| | |
|---|---|
| Name: | Melissa Rost |
| Affiliation: | Georgia Tech Research Institute |
| Email: | melissa.rost@gtri.gatech.edu |
| Presentation type (s): | **Poster, Software Demonstration** |

# Implementing Real-Time Patient Level Predictions Using PLP Models

**C. Scott Brown, Jon D. Duke, MD, James P. Fairbanks, Ph.D., Christine Herlihy, Kausar Mukadam, Jason A. Poovey, Melissa D. Rost,**
**Georgia Tech Research Institute, Atlanta, GA, USA**

**Abstract**

*We built a general framework over the OHDSI patient-level prediction package to publish an API for a predictive analytics model for C. Diff, a deadly bacterial infection. This model makes risk predictions for individual patients based within a precision medicine decision support tool with the ability to toggle off and on the inclusion of potential drugs or treatments. This tool can aid medical practitioners when prescribing drugs with potential complications. With this ability, consideration needs to be given to the quality of the model itself as that effects the quality of the predictions and the inability of the model to predict for new treatments not available in the training data.*

**Introduction**

The OHDSI suite of tools provides a very rich platform for health data standardization, cohort building, advanced data analysis, and more. As an initial test case leveraging these capabilities, we used the OHDSI Patient Level Prediction (PLP) package to extract patient-level data from the MIMIC critical care dataset comprising approximately 40,000 critical care patients[1]. This de-identified data set includes but is not limited to demographic information, vital signs, laboratory tests, and administered medications.

From the MIMIC data, we were specifically interested in building a predictive model for Clostridium difficile, commonly known as C. Diff. This life-threatening infection in the colon is caused by taking antibiotics that kill the good gut bacteria so that bad bacteria can grow unchecked. In addition to the problems for the patient, costs attributable to C. Diff in the US are estimated at $6.3 billion annually and nearly 2.4 million days of inpatient stay annually so understanding the causes and factors involved more thoroughly has the potential to save lives and reduce a costly, resource-intensive problem[2].

The general PLP pipeline involves creating a cohort of patients in the database, extracting features of interest present in individual patients, and analyzing the resulting dataset by building a predictive model. The OHDSI PLP package to perform these tasks is written in R; however, the third step is implemented by having R call specific models in scikit-learn, a state of the art machine learning library written in Python. In order to facilitate access to the entire scikit-learn library, we use scikit-learn directly for the third step of analyzing the resulting data.

**Integration with the Patient Level Prediction Package**

To build the cohorts with the PLP package, SQL queries are dynamically constructed and executed to write to the cohort tables in the database. Our implementation incorporates the queries directly for creating the index and outcome cohorts as well as information for connecting to the database. Since cohort creation is isolated from downstream analysis, OHDSI ATLAS can also be used to create cohorts and the unique cohort ID supplied to the downstream analysis.

With the cohorts created, our implementation used the features defined in the PLP package to extract patient-level information. In order to support both bulk queries for training and individual patient queries for predicting, the features were extracted using SQL queries generated for both purposes. After extracting these features, the resulting data from the cohorts, population, outcomes, and covariates are used in the analytics step.

**Predicting Outcome Risk for a Single Patient Using a Trained Model**

In order to predict C. Diff risk, we built a predictive model using the scikit-learn library. A user interface (UI) shown

in Figure 1 allows researchers to choose any combination of drugs, conditions, procedures, and demographics with which to build either a balanced random forest or logistic regression model. After many iterations on the model type and covariates, the UI was limited to only these most promising subsets for C. Diff.

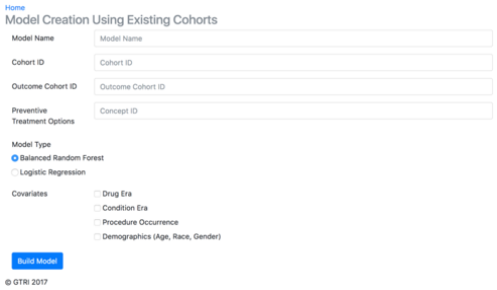


**Figure 1.** User interface for creating a predictive model on cohorts already written to the database.

We extended beyond the PLP pipeline of training a model and evaluating its accuracy by building a real-time system to use the trained model to predict the risk of C. Diff for specific patients using their vector of covariates. This capability is deployed in the Advanced Clinical Decision Support (ACDS) system (Figure 2) to provide precision medicine to individual patients. For potential drugs to prescribe, we can use the model to predict an individualized patient outcome in the presence or absence of a certain drug to aid a physician in clinical decision-making.

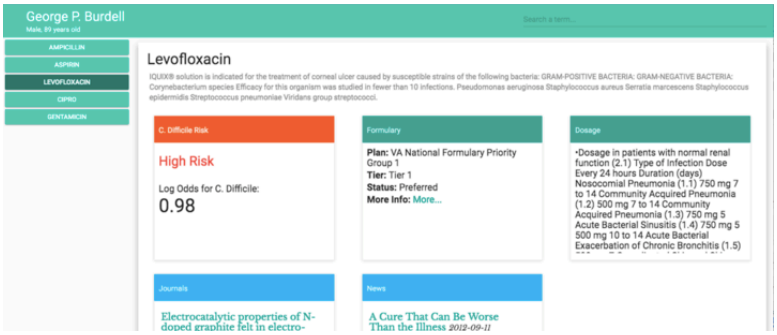


**Figure 2.** Advanced clinical decision support (ACDS) tool showing an individualized patient prediction of c. diff risk in the presence of levofloxacin.

## Conclusion

Utilizing the OHDSI PLP package, we were able to extend the predictive model built from a cohort of patients in order to provide a real-time prediction for new patients with the same selection criteria for the cohort. Such a tool allows health practitioners the ability to gain knowledge of how an individual patient could be expected to react to a new drug or treatment before having to go through prescribing the drug and waiting for it to run its course.

This kind of advancement has the potential to revolutionize how medicine is prescribed; however, the medical community needs to be aware of potential flaws. The predictions are only as good as the model used to create them and creating the model is the real 'science' of the data science. Additionally, a predictive model only works with data that it was trained on, so predicting for new covariates that weren't previously seen yields invalid results. Precision medicine going forward will require collaboration between data scientists and medical professionals to understand the potential benefits and flaws.

## References


1. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific Data (2016). *DOI:* 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635
2. Zhang S, Palazuelos-Munoz S, Balsells EM, Nair H, Chit A, Kyaw MH. Cost of hospital management of Clostridium difficile infection in United States—a meta-analysis and modelling study. BMC Infectious Diseases. 2016;16(1):447. doi:10.1186/s12879-016-1786-6.