Building Deep Learning Models with the OMOP CDM

Kristin Feeney, MPH[1], Dennis Robert, MBBS/MD, MSST[1], Prerna Patil, MHIM[1], Sergey Charkin, MS[1], Dan Housman, BSc[1], Yuval Koren, MSc, [1]
Haiping Xia, MS,PhD,[1] Jinlei Liu, MS[1]
[1]ConvergeHEALTH by Deloitte, Deloitte Consulting LLP, Newton, MA

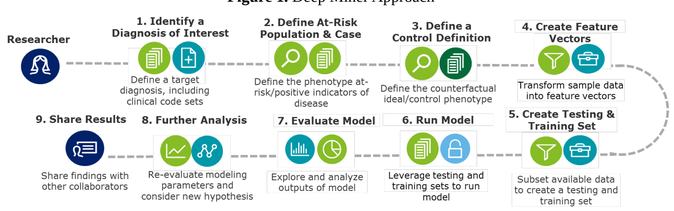
Introduction

Advanced cognitive learning algorithms ("deep learning") are an emerging data science technique with many potential applications in the Life Sciences industry.[1,2] As big data and elastic computing becomes more mainstream, deep learning models like convolutional neural networks (CNN) and recurrent neural networks (RNN) are more accessible than ever. Artificial intelligence experts agree these modeling techniques have the potential to match in predictive accuracy and even overcome the cognitive bias of today's conventional rule-based models.[3] To date, limited application exists in structured observational health data. To test the utility of these models in these data, our team designed a framework for testing RNNs and CNNs in observational health data transformed into the OMOP Common Data Model (CDM). This experiment aims to evaluate the accuracy of using a machine-driven view of mapping the relationship between medications, procedures, diagnoses or labs preceding a diagnosis of interest. It is intended to be extended to other research questions.

Approach

Our package, Deep Miner, combines open source methods, proprietary data transformation, machine learning and neural network algorithms. Deep Miner uses a case-control experiment, guided by subject matter input, built on top of the robust semantic mapping contained in the OMOP CDM to evaluate the accuracy of putting deep learning models head-to-head with traditional models to predict a disease of interest. Figure 1 below details our overall research process irrespective of model deployed.

Figure 1. Deep Miner Approach



Defining a Disease of Interest

For the purposes of this exercise, we created a case and control definition within a single therapeutic area (**Inflammatory Bowel Disease**) using synthetic data (**DeSynPUF data**) in **OMOP CDM v5** format. We defined the true and false cohorts as follows:

Table 1. True Cohort Definition

DISEASE DEFINITION (TRUE COHORT)
Inflammatory bowel disease consists of patients with either (Ulcerative Colitis and Crohn's Disease) who have a concept occurrence of at least one concept code associated to IBD: 4074815 (IBD), 4270915 (Chronic inflammatory small bowel disease)

Table 2. False Cohort Definition

CONTROL DEFINITION (FALSE COHORT)
Patients selected at random from the eligible observation period with greater than five recorded visits within the source data. Patient IDs are discarded from the random sample if they are also included in the true cohort patient list

Leveraging OHDSI Tools and the OMOP CDM

We first selected in eligible cases and controls using relevant concept IDs. We then used a combination of reports from OHDSI ACHILLES (Figure 2) and data characterization scripts to profile the frequency of OMOP concepts (i.e. conditions, drugs, observations) and identify potential relatedness in OMOP concepts. This allowed us to determine how to create hold-out logic to construct our 80-20 test-train splits. We then retain only non-zero feature vectors (i.e. only vectors with concept IDs).

Figure 2. OHDSI ACHILLES

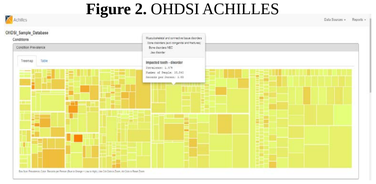


Figure 3. Sample Feature Vector

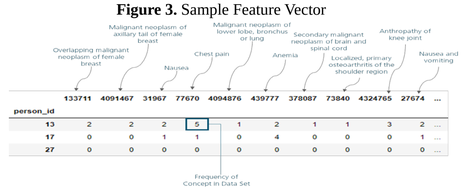


Figure 3 shows a representative feature vector. We experimented with tall-to-wide methods. It captures attributes from multiple domains (DRUG_EXPOSURE, CONDITION_OCCURRENCE, OBSERVATION, MEASUREMENT) as well as OMOP CDM concept mappings (CONCEPT, CONCEPT_ANCESTOR).

In this cohort definition, there were approximately 40,000 Concept IDs represented including approximately 8,000 Conditions, 10,000 Procedures and 19,000 Drug Exposures in our conceptual "bag of features".

Results

We were able to develop a framework to ingest structured observational health data to investigate the features that contribute to development of a specific diagnosis of interest. Table 3 and 4 show a side by side comparison of model performance.

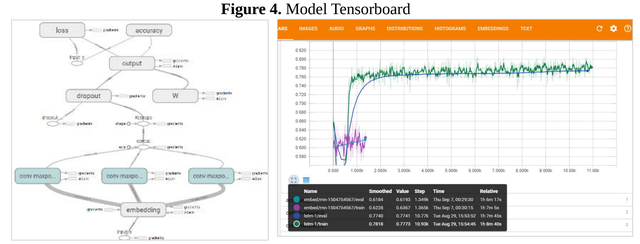
Table 3. Experiment Parameters

	SynPUF Data
Total number of patients	~2.5M
# of patients with IBD	96,578
# of control patients without IBD	100,748
# of All features	13,681
# of filtered Top features	187

Table 4. Model Performance

Models	AUC
Logistic Regression	0.73
Random Forest	0.74
Elastic Net	0.84
Convolutional Neural Networks	0.85
Recurrent Neural Networks	0.79

Figure 4. Model Tensorboard



Discussion

Synthetic data is created with the goal of providing a realistic set of claims data. It may still lack the full depth of authenticity and complexity of a real data set. We anticipate that neural nets will perform better on other real world data sets. Further stringency can be applied during cohort creation, including propensity matching methods, to control for the inherent heterogeneity of a target phenotype.

Conclusion

We were able to build a model pipeline on the OMOP CDM to facilitate rapid evaluation of specific hypothesis. This experiment shows that deep learning models, such as CNNs, have the potential to outperform traditional models when modeling thousands of parameters at the same time. More research is needed on additional data sets to externally validate these models.

References

1. Dong X, Qian L, Guan Y, Huang L, et al. "A multiclass classification method based on deep learning for named entity recognition in electronic medical records", *Scientific Data Summit (NYSDS) 2016 New York*, pp. 1-10, 2016.
2. O' Donoghue J, Roantree M, Van Boxtel M, (2015). "A Configurable Deep Network for high-dimensional clinical trial data", *Neural Networks (IJCNN) 2015 International Joint Conference on*, pp. 1-8, 2015, ISSN 2161-4407.
3. Tishby, N. June 2017. *Information Theory of Deep Learning*. [video] Available at: https://www.youtube.com/watch?v=bLqJHjXihK8&feature=youtu.be [Accessed 10 Oct. 2017].