[Note: All submissions must be in PDF format. Failure adhere to the format requirements may result in rejection of your submission without review]

Name:	Jenna Reps, PhD
Affiliation:	Janssen Research and Development
Email:	jreps@ITS.JNJ.com
Presentation type (select one):	Collaborator Demonstration

Patient-Level Prediction Package Demo

Jenna M. Reps, PhD[1], Martijn J. Schuemie, PhD[1], Patrick B. Ryan, PhD[1], Peter R. Rijnbeek, PhD[2]
[1]Janssen Research and Development, Raritan, NJ; [2]Erasmus MC, Rotterdam, The Netherlands

Abstract

Many clinical prediction models have been developed in the past to support decision making in a wide spectrum of specialties. These models predict a diagnostic or prognostic outcome based on a combination of patient characteristics, e.g. demographic information, disease history, treatment history. The number of publications describing clinical prediction models has increased strongly over the last 10 years. Surprisingly, most currently used models are estimated using small datasets and contain a limited set of patient characteristics. Due to the increasing availability of massive sets of observational health data such as claims or patient records large-scale analytics is now becoming a reality. OHDSI has established an international network of researchers and databases that standardized their data to a common data model and standard vocabularies which enables the application of standardized tools for patient-level predictive modelling. In the software demonstration the first version of the PatientLevelPrediction R library is applied to the problem of predicting the occurrence of myocardial infarction within 1 to 366 days after the first exposure to celecoxib. The library can support local modeling activities but also facilitates network based research that protects patient privacy by only sharing models and evaluation metrics.

Introduction

Effective exploitation of the massive sets of health data demands novel methodology and an interdisciplinary approach. Firstly, the longitudinal data is by nature sparse and irregular-spaced. Secondly, dimensionality reduction methods need to be assessed and developed to deal with the massive amount of data. Thirdly, leveraging the temporal information in the EHR records could possibly improve the performance of prediction models. Fourthly, purely data driven approaches run the risk of resulting in incomprehensible and suboptimal models by completely ignoring the already available background knowledge of the outcome and its etiology. Leveraging background information from literature and product labels is clearly a promising field of research in this context. Finally, proper assessment of the internal and external validity of the estimated models is challenging. The curse of dimensionality or overfitting is a well-known problem which arises from model uncertainty and parameter uncertainty. Additionally, it is important to be able to distinguish noise from true heterogeneity when testing the models externally. The OHDSI data network is very well suited to study the transportability of the models in a wide range of settings.

The focus of the Patient-Level Prediction workgroup is to research and develop solutions for patient-level prediction modelling in massive sets of real-world EHR data, considering the challenges described above. All patient-level predictive model tools will be developed as open-source solutions built against the OMOP common data model.

Patient-Level Prediction tools

In the software demonstration we will show how to develop and compare models to predict the risk of developing of myocardial infarction within 1 to 366 days after first being exposure to celecoxib (see Figure 1) using the *PatientLevelPrediction* package:

https://github.com/OHDSI/PatientLevelPrediction

In short, the user first needs to specify two cohorts and the risk time period:

1. At risk cohort: A cohort of people sharing some baseline health characteristics (e.g., first time exposure to celecoxib). The characteristics should describe the group of people you want to apply the model to in the future.

2. Outcome cohort: A cohort of people who have the outcome (e.g., myocardial infarction) for which we wish to build a predictive model.

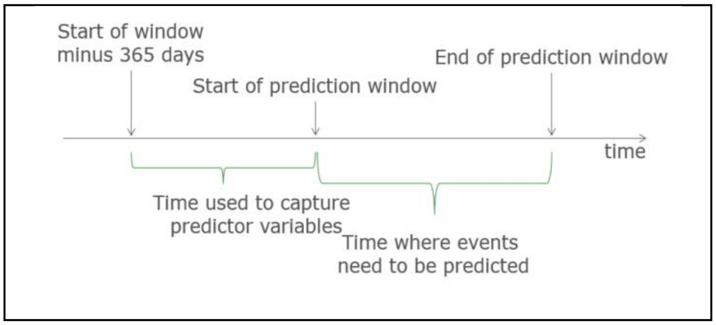
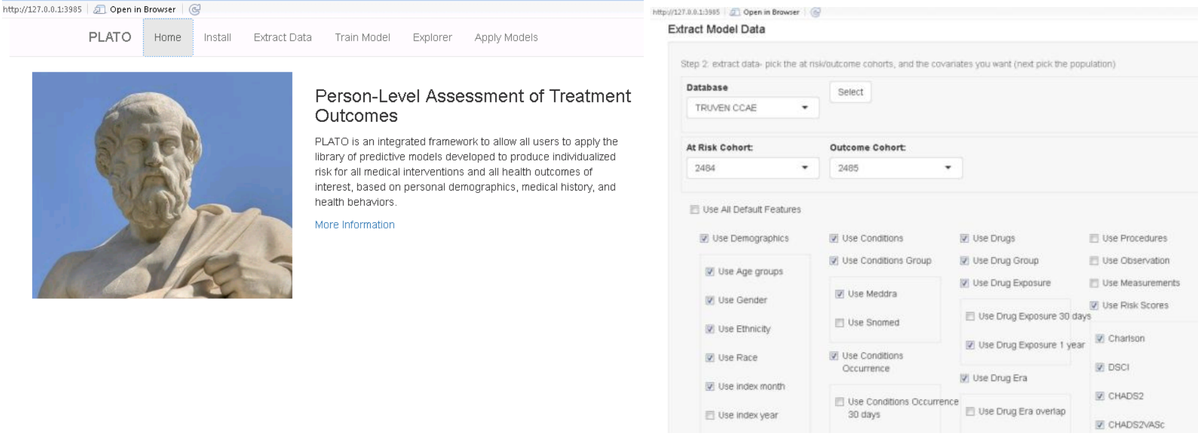


Figure 1. Patient-Level Prediction problem definition.

Using the *FeatureExtraction* package (https://github.com/OHDSI/FeatureExtraction) data is then extracted from the time period preceding (and including) the cohort start date. This large set of features will be used to predict outcomes during the prediction window. These features can include binary indicators for the occurrence of any individual drug, condition, procedures, as well as demographics and comorbidity indices. Functionality is available to create custom features if needed.

The *PatientLevelPrediction* package creates a train-test split of the data by selecting patients randomly or by comparing two different time periods. During model training it applies n-fold cross validation to limit overfitting. It has functionality to run a large set of algorithms to estimate the best models using the training data. These current models include generalized linear models with regularization[1], tree based models such as random forests[2] and gradient boosting machines[3], k-nearest neighbors, naïve Bayes and we are currently adding in neural networks and deep learning[4]. The overall performance, the discrimination, and the calibration of the estimated models is assessed on the test set using patient or time splits. Graphical output is generated, e.g. ROC curves, to facilitate easy model comparison and selection. In addition to the internal validation, the package incorporates a function for applying the trained model on new data, which can be used to readily perform external validation on data in the common data model format across the OHDSI network. An interactive application (Shiny) has been developed as shown in Figure 2.



Person-Level Assessment of Treatment Outcomes

PLATO is an integrated framework to allow all users to apply the library of predictive models developed to produce individualized risk for all medical interventions and all health outcomes of interest, based on personal demographics, medical history, and health behaviors.

Extract Model Data

Step 2: extract data- pick the at risk/outcome cohorts, and the covariates you want (next pick the population)

Database: TRUVEN CCAE

At Risk Cohort: 2484

Outcome Cohort: 2485

☐ Use All Default Features

☑ Use Demographics
☑ Use Age groups
☑ Use Gender
☑ Use Ethnicity
☑ Use Race
☑ Use index month
☐ Use index year

☑ Use Conditions
☑ Use Conditions Group
☑ Use Meddra
☐ Use Snomed
☑ Use Conditions Occurrence
☐ Use Conditions Occurrence 30 days

☑ Use Drugs
☑ Use Drug Group
☑ Use Drug Exposure
☐ Use Drug Exposure 30 days
☑ Use Drug Exposure 1 year
☑ Use Drug Era
☐ Use Drug Era overlap

☐ Use Procedures
☐ Use Observation
☐ Use Measurements
☑ Use Risk Scores
☑ Charlson
☑ DSCI
☑ CHADS2
☑ CHADS2VASc

Figure 2. Interactive patient-level prediction application

OHDSI Network study

We are currently implementing a study to investigate the performance and transportability of models over the OHDSI network and welcome any interested participants to join the study: http://www.ohdsi.org/web/wiki/doku.php?id=research:celecoxib_prediction_models.

Conclusion

A first version of an R package for patient-level prediction has been developed that facilitates model creation and evaluation on massive EHR datasets. In the near future, the team will further expand the toolset with other algorithms, feature selection and engineering methods. The final goal is to develop a web-interface, called Person-Level Assessment of Treatment Outcomes (PLATO), which will be fully integrated in the OHDSI infrastructure. By using this standardized and transparent approach we strongly believe we can advance the field of clinical predictive modelling.

References

1. Tibshirani, Robert. Regression shrinkage and selection via the lasso.*Journal of the Royal Statistical Society. Series B (Methodological)* 1996; 267-288.
2. Breiman, Leo. Random forests. *Machine learning* 2001;45.1: 5-32.
3. Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 2001; 1189-1232.
4. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature 20015;*521.7553: 436-444.