Using a Large Clinical Language Model to Predict HIV Risk From Transformed Structured EHR Data

Yiyang Liu, PhD MPH (yliu26@ufl.edu) 1, Ziyi Chen, MS 2, Suman Pogul, MS 1, Hwayoung Cho PhD, RN 3, Mattia Prosperi, PhD 1, Yonghui Wu, PhD 2

1. Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida.
2. Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida.
3. Department of Family, Community and Health Systems Science, College of Pharmacy, University of Florida.

Key Takeaways

- **By transforming structured tabular EHR data into narrative text, we can leverage the power of large language models for disease prediction.**
- **Our clinical LLM (GatorTron) demonstrates high discrimination power in HIV risk prediction (AUC = 0.88), a comparable performance to traditional ML models but with reduced needs for data preprocessing.**

Background

Why this topic? Identifying people at elevated HIV risk can support timely prevention (e.g., PrEP)

Current challenge: Traditional machine learning prediction rely on 1) feature selection based on expert's domain knowledge, 2) intensive manual preprocessing EHR to extract and code variables

Opportunity: Large language models (LLMs) can directly interpret clinical text and make predict.

Objective: Evaluate whether a clinical LLM (GatorTron) can predict HIV risk from text-transformed EHR data

Methods

Data source: a multi-site clinical network (OneFlorida+)
Outcome of interest: new HIV diagnosis
Sample size: n=54,718 (3,569 people with new HIV diagnosis, 54,718 controls)

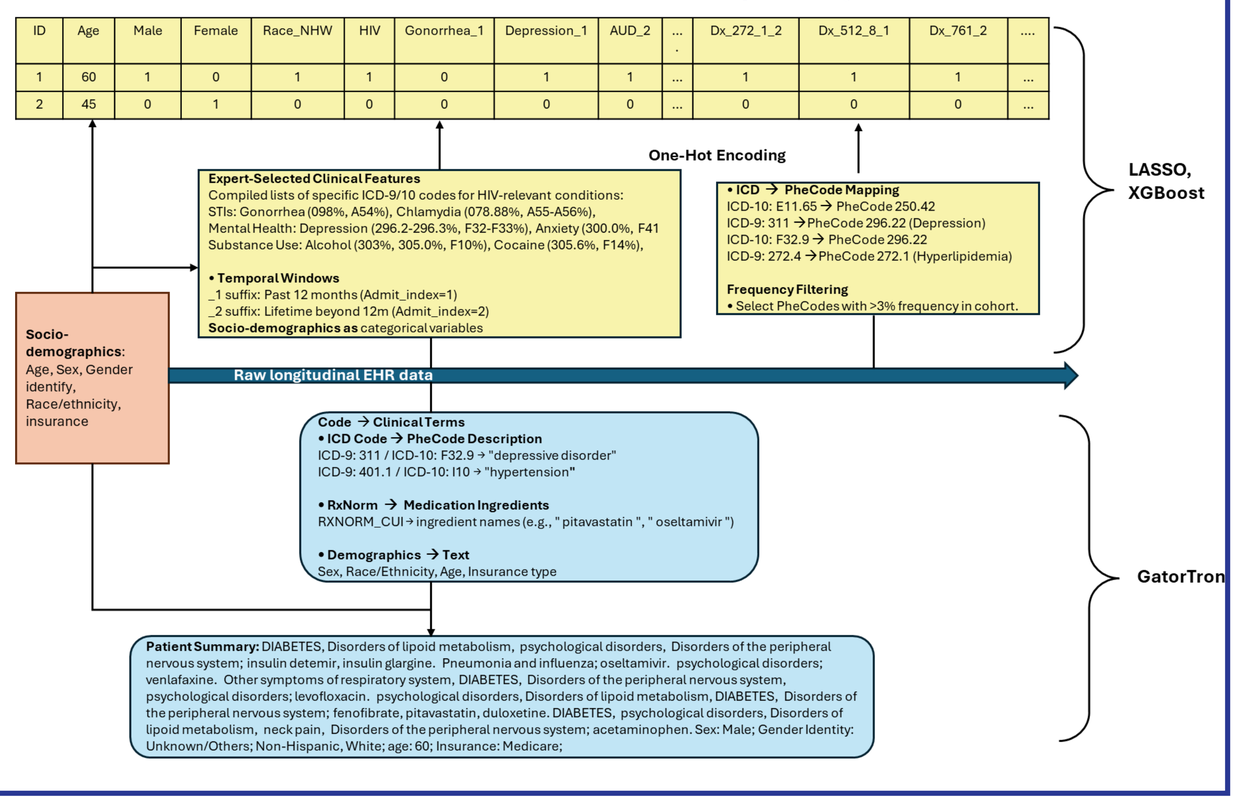
Machine learning methods

- GatorTron: a BERT based clinical LLM trained on de-identified clinical notes and biomedical literature.
- Traditional Machine Learning models: LASSO, XGBosst

Prepare EHR for prediction (also see figure below)

1. LLM:
- Extract socio-demographics into text;
- Transform ICD 9 and 10 diagnosis codes into corresponding text narrative;
- Medication prescription: use APIs provided by the U.S. National Library of Medicine to map all RxCUIs to the names of their active ingredients and convert these into text

2. Traditional ML
- One-hot coding socio-demographics
- Expert select clinical features
- Data-driven feature selection with frequency filter



Results

Table 1. Comparison of model performance metrics for the clinical large language model versus traditional machine learning methods.

	AUC	Precision (PPV)	Recall (sensitivity)	F1 score	Accuracy
LLM model					
Model 1: Dx only	0.74	37.5%	19.0%	25.2%	93.2%
Model 2: Dx+Dem	0.87	74.0%	42.4%	53.9%	95.6%
Model 3: Dx+Dem+Rx	0.88	62.7%	46.7%	53.5%	95.1%
XGBoost model					
Dem+Dx (expert+freq)	0.89	30.0%	69.9%	42.0%	88.3%
LASSO model					
Dem+Dx (expert+freq)	0.87	26.1%	69.8%	38.0%	86.2%

Model fairness consideration: comparing AUC across key demographic groups

- By Sex: male 0.88 vs female 0.87
- By Age group: Aged 13-29 0.84, Aged 30-44: 0.85, Aged 45+ 0.88
- By Race/ethnicity: Non-Hispanic White 0.86, Non-Hispanic Black, Hispanic: 0.86

Acknowledgements This work was supported by the National Institute on Allergy and Infectious Diseases R01AI172875 (MPI Prosperi/Bian), National Institute on Mental Health R34MH135768 (MPI Liu/Cho), and R21MH137736 (MPI Naar/He). Research reported in this publication was supported in part by the OneFlorida+ Clinical Data Network, funded by the Patient-Centered Outcomes Research Institute numbers CDRN-1501-26692 and RI-CRN-2020-005