Best Practices for Patient-Level Prediction in OHDSI

Peter R. Rijnbeek, PhD[1], Patrick Ryan, PhD[2], Hamed Abedtash, PharmD[3], David Dorr, MD, MS[4], George Hripcsak, MD, MS[5], Mandev S. Gill, PhD[6], Kenney Ng, PhD[7], Narges Razavian, PhD[8], David Sontag, PhD[8], James Weaver, MPH, MS[2], Andrew E. Williams, PhD[9], Johan van der Lei, PhD[1], Martijn Schuemie, PhD[2], Jenna Reps, PhD[2]

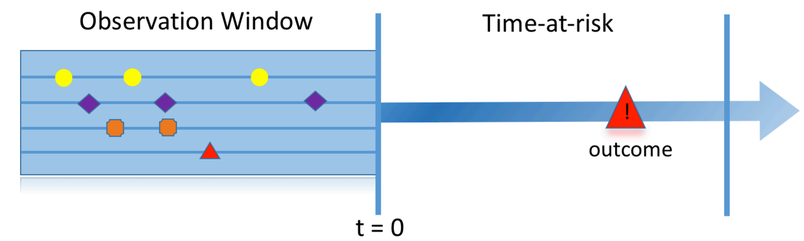
[1]Erasmus MC, Rotterdam, The Netherlands; [2]Janssen Research and Development, Raritan, NJ [3]Indiana University School of Informatics and Computing, Indianapolis, IN; [4]Oregon Health & Science University, Portland, OR; [5]Columbia University Medical Center, New York, NY; [6]Columbia University, New York, NY; [7]IBM Research, Cambridge MA, USA; [8]New York University, New York, NY; [9]Maine Medical Center Research Institute, Portland, Maine

Observational Health Data Sciences and Informatics (OHDSI) holds the promise of making large-scale, patient-specific predictive modeling a reality. The OHDSI network currently contains longitudinal data on over 600 million patients observed for multiple years and comprising over 5 billion clinical observations. The data is stored in a common data model (CDM), enabling uniform and transparent analysis. These large standardized populations contain rich data to build highly predictive large-scale models and also provide immediate opportunity to serve large communities of patients who are in most need of improved quality of care.

Effective exploitation of these massive dataset demands novel methodology and an interdisciplinary approach. The focus of the Patient-Level Prediction workgroup is to build a standardized, fully transparent workflow on top of the OMOP CDM. One of the first steps of the group was to establish best practices for patient-level predictive modelling. This work describes the consensus of the team which will provide a solid foundation for our future research and tool development in this promising field.

Problem definition

Among a population at risk, we aim to predict which patients at a defined moment in time (t=0) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to that moment in time.



General Principles

Transparency. Others should be able to reproduce a study in every detail using the provided information. All analysis code should be made available as open source on the OHDSI Github.

Problem pre-specification. A study protocol should unambiguously pre-specify the planned analyses.

Code validation: Unit tests, code review, or double coding steps are required to validate the developed code base. It is recommended to test the code on benchmark datasets.

Best Practices

Data characterization and cleaning is required before modelling. Tools are being developed in the community to facilitate this. A data cleaning step is recommended, e.g. to remove outliers in lab values.

Handling of missing values should be declared. The workgroup believes handling of missing data in patient-level prediction is an interesting area of future research.

Feature construction and selection should be completely transparent using a standardized approach to enable replication and to enable application of the model on unseen data.

Inclusion criteria should be made explicit. It is recommended to do sensitivity analyses on the choices made. Visualization tools could help and this will be further explored.

Model development is done using a split-sample approach. The percentage used for training could depend on the number of cases, but as a rule of thumb 80/20 split is recommended. Hyper-parameter training should only be done on the training set possibly using cross-validation methodologies. Model development should be an empirical process in which multiple models are evaluated.

Internal validation should only be done once on the test set. The following minimum set of performance measures are required:

- Overall performance: Brier score (unscaled/scaled)
- Discrimination: Area under the ROC curve (AUC)
- Calibration: Intercept + Gradient of the line fit on the observed vs predicted probabilities.

External validation should be done for all studies. Several scenarios are being explored to enable central sharing of models. The goal is to support replication of results, re-calibration of models, but also sharing final models with a wider community outside OHDSI

Dissemination of results should follow the minimum requirements as stated in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (Moons, KG et al. Ann Intern Med. 2015;162(1):W1-73).

Conclusions

Taking into account the team's combination of backgrounds, the unfettered access to a unique data resource, and the substantial collaborative track record of OHDSI, we believe the patient-level prediction workgroup is well positioned to advance the field.

The first set of best practices for patient-level predictive modelling in the OHDSI context have been established and will form the basis of the challenging but extremely interesting road ahead.

Contact:p.rijnbeek@erasmusmc.nl