Sentinel

Predicting Causes of Death from Structured Electronic Health Records Using Machine Learning

Mohammed A Al-Garadi[1], Ruth Reeves[1,2], Rishi J Desai[3], Michele LeNoue-Newton[1], Daniel Park[1], Shirley V. Wang[3], Judith C. Maro[4], Candace C. Fuller[4], Kueiyu Joshua Lin[3,5], José J. Hernández-Muñoz[6], Aida Kuzucan[6], Xi Wang[6], Haritha Pillai[3], Kerry Ngan[3], Jill Whitaker[1], Jessica A. Deere[1], Michael F. McLemore[1], Dax Westerman[1], Michael E. Matheny[1,2]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA,
[2]Geriatrics Research Education and Clinical Care Service & VINCI, Tennessee Valley Healthcare System VA, Nashville, TN, USA,
[3]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA,
[4]Harvard Pilgrim Health Care Institute and Department of Population Medicine, Harvard Medical School, Boston, MA, USA, [5]Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA,
[6]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD

Presented at the 2024 ISPE Annual Meeting by: Mohammed Al-Garadi, PhD
Research Assistant Professor
Department of Biomedical Informatics
Vanderbilt University Medical Center
Email: mohammed.a.al-garadi@vumc.org

Disclosures

- This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).
- The contents are those of the authors and do not necessarily represent the official views of, nor and endorsement, by FDA/HHS, or the U.S. Government.

- J.C.M and C.C.F are employed at HPHCI, an organization which conducts work for government and private organizations, including pharmaceutical companies.

Introduction

- Importance of rapidly identifying causes of death (CoD) for medical product surveillance. For example:
  - Rapidly identifying death and causes of death is important in medical product safety studies
  - In the US, structured EHR data provide a rich source of clinical information to pharmacoepidemiology studies, but death information is often incomplete, and cause of death information is typically not available
- Challenges in Death Reporting in US EHR Systems:
  - Delayed availability of CoD information
  - Death information often incomplete in US EHR systems
  - Variability in data across different EHR systems

Objective

- Evaluate the capacity of machine learning (ML) using structured EHR data to predict CoD.

Methods

Cohort, Data Processing, Feature Extraction, ML Models, and Evaluation Metrics

Data Sources and Reference Standards

- VUMC Cohort
  - Cohort of VUMC patients consisted of 13,796 patients with last encounter at VUMC between 2019 and 2021 with matched records within the National Death Index. Data on patients was extracted from the VUMC Research Derivative, which is a database of clinical data derived from VUMC's enterprise data warehouse and maintained by the Office of Research Informatics.
- National Death Index
  - Database of death records in the US compiled from state vital statistics office. Serves as the reference standard for both date of death and cause of death.
- National Center for Health Statistics
  -

Basic Cohort VUMC NDI Cohort Demographics

Cohort Demographics

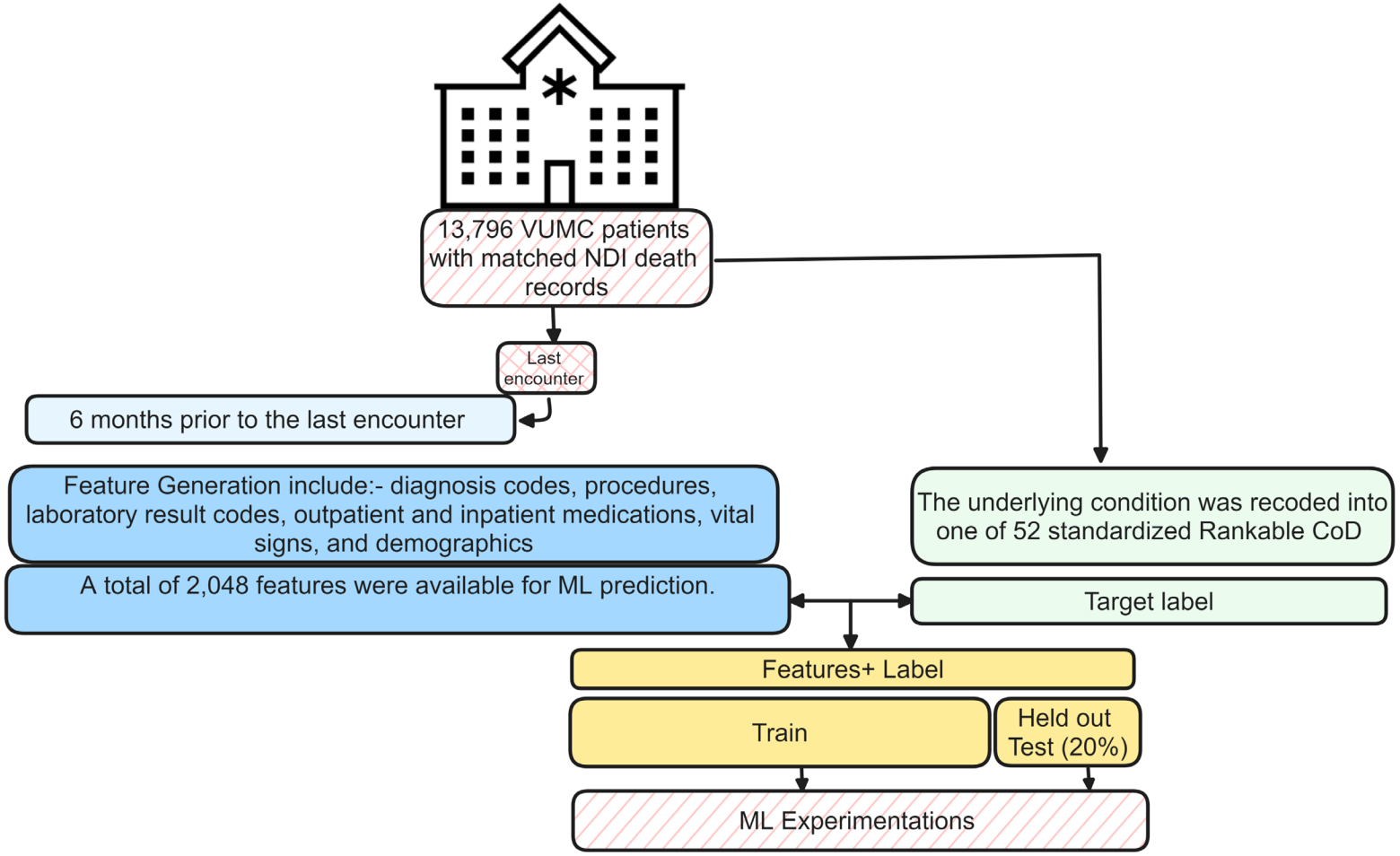
Age	
Mean ± SD	67.76±15.62
Quartiles (25, 50, 75)	59 **70** 79
Race (%)	
White	81.97%
Black	11.89%
Asian	0.88%
American Indian/ Alaskan Native	0.12%
Other	5.12%
Ethnicity (%)	
Non-Hispanic	93.25%
Hispanic	1.34%
Unknown	5.41%
Gender (%)	
Male	57.36%
Female	42.62%
Unknown	0.02%

Cohort: 13,796 VUMC patients with matched NDI death records

Data: VUMC Research Derivative

Data Processing and Feature Extraction

- We collected structured data for 6 months prior to the last encounter.
- Features include diagnosis codes, procedures, laboratory result codes, outpatient and inpatient medications, vital signs, and demographics.
- A total of 2,048 features were available for ML prediction.
- The underlying condition was recoded into one of 52 standardized Rankable CoD (Target label).
- Data split 80/20 into train/test sets, stratified by CoD labels.



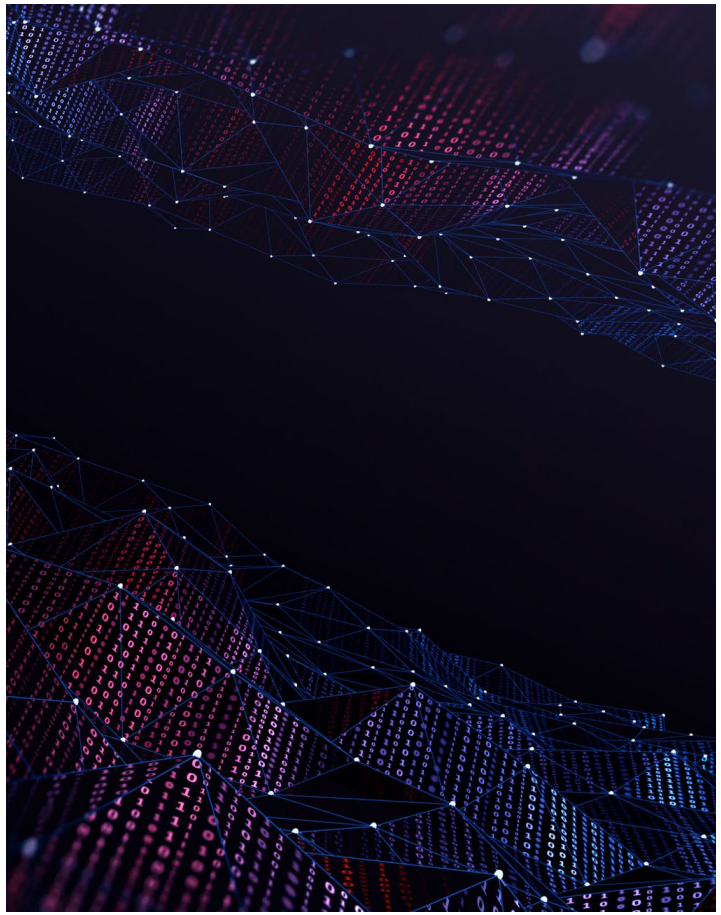
Machine Learning Models

Machine learning

1. **XGBoost (Extreme Gradient Boosting):**
   - Efficient, scalable gradient boosting implementation.
   - Enhances performance by applying boosting techniques.
2. **Random Forest:**
   - Ensemble method with multiple decision trees.
   - Outputs class mode (classification) or mean prediction (regression).
3. **K-Nearest Neighbors (KNN):**
   - Non-parametric, uses k closest examples for classification.
   - Based on feature space proximity.
4. **Support Vector Machine (SVM):**
   - Finds hyperplane to separate classes.
   - Used for classification and regression.

Evaluation Metrics

1. **Weighted AUC (Area Under the Curve):**
   - Measures overall model performance across all thresholds.
   - Accounts for class weights to handle imbalanced datasets.
2. **Individual Class AUC:**
   - AUC calculated for each class separately.
   - Assesses model's ability to distinguish each class from others.
3. **Weighted F-measure Summary**
   - The weighted F-measure balances precision and recall across all classes, accounting for class frequency to provide a comprehensive evaluation of model performance on imbalanced datasets.



Results

Weighted f-measure, and weighted AUC the top 15 in 52 Rankable CoD classification

Structured Data COD

52 COD Name	Counts	Percentage
Malignant Neoplasm	4155	30.12%
Diseases of heart	2192	15.89%
COVID19	1044	7.57%
Unintentional injuries	1042	7.55%
Cerebrovascular disease	612	4.44%
Chronic liver disease and cirrhosis	364	2.64%
Chronic lower respiratory disease	353	2.56%
Diabetes Mellitus	306	2.22%
Nephritis, nephrotic syndrome, and nephrosis	194	1.41%
Influenza and pneumonia	188	1.36%
Septicemia	157	1.14%
Intentional Self Harm	153	1.11%
Parkinson disease	131	0.95%
Essential hypertension and hypertensive renal disease	129	0.94%
Alzheimer	115	0.83%
Other	2,658	19.27%

The selected 52 CODs represents 86% of all the CODs in the selected cohort and the top 15 CODs represent 80%.

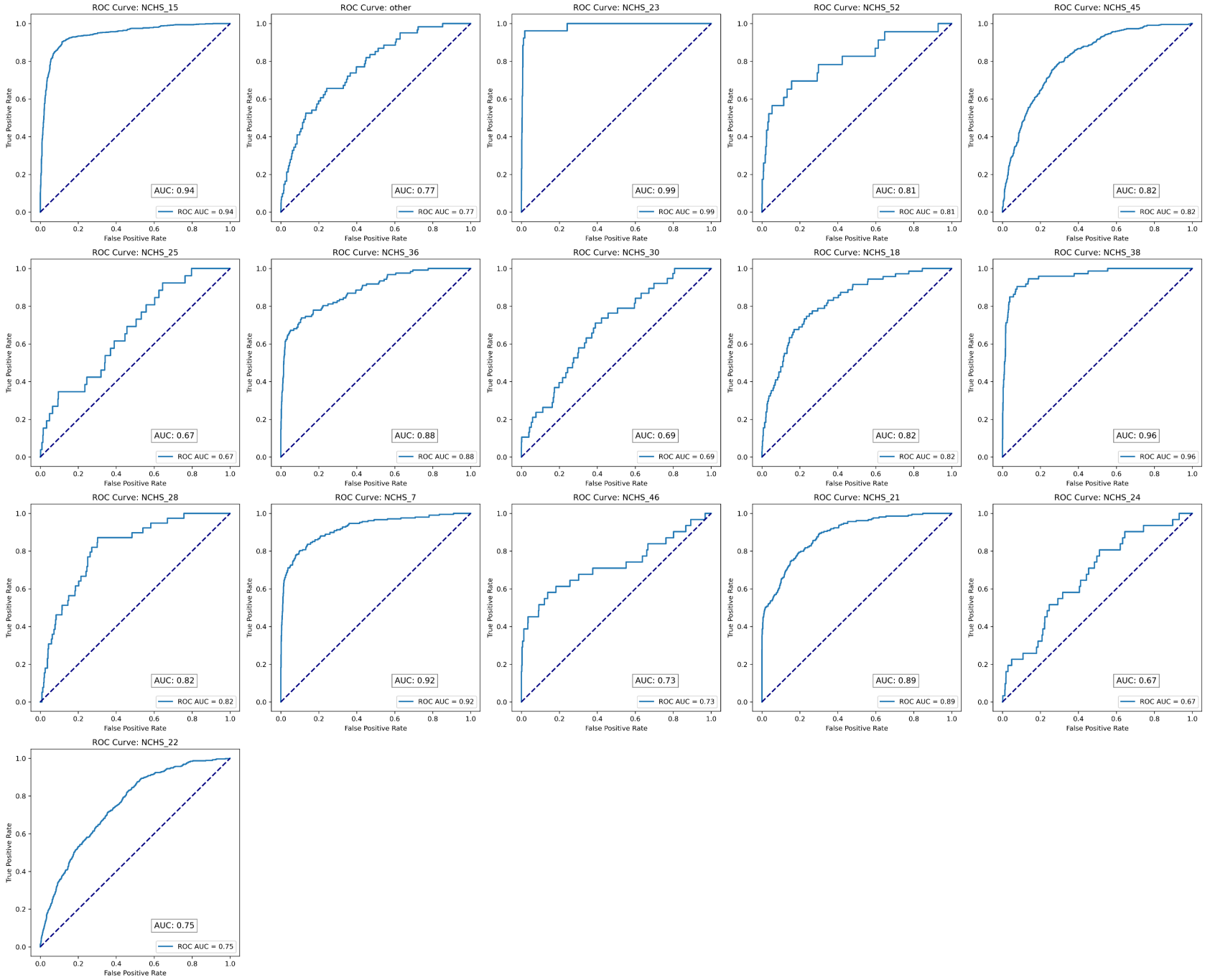
Structured Data COD Prediction Results

- **Results from 52 using structured data only based on test data**

Model	Average Weighted AUC	Weighted F-measure
XGBoost	0.86	0.74
Random Forest	0.79	0.72
KNN	0.65	0.51
SVM	0.73	0.59

AUC Results of XGBoost Algorithm (Best Performing Model) for Top 15 Classes in 52 Rankable Causes of Death Classification Based on Held-Out Test Data

Disease	Counts	AUC
Malignant Neoplasm	4155	0.94
Diseases of heart	2192	0.99
COVID19	1044	0.81
Unintentional injuries	1042	0.82
Cerebrovascular disease	612	0.67
Chronic liver disease and cirrhosis	364	0.88
Chronic lower respiratory disease	353	0.69
Diabetes Mellitus	306	0.82
Nephritis, nephrotic syndrome, and nephrosis	194	0.96
Influenza and pneumonia	188	0.82
Septicemia	157	0.92
Intentional Self Harm	153	0.73
Parkinson disease	131	0.89
Essential hypertension and hypertensive renal disease	129	0.67
Alzheimer	115	0.75
Other	2573	0.77



Discussion

Key findings, future directions

Key Preliminary Findings

- Good discrimination in CoD prediction using high-dimensional features from structured EHR data.
- XGBoost outperformed other models, likely due to its ability to handle sparse data and nonlinear relationships.
- We observed poor discrimination in CoD prediction for some causes, such as cerebrovascular disease which may be due to close correlation with other diseases. In contrast, conditions like heart diseases and nephritis showed excellent discrimination.

Future Directions

- Develop hybrid model: Structured EHR data + Unstructured clinical notes
- Integrate models into probabilistic CoD estimation
- Create more effective mortality predictor
- Focus on generalizable and portable ML models:
  - Cross-validation across different sites
  - Enhance model transferability

Thank You

Questions?