Assessment of Pre-trained Observational Large Longitudinal models in OHDSI (APOLLO)

Martijn Schuemie[1,2], Yong Chen[3], Egill Fridgeirsson[4], Chungsoo Kim[5], Jenna Reps[1,4], Marc Suchard[2], Xiaoyu Wang[1,6], Chao Pang[7]

[1] Johnson & Johnson, [2] UCLA, [3] University of Pennsylvania, [4] Erasmus University Medical Center of Rotterdam, [5] Ajou University Graduate School of Medicine, [6] Duke University, [7] Columbia University

Background

Large language models (LLMs) have recently received significant attention because of their ability to comprehend complex linguistic structures, enabling, among other things, ChatGPT to participate in human-like conversations. These models have been applied to various domains, extending beyond text to include images processing, as exemplified by projects like Dall-E and Midjourney.

The Assessment of Pre-trained Observational Large Longitudinal models in OHDSI (APOLLO) project aims to explore the feasibility of employing pretrained models in the analysis of large healthcare databases, including electronic health records and administrative claims. The main form of these databases is time-stamped sets of codes, such as diagnosis codes, procedure codes, drug codes, and other time-stamped values, such as laboratory measurements.

Deep learning models such as LLMs are commonly applied in two stages: pre-training on a large dataset, followed by fine-tuning for a specific task. In the APOLLO project, pre-training involves using the entire database in the OMOP Common Data Model (CDM), which typically includes millions of persons. During pre-training, the model is supervised by withholding certain information and trained to predict the withheld information. As illustrated in Figure 1, the model can be trained in a forward-only manner, akin to the General Pre-trained Transformer (GPT) algorithm, [1] where it uses a patient's historical data to predict subsequent clinical events. Alternatively, a bidirectional approach, inspired by the Bidirectional Encoder Representations from Transformers (BERT) algorithm,[2] can be employed, where a specific concept is masked, and the model utilizes all available context to predict it. Fine-tuning builds upon the pre-trained model, further training it for a specific task.

These tasks may include:

- patient-level prediction, where a pre-trained model may prove more accurate with less training data than current non-pre-trained models.
- missing value imputation, which is almost identical to the bidirectional pre-training task.
- phenotyping, which can be thought of as a type of imputation.
- patient clustering, where nodes in the hidden layers may represent subgroups of interest.
- causal effect estimation, either by using the model for computing propensity scores, or directly eliciting effects learned by the pre-trained model.
- counterfactual prediction: given a choice between various treatment options, what is expected to happen to a patient in the future, for each treatment option?

We also suspect more potential applications will become apparent in the future.

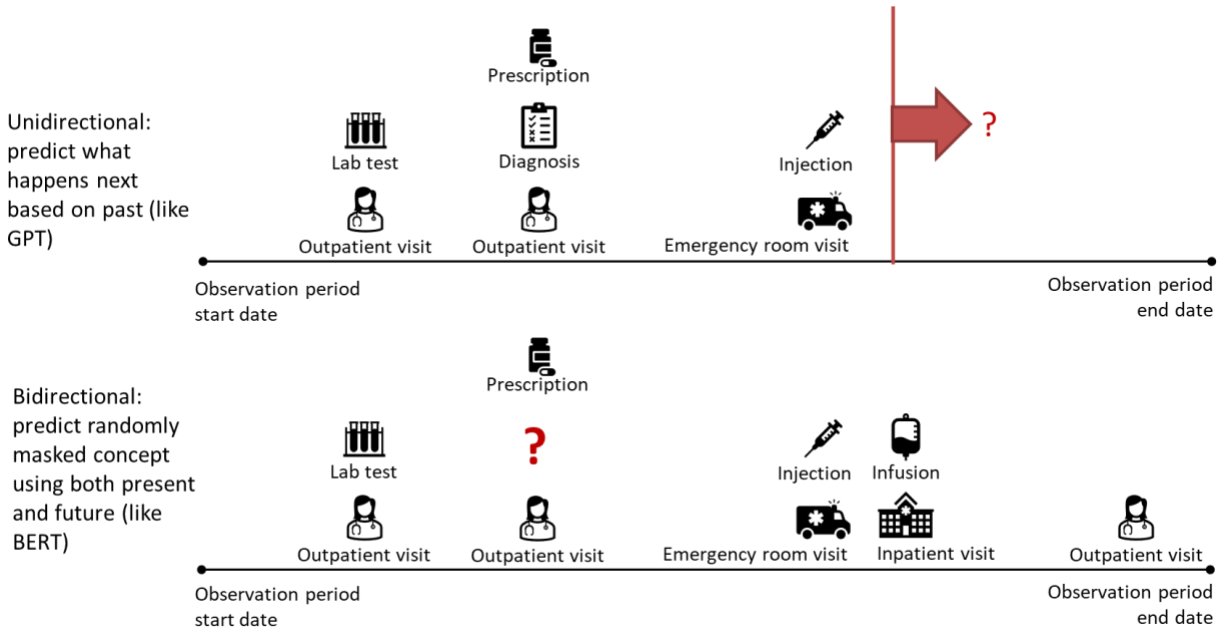

Figure 1. Two possible pre-training tasks for data in the OMOP Common Data Model (CDM): forward-only/unidirectional, and bidirectional

Early success in applying LLMs to healthcare data, such as CEHR-BERT, have demonstrated that using large pre-trained models on healthcare data is feasible.[3] Building upon this success, the APOLLO project aims to facilitate the applications of these models to any database within the OHDSI network, and evaluate their performance across various tasks. APOLLO is currently in progress, with preliminary results to be released in the latter half of 2023.

Methods

Architecture

The current architecture, illustrated in Figure 2, utilizes the OHDSI DatabaseConnector R package to establish a connection with the CDM database and extract data for either a sample or the full set of persons. Subsequently, these data are stored locally in the efficient Apache Parquet format. The stored data comprises a subset of the CDM tables and columns, encompassing all clinical domain tables (except the notes tables), and includes person IDs, visit occurrence IDs, concept IDs, numeric values, and several vocabulary tables

Because research in LLMs is done almost exclusively in Python, the remainder of the pipeline is implemented in Python. A pre-processing script converts the CDM data to sequence-information per person, before fitting the model using the PyTorch library.

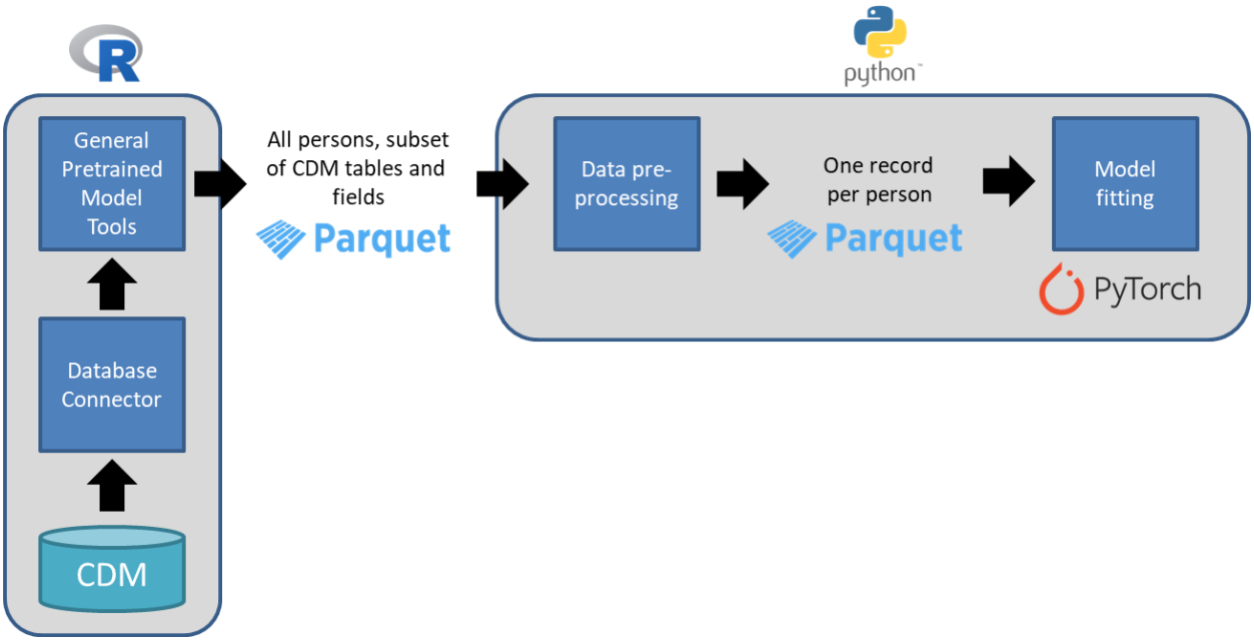

Figure 2. Overall architecture for pre-training

Evaluation benchmarks

Initial evaluations will use simulated data only. We have developed a simple simulator that uses a hidden-state Markov model to generate data in CDM format, including data for fine-tuning prediction tasks.

For patient-level prediction and causal effect estimation we will rely on existing OHDSI benchmarks.[4,5] For other tasks new benchmarks will be developed.

Where possible, performance will be compared to the current state-of-the-art, such as the algorithms implemented in the OHDSI PatientLevelPrediction package, and the CohortMethod package using large-scale propensity scores.

Analyses choices to evaluate

There are many analysis choices when developing general pre-trained models, as well as when fine-tuning. These include:

- Type of pre-training task: unidirectional or bidirectional? Predicting the next /masked event by choosing among all possible events, or by choosing among a limited set of candidates automatically selected for the training?
- Choice of input and output representation, including
    - How to represent elapsed time between events
    - Whether and how to encode and embed time, age, season, the day of the week, etc.
    - Which features to include. Should only the most prevalent concepts be included? Should drugs be mapped to ingredients? Etc.
- Model architecture, such as number of layers and number of nodes per layer, but also choice of activation functions.
- Training parameters, such as regularization, learning rate, and number of epochs.

- Data sources to use.

A set of combinations of these choices will need to be established, and evaluated using the benchmarks.

Results

Feasibility

In a feasibility study, the GeneralPretrainedModelTools R package was used to take a sample of two million persons from the Merative Marketscan CCAE database. Download took 1.8 hours, and Parquet files take 3.1GB of disc space. Pre-processing took 10 minutes, resulting in Parquet files totaling 2.3GB. A single epoch of pre-training took 20 hours on an NVIDIA A10G for a 121-million-parameter model.

Conclusion

APOLLO is a collaborative research project within OHDSI that aims to investigate the potential of leveraging pre-trained large models in the analyses of healthcare data. Drawing inspiration from the advancements in the field of LLMs applied to text and images, APOLLO seeks to extend these successes to healthcare applications. Although formal evaluations are pending, it is worth noting that a member of the APOLLO team has previously documented notable outcomes using this approach. The infrastructure for performing methods experiments on large pre-trained models is currently being developed, and an early feasibility study suggests that this infrastructure will be able to handle the extensive datasets available within OHDSI.

Despite existing uncertainties surrounding the applications of large pre-trained models to healthcare data at scale, the potential for transformative impacts is promising.

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems, 2020, Volume 33, pages 1877-901.
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
3. Pang C, Jiang X, Kalluri KS, Spotnitz M, Chen R, Perotte A, Natarajan K. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. Proceedings of Machine Learning for Health, volume 158 of Proceedings of Machine Learning Research, pages 239–260. PMLR, 04 Dec 2021.
4. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR, Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data, J Am Med Inform Assoc. 2018 Aug 1;25(8):969-975. doi: 10.1093/jamia/ocy032
5. Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, Ryan PB, Madigan D, Hripcsak G, How Confident Are We About Observational Findings in Healthcare: A Benchmark Study, Harv Data Sci Rev. 2020;2(1):10.1162/99608f92.147cc28e. doi: 10.1162/99608f92.147cc28e