How much data do we need for clinical prediction?

L.H. John[1],MSc; P.R. Rijnbeek[1], PhD
[1]Erasmus University Medical Center, Rotterdam, The Netherlands

Background

Historically, clinical prediction models have been developed with limited sample size and relatively few predictor variables [1]. In recent years, however, large observational health databases have opened up possibilities to develop clinical prediction models at a much larger scale, for a host of target populations and health outcomes and using large sample sizes [2].

To take advantage of the predictive power of such data sources the Observational Health Data Sciences and Informatics (OHDSI) initiative has designed and implemented a prediction framework for observational health data standardized to a common data model [3]. This framework can be used to develop and validate clinical prediction models across disparate observational health databases, for example from administrative claims and electronic health records, effectively enabling the development of models at a global scale.

Objective

It is unknown whether these large amounts of observational health data help to improve model performance or unnecessarily put constraints on computing resources and computation time. Constraining large datasets based on an optimal sample size may facilitate the development of clinical prediction models at scale.

In this retrospective study we investigate sample size requirements for developing clinical prediction models using routinely collected observational health data. We empirically assess the effect of sample size on discrimination performance (AUROC) by generating learning curves for a diverse set of clinical prediction problems. We also compare our results to the popular 10 Events Per Variable (EPV) rule, which can be used to estimate an optimal sample size.

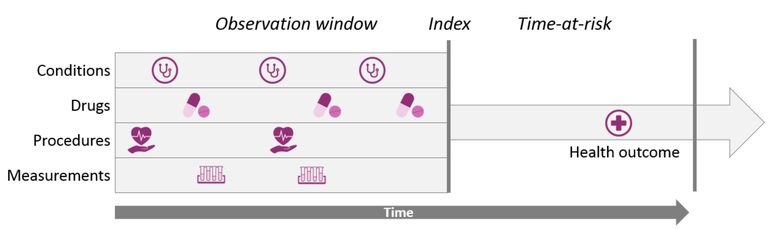

Figure 1. Patient-level prediction as defined by OHDSI.

Methods

We define *patient-level prediction* as a modeling process wherein a *health outcome* is predicted within a *time-at-risk* period relative to a *target cohort* start date (**Figure 1**). Prediction is performed using a set of predictors derived from patient data in an observation window prior to the start of the target cohort. The prediction algorithm of choice is logistic regression with L2 regularization (LASSO), which also performs automatic variable selection.

We empirically assess the effect of sample size on prediction performance by generating learning curves for 81 prediction problems of which 23 originated from the patient-level prediction framework study in patients with depression. The remaining 58 prediction problems were defined in patients with hypertension as part of the LEGEND study.

The analysis is performed on multiple data sources.

- IBM® MarketScan® Commercial Database (CCAE)
- IBM® MarketScan® Multi-State Medicaid Database (MDCD
- IBM® MarketScan® Medicare Supplemental Database (MDCR)

Results

Learning curves show the effect of sample size on discrimination performance (AUROC). We observe diminishing performance returns as we add more data to a model. We also observe that in our highly imbalanced data sets the learning is strongly correlated to the number of outcome events (positive cases) rather than the absolute sample size. **Figure 2** shows the learning curves plateauing with 5,000 to 20,000 outcome events. Interestingly, the effect is the same across all three databases

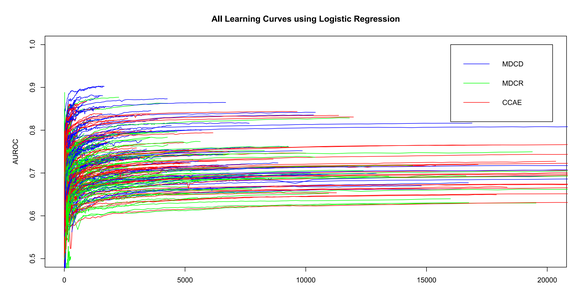

Figure 2. Learning curves generated using Logistic Regression and color coded by database.

Discussion

To make recommendations on the optimal sample size we can empirically estimate the required number of outcome events as a function of the expected performance improvement ΔAUROC (**Figure 3**).

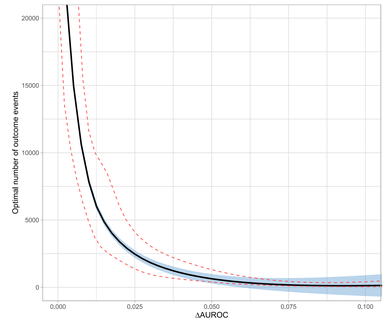

Figure 3. Optimal sample size as a function of ΔAUROC.

We observed that the automatic variable selection selects more variables as sample size increases. This results in unnecessarily growing model complexity, as discrimination performance can plateau with several thousand outcome events. A model with fewer variables can be transported, applied, and "interpreted" more easily, which outlines the relevance of estimating an optimal number of outcome events using the curve in **Figure 3**.

Conclusions

We do not need millions of patient records for clinical prediction. An optimal sample size should be defined in terms of the number of outcome events, which agrees with commonly used formulas such as the 10 EPV rule. For iterating over design choices, e.g. finding optimal hyperparameters, one may refer to **Figure 3** and choose to use a larger ΔAUROC, while for model deployment a smaller ΔAUROC is recommended.

[1] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review," *J Am Med Inform Assoc,* May 17 2016.
[2] I. W. Health, "IBM MarketScan Research Databases for Health Services Researchers (White Paper)," 2018.
[3] J. M. Reps, M. J. Schuemie, M. A. Suchard, P. B. Ryan, and P. R. Rijnbeek, "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data," *J Am Med Inform Assoc,* vol. 25, pp. 969-975, Aug 1 2018.

Contact: contact@ohdsi.org