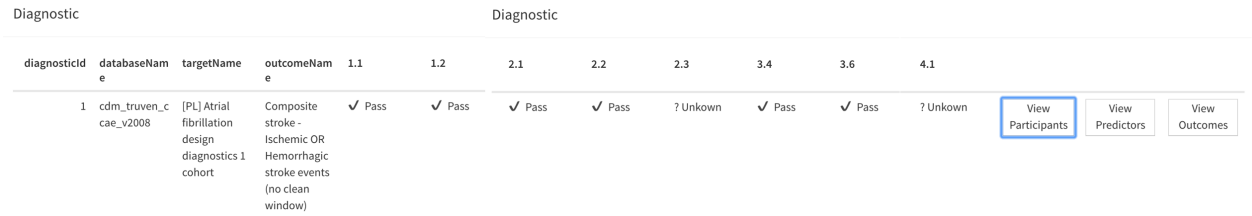

Figure 2 - Diagnostic summary in shiny app. After selecting the model design the PROBAST diagnostic results for each databases investigated will be summarized with a pass/fail/unknown. The user can also use the buttons on the right to select different tables/plots for empirical assessment of the PROBAST criteria deemed unknown.

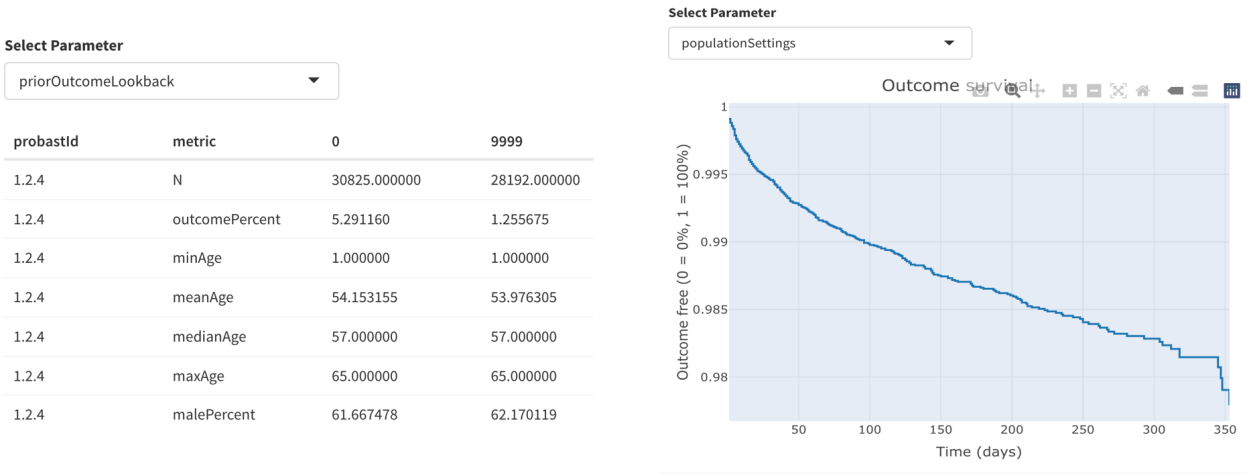

Figure 3 - The interactive tables/plots to empirically explore potential issues. The table on the left shows the number of participants, the outcome rate, the age summary of the participants and the male % when including all patients with atrial fibrillation (column 0) vs just the atrial fibrillation patients who did not have stroke within 9999 days prior to initial atrial fibrillation. The plot on the right shows the fraction of the participants outcome free across the 365-day time-at-risk.

The table on the left of Figure 3 shows excluding patients who had stroke prior to index remove ~2000 participants. Due to excluding patients with prior stroke, the rate of stroke decreased as expected. Excluding patient with prior stroke did not appear to impact the age and gender distributions of the included participants. This indicates PROBAST 1.2 passes as excluding participants with prior stroke did not appear to change the demographics. In our example we only had an exclusion based on having the outcome prior, but if other exclusions are included in a design, users can explore the impact of these to see whether the age/gender distributions are impacted. It can be seen that ~280 included participants had the outcome, this is why PROBAST 4.1 did not fail. However, it is up to the user to decide whether 280 outcomes are sufficient. The right plot in Figure 3 shows that the time to outcome is spread out over the 365-day follow-up. This is used to guide PROBAST 3.3. If there was a large drop around day 0, this could indicate a design issue as the outcome is occurring closely in time to the index and the predictors are often created using records up to index.

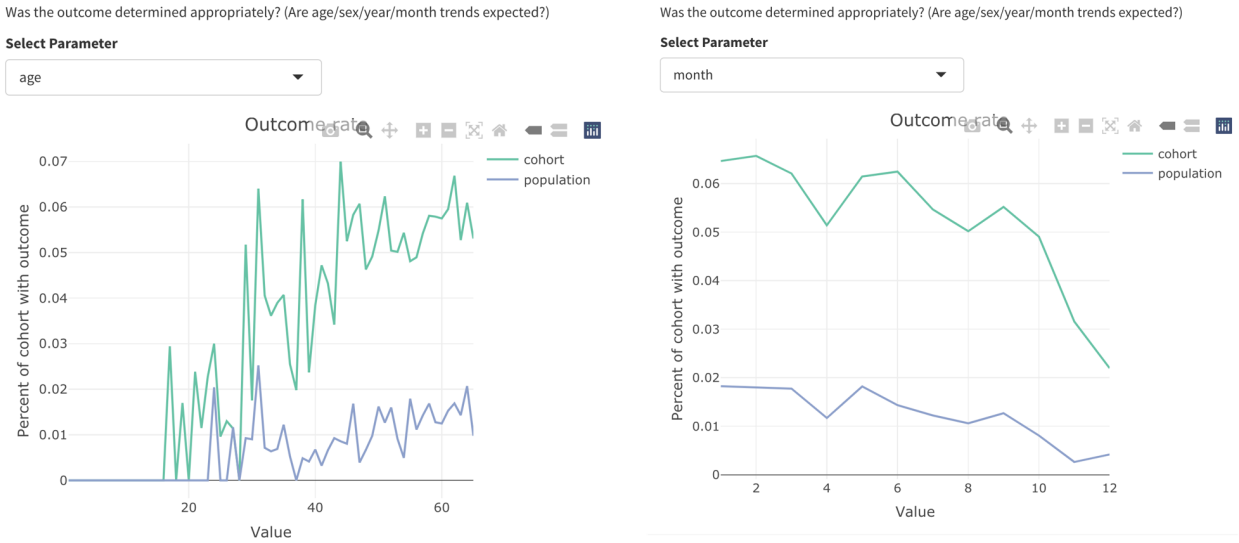

Figure 4 - The outcome rate by age and index month for all patients with atrial fibrillation (cohort) and only those who never had a prior stroke (population). These plots can help identify what impact the inclusions criteria has.

Figure 4 shows the interactive plots that let you explore the outcome rate between the full cohort of patients and those who satisfy the inclusion criteria. This is to check PROBAST 3.1. The purpose of this plot is to guide whether the outcome appears to be appropriately defined (i.e., is the outcome rate trend by age and index month as expected) and what impact the inclusion criteria has on the outcome rate (is the difference between the full cohort and subset satisfying the inclusions criteria expected).

Conclusion

We show that is it possible to develop some objective checks for the PatientLevelPrediction model design to inspect whether the design may be biased. In addition, we proposed three different tables/plots that may highlight design issues. When demonstrated for the task of predicting stroke in patients newly diagnosed with atrial fibrillation, we showed that in general our design looked feasible for the characteristics evaluated with potentially insufficient number of outcomes. We recommend that users of the PatientLevelPrediction generate and explore the new diagnostics to ensure their model design is issue free. In addition, users can automatically generate a report with the design and diagnostic results using the shiny app (summary table and all the tables/plots). For transparency, we recommend that this report be submitted as a supplement when any model developed using the PatientLevelPrediction is submitted as a publication to a journal. This will enable reviewers to explore any model design issues.

References/Citations

1. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, Hooft L, Kirtley S, Riley RD, Van Calster B, Moons KG. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC medical research methodology. 2022 Dec;22(1):1-6.
2. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MM, Dahly DL, Damen

JA, Debray TP, De Jong VM. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj. 2020 Apr 7;369.

3. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG, PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS medicine. 2013 Feb 5;10(2):e1001381.
4. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine. 2019 Jan 1;170(1):51-8.
5. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association. 2018 Aug;25(8):969-75.