Sentinel

Enhancing Causes of Death Prediction from Electronic Health Records through Multi-Modal Integration of Structured and Unstructured EHR Data

Mohammed Al-Garadi, PhD
Research Assistant Professor
Department of Biomedical Informatics
Vanderbilt University Medical Center
Email: mohammed.a.al-garadi@vumc.org

Mohammed A Al-Garadi[1], Rishi J Desai[3] , Kerry Ngan[3] , Michele LeNoue-Newton[1] , Ruth Reeves[1,2], 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], 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

November 8th, 2024

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

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 and unstructured notes 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
  - Significant variability in the quality of data across different EHR systems.

Objective

Objective

- Leverage both structured and unstructured EHR data in a multi-modal approach to predict CoD

- Explore the potential of integrating textual data embeddings with traditional structured features for more comprehensive and accurate predictions

Methods

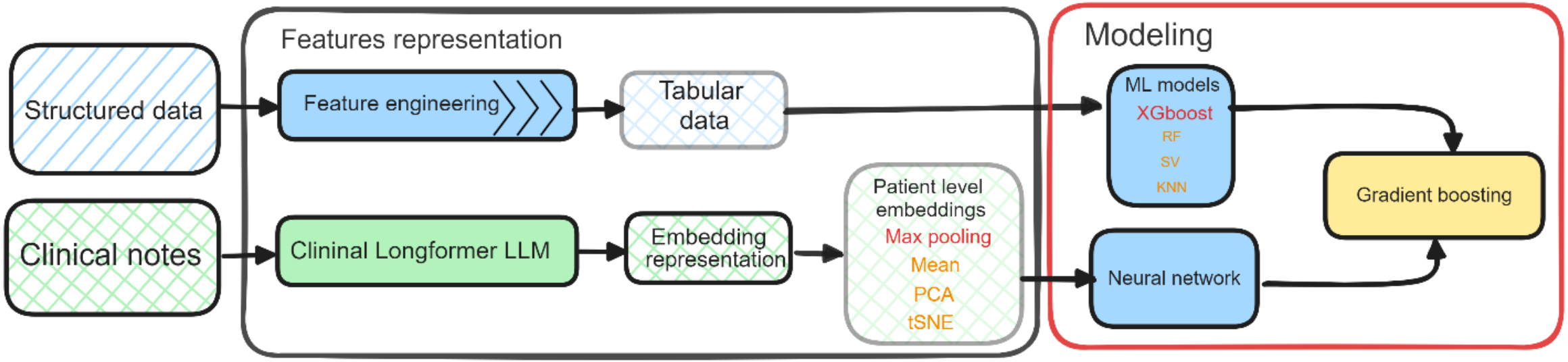
Cohort, Data Processing, Feature Extraction, ML Models, and Evaluation Metrics

Data Sources and Reference Standards

- VUMC Cohort
  - Cohort of VUMC patients consisted of 13,708 patients with last encounter at VUMC between 2019 and 2021 with matched records within the National Death Index. National Death Index

- National Center for Health Statistics
  - Compiles annual reports on births and deaths within the US. Mortality data is grouped into 52 Ranked Causes of Death for reporting purposes that are defined by specific ICD-10 codes.

Multi-Modal Machine Learning Framework for Cause of Death Prediction



Data Sources

- Structured Data (2048 dimensions): Represents clinical categories such as ICD codes, lab test results, medications, vital signs, and demographic data (e.g., age, gender, race).
- Unstructured Clinical Notes (768 dimensions): Derived from free-text clinical notes in the EHR. These notes are processed using Clinical Longformer LLM to generate 768-dimensional patient-level embeddings.

Feature Extraction

- Structured Data: Transformed through feature engineering methods, extracting variables from EHR
- Unstructured Data: Clinical Longformer is used to process text and extract relevant patient-level features, converting them into embeddings for use in predictive models.

Modeling

- Both structured features and unstructured embeddings are combined and fed into various ML models including XGBoost, RF, SVM), KNN, and a neural network. This diverse set of models allows for robust performance evaluation.

- **Top 15 Causes of Death (Labels):** Includes major causes such as heart disease, cancer, stroke, respiratory disease, Alzheimer's, diabetes, and self-harm.

Structured Features

- **Data Source** *(6-month lookback from the last encounter):-*
  - *Collected Variables:*
    - Diagnoses (ICD codes)
    - Procedures (CPT/HCPCS codes)
    - Lab tests (LOINC codes)
    - Medications (RxNorm)
    - Vitals (e.g., blood pressure, heart rate)
    - Demographics (age, gender, race)
  - *Controlled Vocabularies:*
    - LOINC
    - RxNorm,
    - Clinical Condition Grouping (Clinical Classifications Software (CCS))

Unstructured Features

- **Data Source**: 2,710,729 narrative clinical notes
- **Document Embeddings**:
  - Generated using **Longformer** (processing up to 4,096 tokens) and the output 768 Embeddings
- **Patient-Level Aggregation**:
  - Max pooling
  - Averaging
  - Principal Component Analysis (PCA)
  - t-SNE

Data Split and labels

- Train/Test Split: 80% training set and 20% testing set
- CoD labels:
  - NDI coded into 52 Rankable Causes of Death
  - Top 15 leading causes predicted

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

Top 15 CODs

52 COD Name	Counts	Percentage
Malignant Neoplasm	4155	30.3%
Diseases of heart	2192	16.0%
COVID19	1044	7.6%
Unintentional injuries	1042	7.6%
Cerebrovascular disease	612	4.5%
Chronic liver disease and cirrhosis	364	2.7%
Chronic lower respiratory disease	353	2.6%
Diabetes Mellitus	306	2.2%
Nephritis, nephrotic syndrome, and nephrosis	194	1.4%
Influenza and pneumonia	188	1.4%
Septicemia	157	1.1%
Intentional Self Harm	153	1.1%
Parkinson disease	131	1.0%
Essential hypertension and hypertensive renal disease	129	0.9%
Alzheimer	115	0.8%
Other (These represent the remaining 37 CODs)	2,573	18.8%

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

COD Prediction Results

Models	Average F-measure	Average AUC
Structured data Alone (SVM)	0.59	0.73
Structured data Alone (RF)	0.72	0.79
Structured data Alone (KNN)	0.51	0.65
Structured data Alone (XGboost)	0.74	0.86
Structured (XGBoost) and Unstructured Data (Max Pooling)	0.79	0.90
Structured (XGBoost) and Unstructured Data (Mean)	0.78	0.89
Structured (XGBoost) and Unstructured Data (PCA)	0.75	0.87
Structured (XGBoost) and Unstructured Data (tSNE)	0.78	0.90

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 (S) (95% CI)	AUC (S+U) (95% CI)
Malignant Neoplasm	4155	0.94[0.93-0.92]	0.95 [0.94-0.96]
Diseases of heart	2192	0.99[0.99-0.97]	0.98 [0.99-0.97]
COVID19	1044	0.81[0.79-0.83]	0.86 [0.85-0.87]
Unintentional injuries	1042	0.82 [0.81-0.83]	0.86 [0.85-0.87]
Cerebrovascular disease	612	0.67[0.65-0.69]	0.75 [0.73-0.77]
Chronic liver disease and cirrhosis	364	0.88[0.86-0.90]	0.89 [0.87-0.91]
Chronic lower respiratory disease	353	0.69[0.70-0.68]	0.78[0.77-0.79]
Diabetes Mellitus	306	0.82[0.80-0.84]	0.80 [0.82-0.78]
Nephritis, nephrotic syndrome, and nephrosis	194	0.96[0.95-0.97]	0.97 [0.96-0.98]
Influenza and pneumonia	188	0.82[0.81-0.83]	0.86 [0.84-0.88]
Septicemia	157	0.92[0.89-0.95]	0.91 [0.89-0.95]
Intentional Self Harm	153	0.73[0.71-0.75]	0.82[0.79-0.85]
Parkinson disease	131	0.89[0.93-0.85]	0.89[0.93-0.85]
Essential hypertension and hypertensive renal disease	129	0.67[0.62-0.71]	0.76[0.73-0.79]
Alzheimer's disease	115	0.75[0.73-0.77]	0.79[0.77-0.81]
Other (These represent the remaining 37 CODs)	2573	0.77[0.74-0.80]	0.85[0.83-0.87]

S=structured and U=unstructured

Discussion

Key findings, future directions

Key Findings

•**Enhanced Mortality Prediction through Multi-Modal Modeling**: Integrating structured data with unstructured clinical notes significantly improves the accuracy of predicting CoD.

•**Performance Improvements**:

- •Weighted AUC improved from 0.86 [0.84-0.87] using only structured data to 0.90 [0.89-0.92] when unstructured data was incorporated.

- •Out of the 15 conditions, 10 show statistically significant improvements (non-overlapping 95% confidence intervals) when unstructured data is added to the mode

•**Significant Gains in Less Common Conditions**:

- •Notable AUC increases were found for chronic lower respiratory disease (9%), cerebrovascular disease (8%), essential hypertension (9%), and intentional self-harm (9%), all of which showed statistically significant improvements.
- •Unstructured notes provided valuable signals, especially for conditions with fewer samples, addressing underestimation in structured data.

Future Directions

Develop a Unified LLM Model:

- Explore a single Large Language Model (LLM) capable of understanding both structured and unstructured data for improved outcome predictions.

Expand Population Diversity and Scale:

- Apply the multi-modal approach to larger, more diverse cohorts to enhance model generalizability and robustness.

Address Data Quality and Completeness:

- Develop strategies to mitigate the impact of incomplete or poor-quality data on model performance.

Thank You

Questions?