PP23

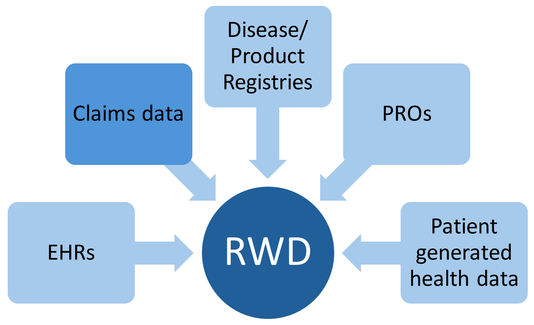
From Data Iron to Data Gold

Ashwini Yermal Shanbhogue

Abstract

In recent years, Real World Data (RWD) has gained importance in the clinical trial space. One of the sources of RWD is claims data. Claims data is usually present in large datasets, which until recently were hard to process and gain any meaningful insights from. However, recent advances in Machine Learning (ML) have made this task easier. In this poster, I will demonstrate how an R package that includes several ML algorithms can be used to make predictions and thus gain insights from a claims data set, the CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF).

Introduction



Claims data consists of billing codes that a patient's healthcare provider submits to insurance providers so that the healthcare provider may get reimbursed for services rendered.

Advantages of claims data:

1. Its large sample size, which can be used to gain insights into rare diseases.
2. Its diversity.
3. Ability to follow a patient's journey across multiple healthcare providers as long as the insurance provider remains constant.

Disadvantages of claims data:

1. Miscoding may occur due to human error.
2. Large size makes it harder to analyze and gain insights from.

However, due to recent advances in Machine Learning (ML), analyzing large data sets and gaining insights from them is becoming easier to accomplish.

'PatientLevelPrediction' is an R package containing several ML algorithms developed by the Observational Health Data Sciences and Informatics (OHDSI) program which can be used to build and validate patient-level predictive models using data in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Since CDMs have standardized structure and vocabulary, using this standardized package to analyze data present within a CDM will result in models and results that are standardized and easy to share or compare across studies and data sets.

CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) is a realistic set of synthetic claims data released by the Centers for Medicare and Medicaid Services (CMS). It has been converted into OMOP CDM by the open-source community.

I will use this data set along with the 'PatientLevelPrediction' package to build an ML model to determine-

Which patients with new onset Type II Diabetes Mellitus (T2DM) will go on to develop Coronary arteriosclerosis in native artery (CA) within 2 years?

Study Design

Target cohort	New Onset Type II Diabetes Mellitus (T2DM)
Outcome cohort	Diagnosis of Coronary arteriosclerosis in native artery
Time at Risk	2 years
ML Model	Regularized Logistic Regression
Covariates	Age, Gender, Race

Custom cohorts using SQL

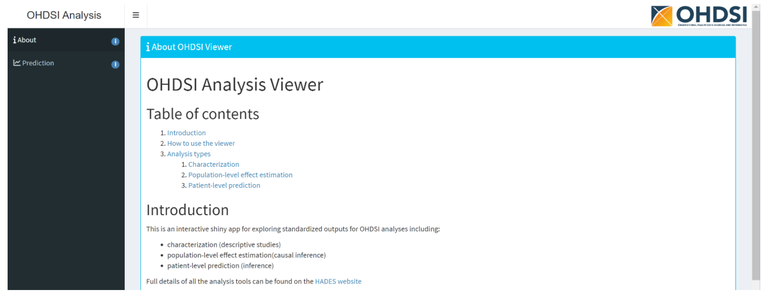
I created custom cohorts for my study using SQL and stored them in the 'dmcohorts' table of the 'results' schema of my PostgreSQL database. The code is available in the paper accompanying this poster and on the Author's GitHub account.

Model building

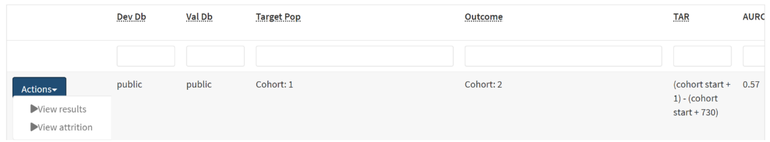
I then used the 'PatientLevelPrediction' R package and CMS 2008-2010 DE-SynPUF data in OMOP CDM to create an ML model in R studio. The code is available in the paper accompanying this poster and on the Author's GitHub account.

Results

I was then able to view my results in the OHDSI Analysis Viewer Shiny application.



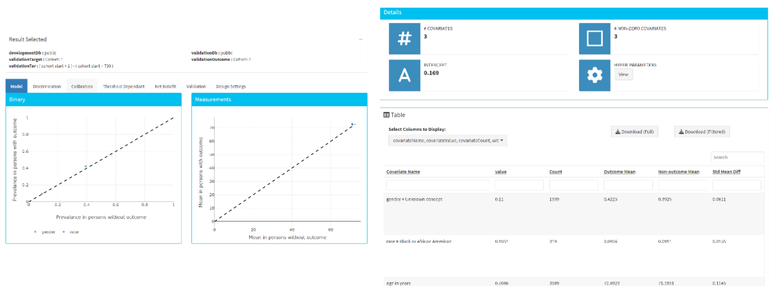
Clicking on the Prediction tab on the left leads to a page where all the models are listed by design and summarized. Selecting the 'View Model' option on the 'Action' button to the left leads to that model's summary page.



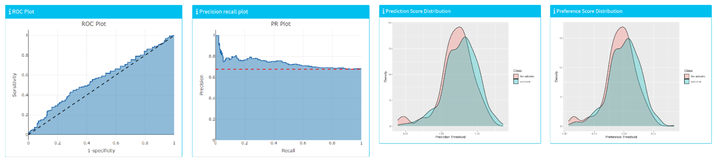
Here, we can see that my model had an AUROC (an indicator of model performance) of 0.57 and that out of my cohort size (T size) of 3389, 2291 showed the outcome (O size), which is an incidence rate of 67.6% (O Incidence (%)).

Clicking on the 'View Results' option on the 'Action' button brings up a Results page with several tabs.

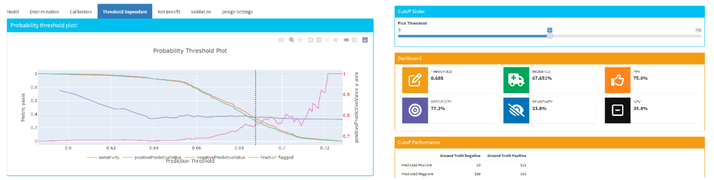
On the 'Model' tab we can see the coefficients of all the covariates and how common each covariate is among the people who had the outcome versus those who didn't (Outcome Mean and Non- Outcome Mean).



On the 'Discrimination' tab, we can see how the model is performing with the help of several plots- the ROC plot, which shows the tradeoff between specificity and sensitivity, the Precision-Recall curve, which shows the tradeoff between Positive Predictive Value (PPV) and sensitivity across thresholds, the F1 score plot, a box plot showing predicted risk for those with (class 1) and without (class 0) the outcome, prediction score distribution and preference score distribution, showing how well the model is able to discriminate between those with and without outcome. The metrics on this tab can be used to fine-tune and optimize the model.



The 'Threshold dependent' tab has the Incidence, Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) at different thresholds. We can pick a threshold to say- if the predicted risk is above that value, the outcome should be considered positive. The cutoff slider can be used to pick the threshold.



Once the threshold has been picked, the ML model we have trained so far can be applied to a fresh data set to create patient-level predictions.

Conclusion

In this poster, I have demonstrated how an R package developed following TRIPOD guidelines, the 'PatientLevelPrediction' can be used along with data in a standard format- CMS 2008-2010 DE-SynPUF in OMOP CDM- to develop an ML model which is easy to share and validate across data sets. Patient-level prediction could have several **applications** in the realm of clinical trials, some of which are-

1. Drug re-purposing: To determine if a drug the patient is being given already is likely to show an effect that was previously unknown. If the effect is therapeutic in nature, clinical trials can be carried out to gain regulatory approval for the new indication.
2. Post-marketing safety surveillance: To monitor patients who are prescribed specific drugs, in order to predict and mitigate the risk of adverse events, enhancing patient safety.
3. Personalized medicine: To determine if a patient's profile before a certain point in time is likely to make them react in a particular way to a drug/ therapy beyond that point and tailoring medical treatment to optimize outcomes.

Through these diverse applications, the integration of ML models with standardized clinical data has the potential to revolutionize the field of clinical trials, making it more data-driven, efficient, and personalized.

References

1. Framework for FDA's Real World Evidence Program. (2018, December). FDA. https://www.fda.gov/media/120060/download
2. Reps J. M., Schuemie M. J., Suchard M. A., Ryan P. B., & Rijnbeek P. (2018). 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, 25(8), 969-975. https://doi.org/10.1093/jamia/ocy032.
3. Centers for Medicare & Medicaid Services. (2013, January 15). CMS 2008-2010 Data Entrepreneurs' synthetic public use file (DE-SynPUF). CMS.gov. https://www.cms.gov/data-research/statistics-trends-and-reports/medicare-claims-synthetic-public-use-files/cms-2008-2010-data-entrepreneurs-synthetic-public-use-file-de-synpuf
4. AWS (n.d) CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) in OMOP Common Data Model. Retrieved on December 29, 2023 from https://registry.opendata.aws/cmsdesynpuf-omop
5. Google Cloud console. (n.d.). Synthetic Patient Data in OMOP. Retrieved on December 29, 2023 from https://console.cloud.google.com/marketplace/product/hhs/synpuf

Contact Information

Ashwini Yermal Shanbhogue
Email: ash23shan@yahoo.com
GitHub: https://github.com/ash23shan/From-Data-Iron-to-Data-Gold