Heart Failure Prediction with Natural Language Processing using Electronic Health Records

Michael Cho, Keith Woh, Samuel Yusuf

Motivation and Goal

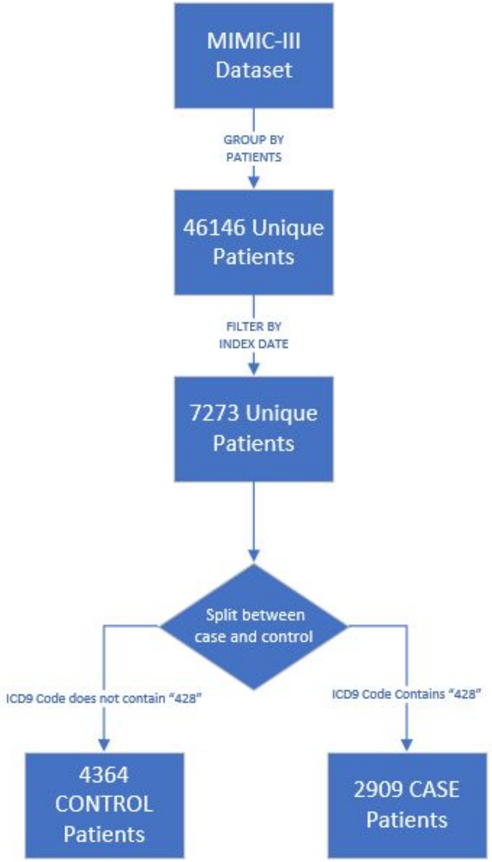
- Heart Failure affects an estimated 6.5 million people in the United States, with projections showing a 46% increase in prevalence from 2012 to 2030
- While prediction of heart failure using structured data in Electronic Health Records (EHR) has yielded promising results, there is still a wealth of unstructured text data in EHRs to explore that can be used to improve predictions
- **Goal:** utilize the unstructured text data in EHRs, in addition to structured EHR data, to better predict heart failures and thus improve patient outcomes.

Method

- Data Filtering
- Preprocessing Structured Data
- Preprocessing Unstructured Data
- Feature construction
- Modelling

Data Filtering

1. Filtered by Index Date
2. Grouped MIMIC-III dataset by Patients
3. Split between case and control patients
4. Randomly removed control patients to balance the dataset



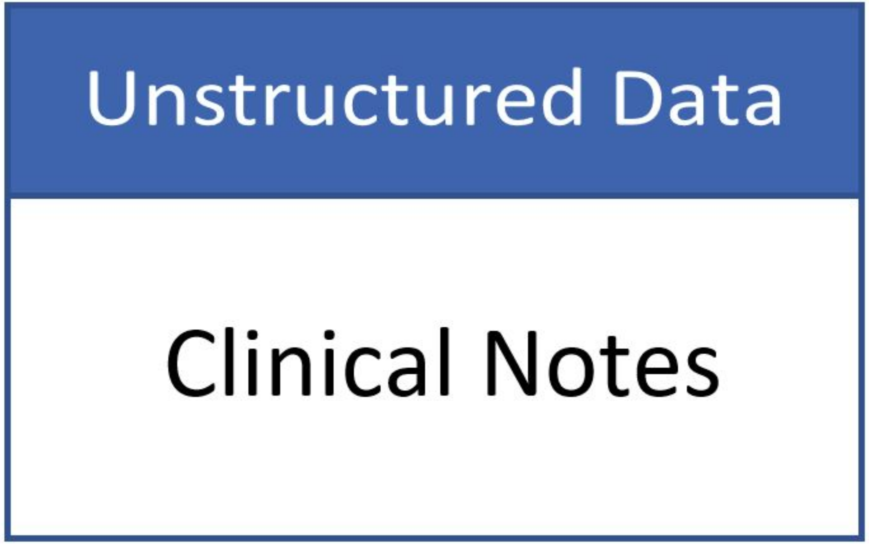
Preprocessing Structured Data

- Filtered according to index date
- Aggregated across all visits
- **Medication, diagnostic, and procedure** - counted occurrence of each event
- **Lab** - average lab value

Structured Data
Medication
Diagnostic
Lab
Procedure

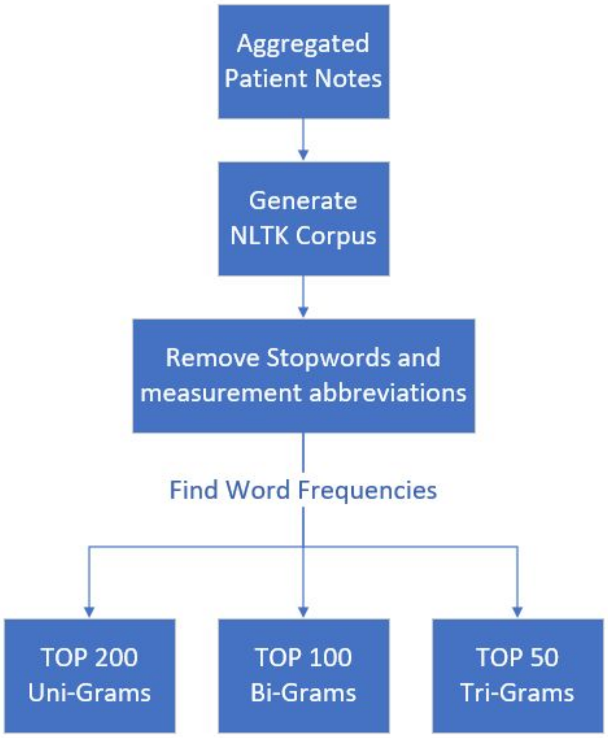
Preprocessing Unstructured Data

- Taken from “TEXT” column of clinical notes
- Notes aggregated by Patient
- NLTK Python library used to find features by word frequency



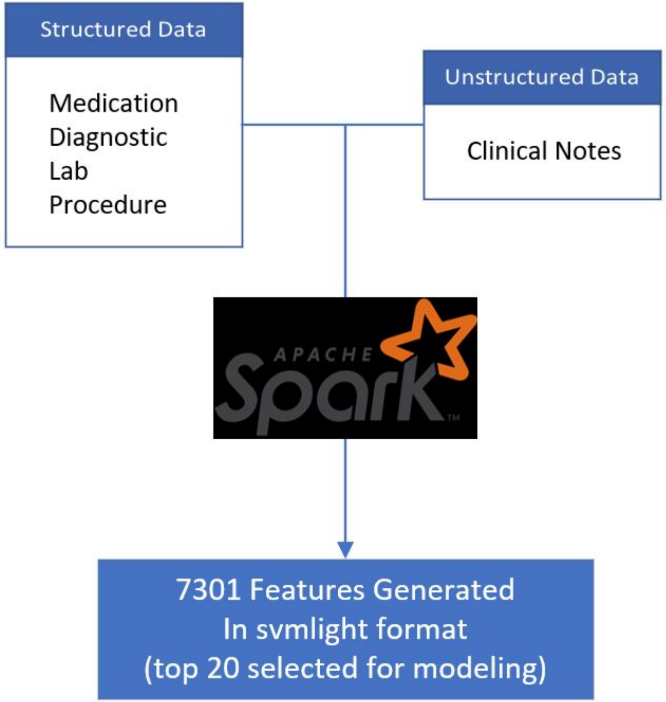
Feature Construction - Notes Analysis

- NLTK used to generate corpus for all patient notes
- Stopwords and Measurement Abbreviations removed (i.e. mg, pt, cm)
- Top word frequencies extracted as features



Feature Construction - Aggregation

- All features combined from structured and unstructured data at the patient-level using Spark
- Svmlight file format was generated for use in our prediction modeling step.



Modelling

- Linear Support Vector Machine Classifier (baseline)
- Support Vector Machine Classifier with L1 Regularization
- Decision Tree Classifier
- Random Forest Classifier
- Feedforward Neural Network
- Majority Voting Ensemble

Results - Classifiers

Model Name	**Accuracy**	**AUC**	**Precision**	**Recall**
Linear Support Vector Machine Classifier (baseline)	61.1%	64.2%	56.31%	53.86%
Support Vector Machine Classifier with L1 Regularization	**69.90%**	74.34%	68%	53.38%
Decision Tree Classifier	65.6%		59.13%	54.5%
Random Forest Classifier	70.4%		**71.13%**	**47.87%**

Results - Feedforward Neural Network

Neural Network			Accuracy	AUC	Precision	Recall
Hidden Layers	Nodes in each layer	Learning Rate				
2	256	0.001	0.672	0.720	0.658	0.434
3	256	0.001	**0.691**	0.734	**0.684**	0.472
4	256	0.001	0.682	**0.748**	0.601	**0.690**
4	512	0.001	0.690	0.742	0.634	0.595

Results - Predictive Features

- Top 20 most predictive features when using SVM L1
- 15 Structured features
- 5 Unstructured features

Feature Type	Feature Name
Structured	proc99291
Unstructured	textmoderate
Structured	medtuss5l
Unstructured	textventricular
Structured	meddesi10
Structured	diag9351
Structured	diag4255
Structured	diag1573
Structured	proc32020
Unstructured	textcompared
Structured	diag42971
Unstructured	textvalve
Structured	diag25082
Structured	lab51006
Structured	medenal5
Structured	proc99261
Structured	proc99232
Structured	proc99239
Structured	diag3970
Structured	medntg100pb

Results - Majority Vote Ensemble

Models	Accuracy	Precision	Recall
SVM Classifier (L1 Regularization) Random Forest Classifier Neural Network	72.0%	69.7%	50.2%

Conclusion

- Models show promising results, with up to 72.0% accuracy
- Unstructured data features were predictive, and added to the quality of the models
- **Future Work** - improvement of model development and improvement of feature extraction methodologies.
    - Recall is low, neural nets could be used to improve model

References

1. Ananthakrishnan AN, et al. Improving Case Definition of Crohn's Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing: A Novel Informatics Approach. Inflamm Bowel Dis. 2013 June; 19(7): 1411-1420.
2. Bashir S, Qamar U, Javed MY. An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis. International Conference on Information Society. 2014; 38(4): 259-264
3. Benjamin EJ, et al. Heart Disease and Stroke Statistics 2017 Update [Internet]. American Heart Association, Inc. 2017; 135:e379. Available from: circ.ahajournals.org/content/circulationaha/early/2107/01/25/CIR.0000000000000485.full.pdf
4. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text [Internet]. 2018 Feb. Available from: https://arxiv.org/pdf/1802.05695.pdf
5. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early Detection of Heart Failure Using Electronic Health Records: Practical Implications for Time before Diagnosis, Data Diversity, Data Quantity and Data Density. Circ Cardiovasc Qual Outcomes. 2016 Nov; 9(6): 649-658.
6. Pakhomov S,Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure. The American Journal of Managed Care. 2007 Jun; 13(6): 281-288.
7. Shu D, R Kannan M, Siddhartha J. Using Natural Language Processing to Screen Patients with Active Heart Failure: An Exploration for Hospital-wide Surveillance [Internet]. Division of Health and Biomedical Informatics, Northwestern University. Available from: https://arxiv.org/pdf/1609.01580.pdf
8. Thomas AA, Zheng C, Jung H, Chang A, Kim B, Gelfond J, Slezak J, Porter K, Jacobsen SJ, Chien GW. Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World J Urol. 2014 Feb; 32(1): 99-103.