Assessment of Content and Structure of Clinical Notes at an Academic Medical Center

Yixin Li MS[1], Harlan M. Krumholz MD SM[1,2], and Wade L. Schulz MD PhD[1,3]

[1]Yale-New Haven Hospital Center for Outcomes Research and Evaluation; [2]Yale School of Medicine, Department of Internal Medicine; [3]Yale School of Medicine, Department of Laboratory Medicine

Background

- Data from the EHR provides the opportunity to use real-world and real-time information to assess outcomes and improve predictive models
- It is estimated that 80% of healthcare data are unstructured sources, such as clinical notes
- Extracting relevant features from unstructured sources is a complex process, and descriptive statistics about the content are not well-described
- The ability to copy-forward notes within the EHR potentially introduces outdated, inaccurate, or unnecessary information

Objectives

- Describe the content and diversity of clinical notes within a large, academic healthcare system
- Identify potential features that can be used for feature engineering in downstream models, such as note similarity, frequency, and distribution

Methods

- All clinical notes from Yale New Haven Hospital from January 2014 through December2015
- Notes extracted in delimited file and converted to JSON, then analyzed with custom Python scripts

- Assessed basic descriptive statistics and lexical content stratified by note type, encounter type, and author specialty

$$\text{Type}-\text{Token Ratio(TTR)} = \frac{Number\ of\ \boldsymbol{unique}\ tokens\ (Vocabularies)}{Number\ of\ token(wordS)}$$

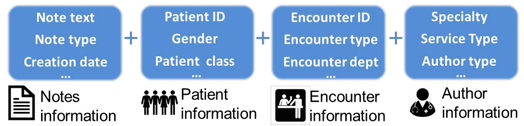
- Calculated the similarity for each combination of H&P, ED notes, and progress notes for two patients to assess the feasibility of similarity analysis in free-text notes

Results

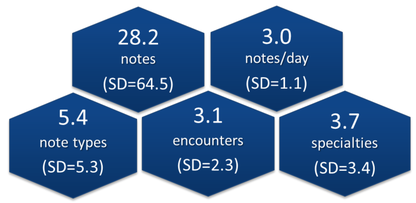
Data Description

- Nearly 1 million unique patients
- 25 million clinical notes
- 20 note features extracted, including note text

Features for YNHH Clinical notes



Number of data elements per patient



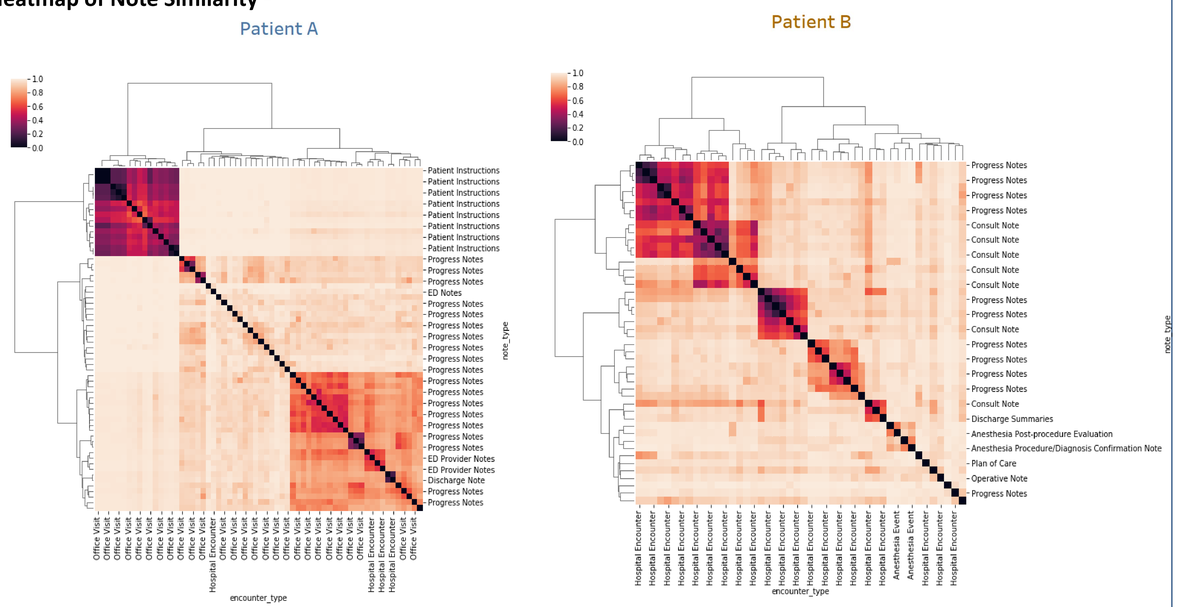
Note Features

Category	Number of types	Percent of top 10 types	Percent missing
Note types (progress, discharge, etc.)	121	87.4%	<0.01%
Encounter types (hospital, outpatient, etc.)	77	95.5%	<0.01%
Author specialties (IM, EM, etc.)	151	32.4%	48.0%

Data and Lexical Diversity by Top 10 Note Types

Note type	% Notes	Average #vocab	Average #words	Average #sentence	Average #word /sentence	Forest Plot of Lexical Diversity by Top 10 Note Types
Progress Notes	30.9%	244.3	478	25.3	22.4	
Telephone Encounter	20.8%	25.8	33.1	13.6	2.5	
Plan of Care	17.1%	141.4	352.3	12	33.6	
ED Notes	5.7%	36.2	52.5	4.3	12	
Patient Instructions	4.9%	168.1	402.1	30.6	16	
ED Provider Notes	2.6%	378.8	777	60.2	14.2	
H&P	2.0%	342.9	754.8	34.3	18.4	
Consult Note	1.3%	423.9	942.5	40.4	26	
Discharge Summaries	1.1%	433.8	935.6	40.6	24.3	
All Notes	100%	158.6	325.4	23	16.7	

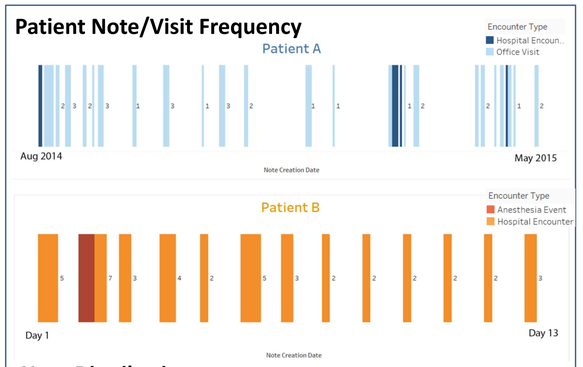
Heatmap of Note Similarity [1]



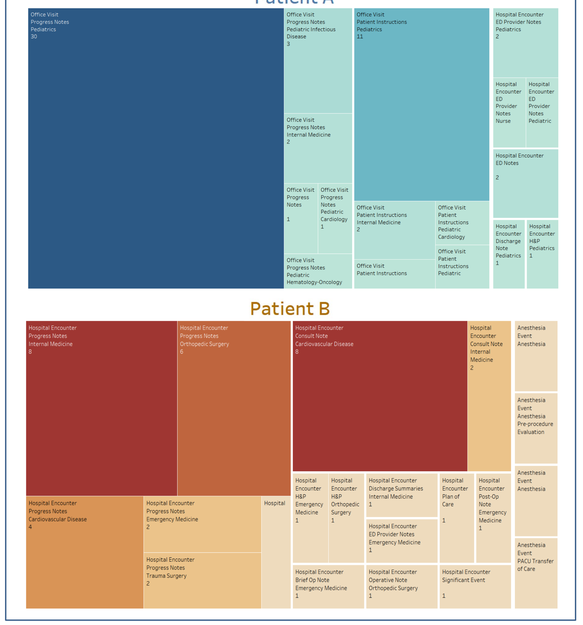
[1] The heatmap plot demonstrates the similarity of note content between every two notes during an admission or a period of outpatient visits. Each row/column represents an individual note. Darker color in the cross area shows higher similarity between the content of two notes.

Results

Patient Note/Visit Frequency



Note Distribution



Conclusions

- These data provide a comprehensive, descriptive assessment of the diversity in unstructured notes
- Multiple features can be rapidly extracted which may be beneficial in downstream analytic models
- Future work will apply these foundational data and results to predictive models, such as operative risk scores, to assess whether unstructured content can improve predictive accuracy