Machine Learning on Structured EHR Data for Prediction in Schizophrenia: Feature engineering and pipeline construction

Kamyar Yazdani, Abhi Jadhav, Aakash Thumaty, Sanya Kochhar, Pranav Warman, Jane P. Gagliardi, MD, MHS[2], Jessica D. Tenenbaum, PhD[1]
[1]Department of Biostatistics & Bioinformatics and [2]Psychiatry, Duke University Medical Center, Durham, NC
✉ jessie.tenenbaum@duke.edu

DukeHealth

Background/Objective

- Schizophrenia is a severe mental disorder that affects ~1.1% of the US population and often requires lifelong treatment.[1]
- Improvement in the ability to predict which patients will need more intensive care in the near future could have a significant clinical value to prevent ED visits.
- This research project aims to apply machine learning to structured clinical data to predict patient ED visits and hospital admission.
- The first semester of this Bass project was dedicated to data acquisition, exploration, and wrangling as well as designing a feature matrix including both direct data and derived features using hierarchical terminologies and clinical prior knowledge.

Data

- A de-identified version of structured EHR data were extracted from Duke's data warehouse using DEDUCE (Duke Enterprise Data Unified Content Explorer) under Duke IRB protocol (Pro00081628)
- Inclusion criteria was at least one inpatient or two outpatient schizophrenia-related diagnosis codes in encounters from 1/1/2014 to 4/10/2017.
- Data types included demographics, diagnosis codes, and medications for encounters during the period in question.
- Overall, our dataset included ~115,000 encounters for 2,800+ patients.

[1]Schizophrenia. (n.d.). Retrieved from https://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml

Patient Population

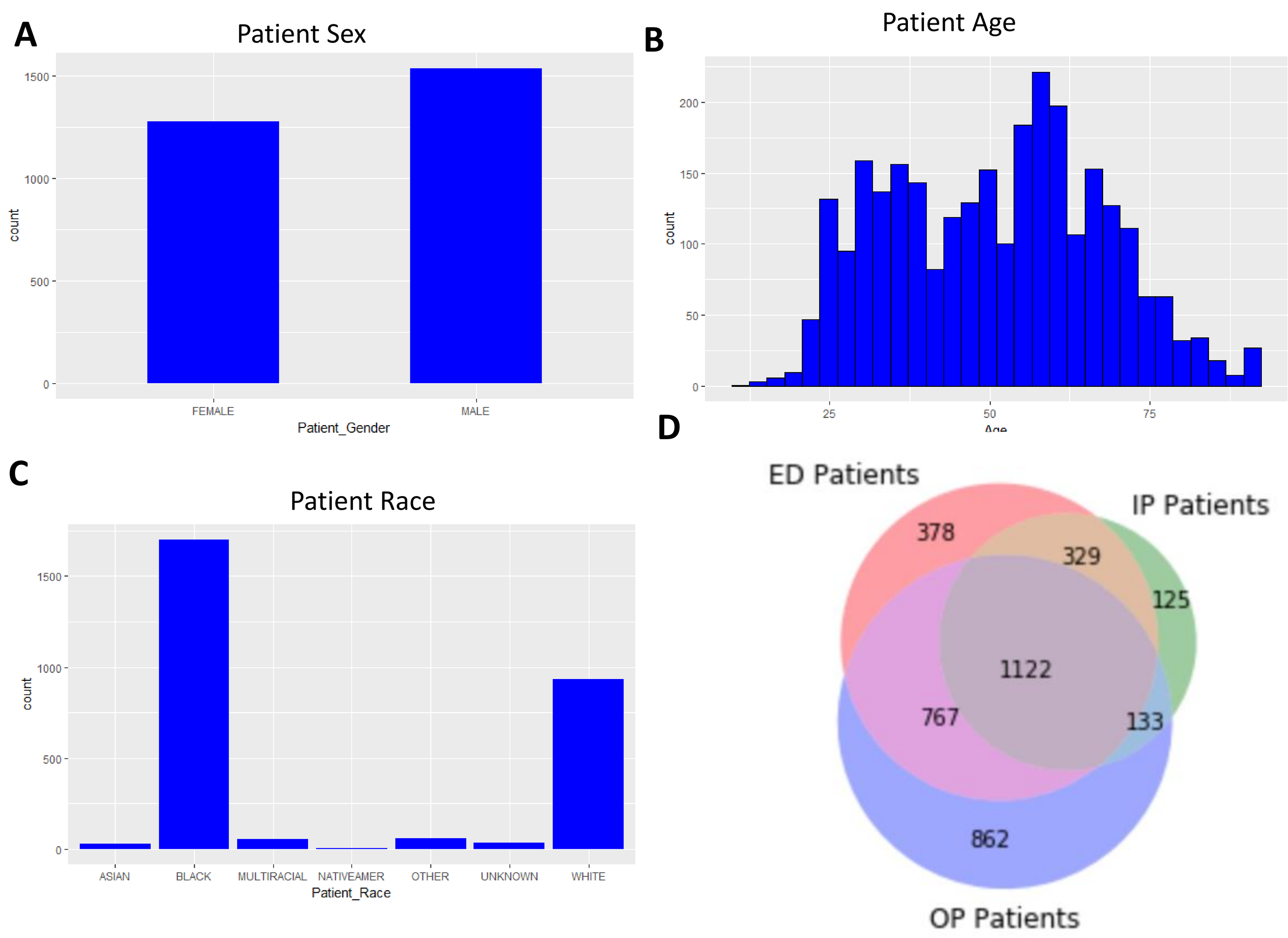


Figure 1: A-C. cohort demographics. D. Counts and overlap of inpatient, outpatient, and ED encounters.

Feature Matrix Construction

- Each row of the feature matrix represents a patient encounter
- Direct features include demographics, insurance information, visit type- IP (inpatient) vs OP (outpatient), total hospital stay, and psychiatric medication data.
- Computed features include the trend of patients' encounters over time, drug classes, and disease classes, which were derived by mapping ICD codes to hierarchical SNOMED codes.
- Medication and diagnosis data are represented as binary values based on whether that diagnosis or drug was recorded at a given encounter.
- Structured data will ultimately be combined with NLP-based symptom data.

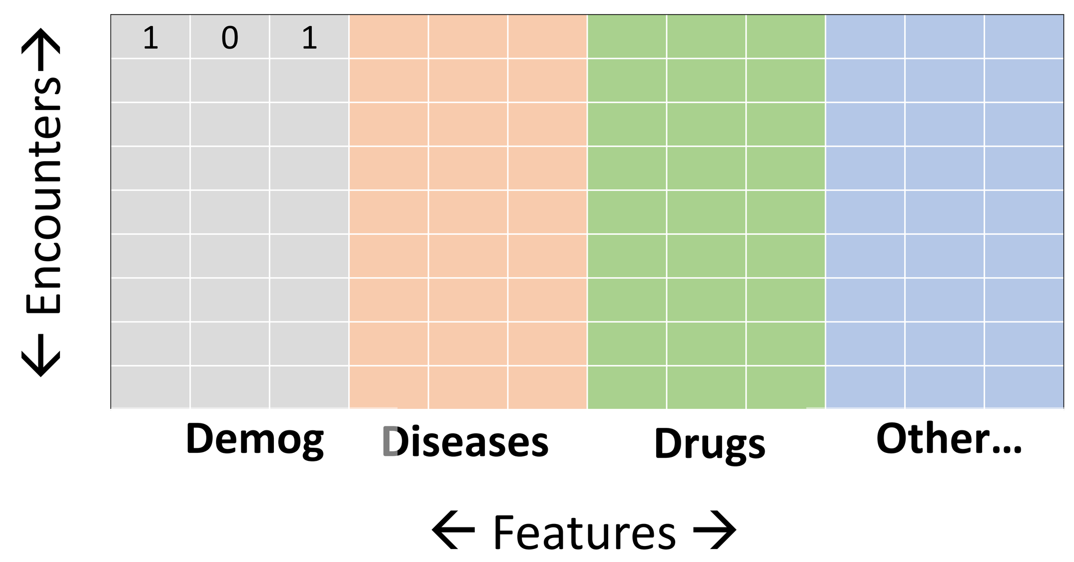


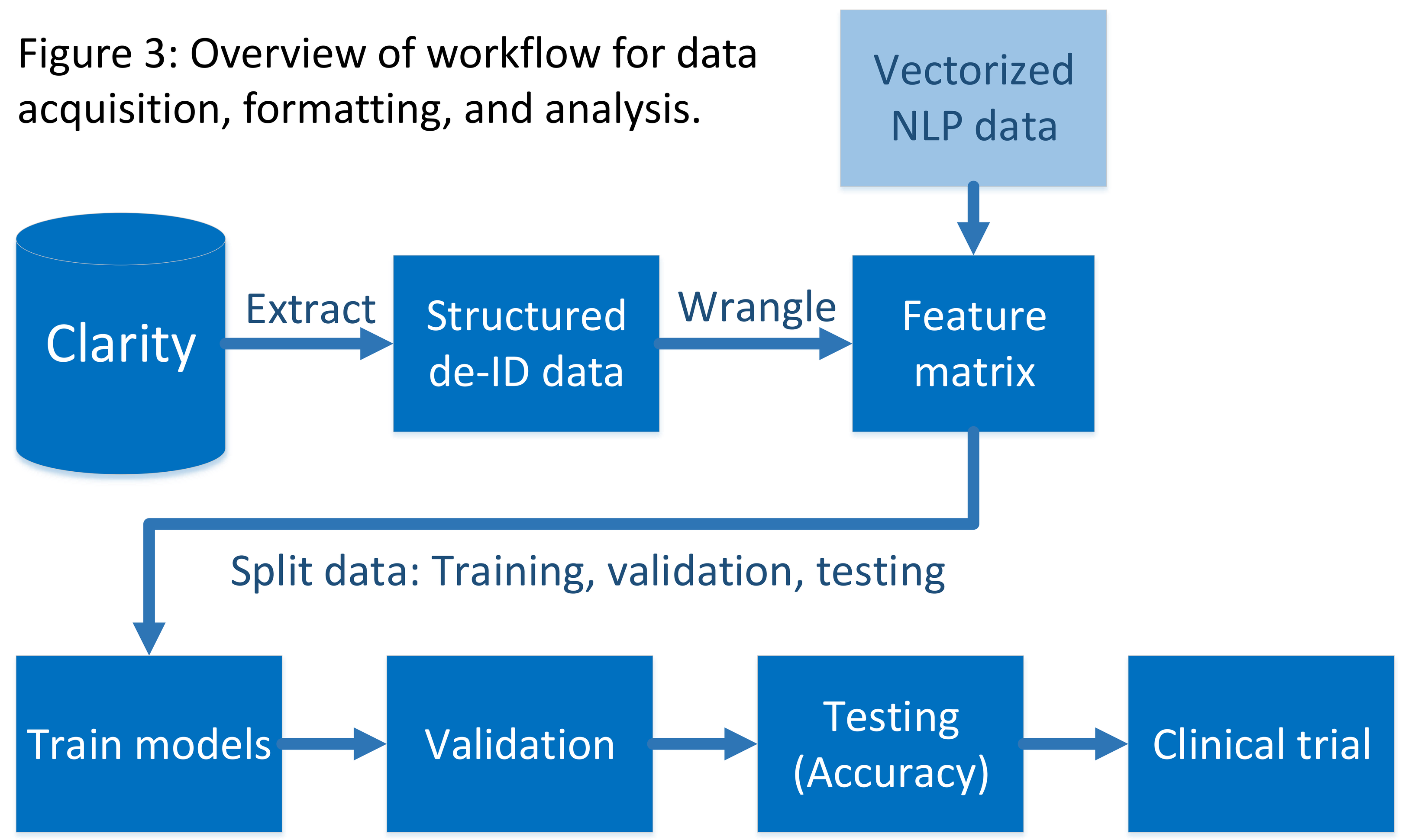
Figure 2: Feature matrix for data analysis. Rows represent encounters. Columns represent demographic data and the presence/absence of clinical elements including medications and diagnoses.

Major Challenges

- Noisy EHR data- missing values, nonsensical dates, etc.
- Project done in parallel with NLP analysis, requiring both de-identified and identified data with different permissions for different team members, jittered dates, and post-hoc data mapping
- Mapping ICD codes to SNOMED, given codes are not 1:1
- How to handle data across time

Pipeline

Figure 3: Overview of workflow for data acquisition, formatting, and analysis.



Conclusions and Future Directions

- Project progress to date has focused on data acquisition, data wrangling, and feature engineering.
- Next steps will be to perform supervised machine learning, evaluating accuracy in predicting which patients would benefit from more intensive resources following an encounter with the health system.
- Once accuracy is assessed and deemed clinically valuable, we aim to expose our algorithm to clinicians through a client-facing application (e.g. through SMART on FHIR) to assist in planning follow-up care and resource allocation.

Acknowledgements

The authors would like to acknowledge Bass Connections; team members Scarlett Zhou, Dr. Myung Woo, Dylan Liu, Stephen Evans, Casey Riffel, Allison Young, Dr. Gopalkumar Rakesh, Dr. Nigam Shah, Dr. John Beyer, Olamiji Sofela, and Colette Blach. JDT was funded by grant K01 LM012529 from the National Library of Medicine.