Characterizing Physicians Practice Phenotype from Unstructured Electronic Health Records

Sanjoy Dey, PhD[1], Yajuan Wang, PhD[1], Roy J. Byrd[1], Kenney Ng PhD[1], Steven R. Steinhubl, MD[2], Christopher deFilippi, MD[3], Walter F. Stewart, PhD[4]

[1]IBM Research, T.J. Watson Research Center, Yorktown Heights, NY USA
[2]Geisinger Health System, Danville, PA USA and Scripps Health, San Diego, CA USA
[3]Inova Heart and Vascular Institute, Fairfax, VA USA
[4]Sutter Health Research, Walnut Creek, CA USA

Abstract

Clinical practice varies among physicians in ways that could lead to variation in what is documented in a patient's electronic health records (EHR) and act as a source of bias to predictive model performance that is independent of patient health status. We used EHR encounter note data on 5,187 primary care patients 50 to 85 years of age selected for a separate case-control study covering 144 unique primary care physicians (PCPs). A validated text extractor tool was used to identify mentions of Framingham heart failure signs and symptoms (FHFSS) from the notes. Hierarchical clustering analyses were performed on the encounter note data for finding subgroups of PCPs with distinct FHFSS documentation behaviors. Three distinct PCP groups were identified that differed in the rate of documenting assertions and denials of mentions. Physician subgroup differences were not explained by patient disease burden, medication use, or other factors related to health.

Introduction

The widespread adoption of electronic health records (EHR) by US health care providers [1] is motivating a rapid growth in the use of predictive models to guide clinical decisions [2], to identify patients at high risk of future events (e.g., 30-day readmission) [3], and to detect disease early [4], among other applications. Copious longitudinal structured and unstructured data are captured by EHRs to characterize the patient's demographic (e.g., age, sex, address), health and treatment status, diagnoses, lab test results, and medication orders. As much as 80% of the EHR data is thought to be in unstructured form [5]. To effectively use EHR data it is important to understand how the data comes to be.

Physicians are the dominant sources of the data captured in EHR. However, physicians vary substantially and systematically in their clinical practices [6] resulting in variation in what is ordered, diagnosed, and documented for each patient, in medication prescribing and in preferences for the intensity of practice. Such physician practice styles are not idiosyncratic. Rather, practice style is known to be directly or indirectly influenced by medical school training, regional practice standards, local practice standards, and performance incentives, among other factors. However, most of these clinical practice differences are independent of the underlying health status of the patient or other characteristics of patients such as demographics or prior genetic predisposition to the disease.

These "practice phenotype" differences can significantly influence the quality of both structured and unstructured data in the EHR and act as a source of potential bias for any downstream analysis of EHR data. To build accurate computational models, we need to detect and normalize for such variances in physician behavior. However, previous studies have largely focused on differences in physicians' practices using structured EHR data to characterize diagnostic practice [6], regional practice pattern and standards [7], expertise, prior educational and training background of doctors [8], and the patient's treatment plan [9]. Differences may be identifiable by a limited number of practice phenotypes. Prior studies indicate that patient utilization phenotypes can be identified from structured data that are strongly influenced by provider preferences [10]. But, documentation in unstructured data has not been examined for such patterns.

In this study, we aim to explore whether there are practice phenotypes that characterize differences among physicians in how information gets into the unstructured EHR data. To test the phenotype hypothesis we use physicians' notes to determine whether there are practice phenotypes in the documentation of Framingham heart failure signs and symptoms (FHFSS). FHFSS are frequently documented in progress notes by PCPs, often years before HF diagnosis

[11]. We describe a systematic framework, based on a clustering approach, for characterizing the practice phenotype of PCPs using a large scale unstructured EHR data.

Methods

We focused on FHFSS to explore variations because physicians routinely document the presence or absence of the symptoms among older primary care patients independent of HF diagnosis. There are a number of challenges in extracting FHFSS from clinical notes that are actually related to practice phenotype. First, FHFSS are often documented in clinical notes much earlier than the clinical diagnosis of HF, which means that tracking the clinical notes longitudinally throughout the patient's medical history is needed. Second, a physician's practice might be affected by other confounding factors such as the patient's age, sex, prior medical history, other co-morbidities and the physician's expertise. Such confounding factors have to be removed in order to obtain an unbiased estimate of the actual practice phenotype defined by documentation behavior of FHFSS.

We carefully performed the study design and cohort selection to remove the effects of confounding factors during the feature extraction from clinical notes. A previously validated natural language processing (NLP) tool [12] was used for extracting FHFSS from clinical notes. For convenience, we used a large well characterized sample of primary care patients selected as controls for a prior nested case control study of heart failure [4]. We focused on the control group of patients to avoid any potential practice behaviors that might be a result of disease onset rather than actual practice phenotype of physicians. We were interested in broadly testing the hypothesis of physician documentation phenotypes in a representative sample of patients and not in a sample that was defined by any specific disease. Details of the patient sample and data source are summarized below along with the feature construction and analytic methods. This study was approved by the Geisinger Institutional Review Board.

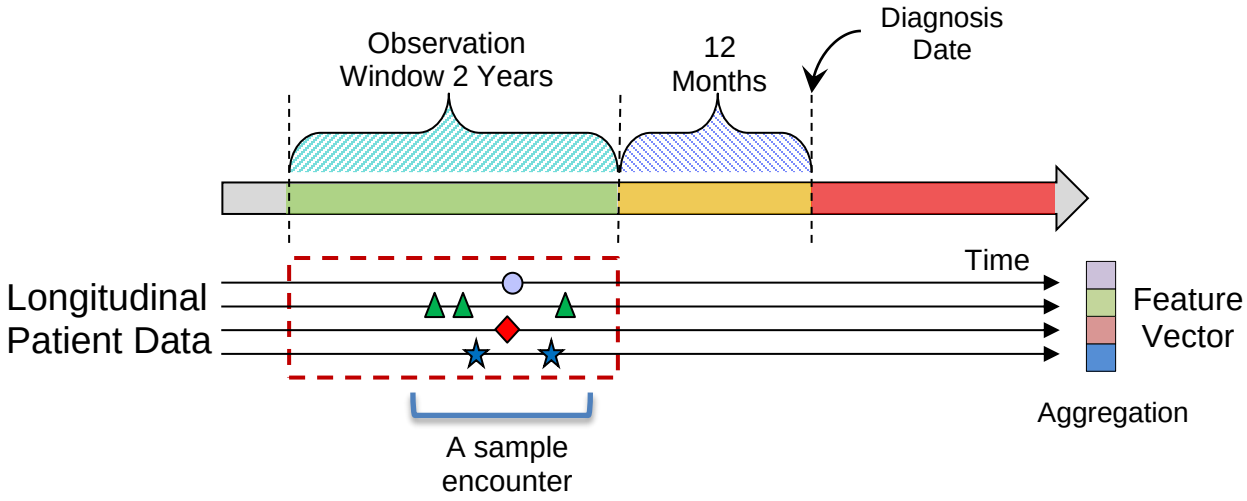

Figure 1. *Relation of observation window for use of data and the diagnosis date for cases and the same relative times for controls. For each patient, a feature vector is generated by applying appropriate aggregation functions on the longitudinal EHR patient data in the observation window.*

Study Design, Population, Setting, Source of Data

Longitudinal EHR data were obtained on patients, 50 to 85 years of age, from the Geisinger Clinic, a multispecialty group practice that provides care to approximately 400,000 residents in central and northeastern Pennsylvania. EpicCare EHR was installed at Geisinger before 2001.

A total of 1684 incident HF cases among Geisinger primary care patients were identified over the time period from 2003 to 2010 [4, 13, 14]. Up to 10 eligible primary care clinic-, sex-, and age-matched (in 5-year age intervals) controls were selected for each incident HF case for a total of 13,525 Geisinger control patients. Primary care patients were

eligible as controls if they had no HF diagnosis up to the one year post-HF diagnosis of the corresponding HF case. Control subjects were required to have their first office encounter within one year of the incident HF patient's first office visit and have ≥1 office encounters 30 days before or any time after the case's HF diagnosis date to ensure similar duration of observations among cases and controls. Nine or 10 controls were identified for 49% of the Geisinger cases; 1.5% of Geisinger cases had only 1 to 2 controls.

The primary care physician (PCP) was the unit of analysis and patient data were nested within each physician. Patient data were assigned to the physician who was the designated primary care provider as documented in the EHR. Assignment of patients to a PCP is documented in an EHR structured field. Note that different encounters of the same patients might correspond to different PCPs, if that patient utilized care from multiple providers. In that case, we mapped each patient to the unique PCP who treated that patient for the longest period of time. Of the total 13,525 patients, 11,268 were explicitly assigned to a PCP.

The EHR data used for this study were selected from the time period in controls that was 12 to 36 months before the incident diagnosis of HF in the corresponding matched case. This time period prior to the diagnosis of HF of the matched case is denoted as the "observation window" (Figure 1). Patients were excluded if: 1) they did not have an encounter note in the12-36 month observation window (this observation window retained the largest amount of patient data as described elsewhere [4]); 2) if the coverage duration period (i.e., total time span covering patient encounters with the PCP that are within the observation window) for the patient was less than or equal to 180 days; and 3) if a patient had a substantial number of documented chronic diseases during the observation window. This last step was used to minimize documentation variation among physicians that could be explained by patient's comorbidies other than disease burden. We considered 1148 ICD-9 diagnosis codes coming from three different types such as Chronic Disease (200 codes), Cardiometabolic Chronic Disease (743 codes) and Chronic Episodic Disease (214 codes) with only 9 codes being common between the first two categories. Most of the patients have very low chronic conditions as shown in Figure 2. Therefore, we only kept patient with at most five (out of 1148 codes) chronic conditions to retain maximum number of samples. The first two criteria of the obserbation windows of 12 to 36 months and the minimum coverage of 6 months led to 6862 patients in total, while imposing all three criteria reduced the samples to 5,187 qualified control samples to 144 PCPs in total. The data used for analysis was confined to this subgroup of PCPs.

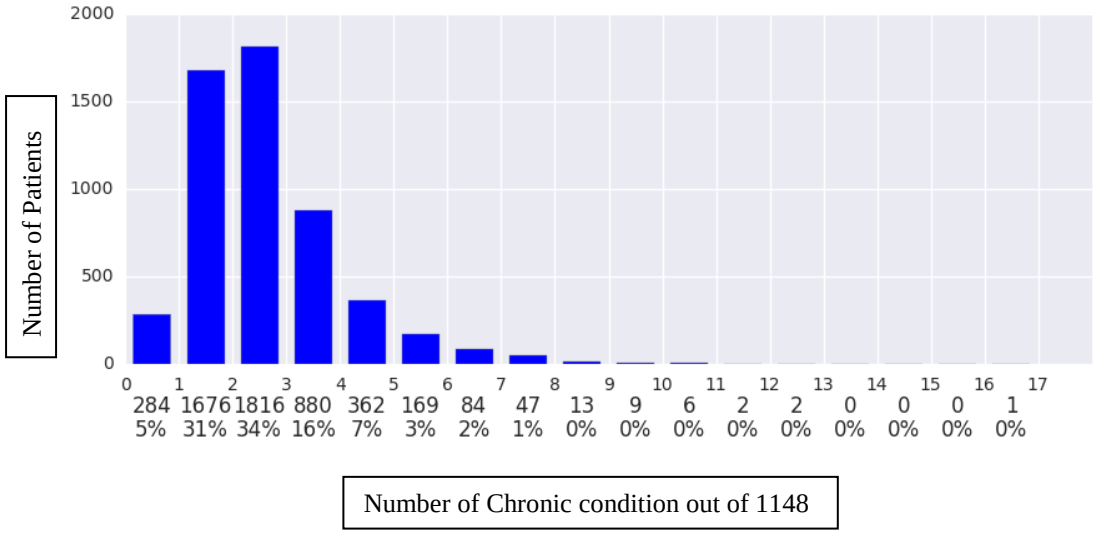

Figure 2: *Qualified patient distribution based on the frequency of chronic diseases from three different types: Chronic Disease (200 ICD-9 diagnosis codes), Cardiometabolic Chronic Disease(743 codes) and Chronic Episodic Disease (214). 284 out of 5351 patients (5%) had no chronic disease. Majority of patients (34%) had only two out of 1148 diagnosis codes.*

Unstructured Data Extraction from Longitudinal EHR Patient Data

We used physician encounter note data to extract mentions of FHFSS. The FHFSS were originally published in 1971 [15] and are often a focus of encounter documentation when physicians assess a patient's cardiovascular health. A

hybrid natural language processing (NLP) tool called PredMED [12] was used to identify mentions of FHFSS within the notes and to label notes according to whether, at the encounter level, each FHFSS was asserted or denied. PredMED achieves an F-score of 0.910 for mention extraction and an F-score of 0.932 for encounter labeling. Among the 17 FHFSS described in [15], 15 were deemed clinically relevant based on prior work [4, 12] and given the known frequency of documentation (Table 1). PredMED extracted mentions of these 15 FHFSS, along with modifiers that indicated the assertion (presence) or denial (absence) of each condition [except Tachycardia and WeightLoss, for which only assertion was extracted]. The resulting 28 FHFSS were used as features, for further analysis of PCP documentation.

PredMED was applied to all progress notes that were created in the 24-month observation window. Summary statistics were derived for each patient by counting of number of FHFSS mentions in the observation window. Moreover, the counts were normalized for the total timespan of all encounters and the total number of encounters within the observation window (Figure 1). This resulted in a feature vector for each patient containing the fraction of total encounters per year that contains a FHFSS mention. In a next step, this patient level information was summarized into the physician level information. Specifically, the averages of all these FHFSS fractions of all the patients belonging to a particular physician was used to build the final feature vector for each physician.

Statistical Analysis

For analysis, each physician is represented as a feature vector of 28 FHFSS. We used unsupervised clustering to determine if there were natural groupings of the physicians. Clustering algorithms use a distance metric for computing the similarity between two sample points. For this analysis, we used Euclidean distance as the distance metric, since the features are represented as fractions of total encounters and normalized by the span of duration of care.

We explored several agglomerative hierarchical clustering analysis (HCA) techniques [16] to determine if natural groupings were identifiable. These techniques are easier to interpret than other clustering approaches such as agglomerative and density based clusterings [16]. For example, the hierarchy generated by HCA provides relationships among different samples and the obtained physician groups, which will be useful to characterize the physician groups. Among various versions of HCA techniques, we used the Ward based algorithm due to its inclination of finding globular clusters similar to K-Means, while preserving the hierarchy of the obtained clusters [16].

Next, we compared the physician groups against each other to characterize the different documentation behaviors of each cluster. We used several descriptive statistics and visualizations to facilitate interpretation. For example, principal component analysis (PCA) [17] was performed to visualize the identified clusters. Also, the mean documentation behaviors of each FHFSS was compared among the clusters using statistical t-tests.

Table 1: *28 Framingham heart failure signs and symptoms (FHFSS) extracted from text notes using the PredMED text analysis tool. The mean, median and standard deviation of the fraction of encounters where each of these 28 features was reported by PCPs is shown.*

FHFSS Description	FHFSS Code	Assertion or Denial	Mean	Median	Std
Bilateral ankle edema	AnkleEdema (ANKED)	Assertion	0.157	0.120	0.142
Bilateral ankle edema	AnkleEdema (ANKED)	Denial	0.534	0.565	0.176
Acute pulmonary edema	APEdema (APED)	Assertion	0.006	0.000	0.032
Acute pulmonary edema	APEdema (APED)	Denial	0.006	0.000	0.032
Dyspnea on ordinary exertion	DOExertion (DOE)	Assertion	0.104	0.100	0.090
Dyspnea on ordinary exertion	DOExertion (DOE)	Denial	0.402	0.385	0.198
Hepatomegaly	Hepatomegaly (HEP)	Assertion	0.005	0.000	0.022
Hepatomegaly	Hepatomegaly (HEP)	Denial	0.365	0.330	0.208
Hepatojugular reflux	HJReflux (HJR)	Assertion	0.000	0.000	0.005
Hepatojugular reflux	HJReflux (HJR)	Denial	0.064	0.050	0.074
Central venous pressure > 16 cm H2O	ICV Pressure (ICV)	Assertion	0.000	0.000	0.000
Central venous pressure > 16 cm H2O	ICV Pressure (ICV)	Denial	0.001	0.000	0.012
Neck vein distention	JVDistension (JVD)	Assertion	0.003	0.000	0.019

Neck vein distention	JVDistension (JVD)	Denial	0.239	0.180	0.193
Nocturnal cough	NightCough (NC)	Assertion	0.012	0.000	0.033
Nocturnal cough	NightCough (NC)	Denial	0.204	0.140	0.165
Pleural effusion	PleuralEffusion (PLE)	Assertion	0.006	0.000	0.024
Pleural effusion	PleuralEffusion (PLE)	Denial	0.086	0.070	0.088
Paroxysmal nocturnal dyspnea	PNDyspnea (PND)	Assertion	0.028	0.000	0.054
Paroxysmal nocturnal dyspnea	PNDyspnea (PND)	Denial	0.177	0.140	0.141
Rales	Rales (RALE)	Assertion	0.049	0.000	0.079
Rales	Rales (RALE)	Denial	0.662	0.670	0.174
Radiographic cardiomegaly	RCardiomegaly (RC)	Assertion	0.005	0.000	0.034
Radiographic cardiomegaly	RCardiomegaly (RC)	Denial	0.104	0.065	0.152
S3 gallop	S3Gallop (S3G)	Assertion	0.007	0.000	0.025
S3 gallop	S3Gallop (S3G)	Denial	0.455	0.455	0.256
Tachycardia (rate of ≥120 min−1)	Tachycardia (TACH)	Assertion	0.055	0.025	0.080
Weight loss of 4.5 kg in 5 days, in response to HF treatment	WeightLoss	Assertion	0.013	0.000	0.054

Results

Summary statistics were generated for each of the FHFSS. Table 1 contains the mean, median and standard deviation of the fraction of encounters where each of these 28 features was reported by PCPs. Overall, there are many more encounters with denials of the signs and symptoms than with assertions. A few of the signs and symptoms (e.g., AnkelEdema, DOExertion, Rales-denial, S3Gallop-denial) are relatively frequent. Several signs and symptoms (e.g., ICV Pressure, APEdema, WeightLoss) are very rare.

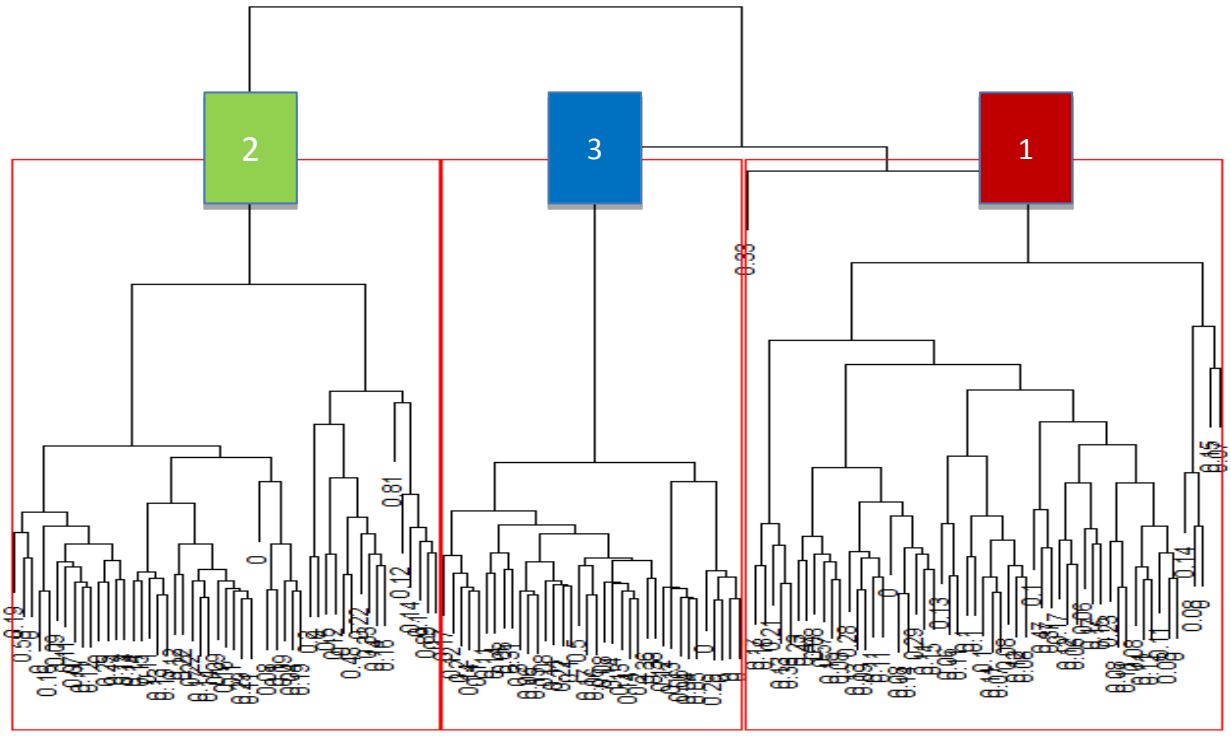

Figure 3: *The dendrogram obtained from HAC with the partitioning for K=3 clusters based on Ward algorithm and Eucledian distance metric.*

Table 2: *Descriptive statistics for the three PCP clusters*

Characteristics	Group 1	Group 2	Group 3	p value (Group 1 vs 2)	p value (Group 2 vs 3)	p value (Group 1 vs 3)
Number of PCPs	63	61	20			
Total Number of Patients	2860	1882	445			
Avg. Number of Patients per PCP	45	31	22			
Avg. Coverage Days	577	579	571	0.77	0.38	0.47
Avg. Age (Years)	70	70	69	0.99	*0.01*	*0.01*
Female Gender (%)	51	51	57			
Avg. Count of Chronic Disease	0.2	0.2	0.2	0.30	0.31	0.68
Avg. Count of Cardio-metabolic Chronic Disease	1.1	1.1	1.0	0.78	0.26	0.18
Avg. Count of Chronic Episodic Disease	1.0	1.0	1.0	0.08	0.15	0.68

Some basic descriptive statistics regarding the patients and physicians assigned to each of these three clusters are given in Table 2. Cluster 1 and cluster 2 were larger than cluster 3 in terms of number of patients and PCPs, but no significant differences were observed among the clusters in terms of coverage days, average age, gender, and comorbid chronic diseases. In this analysis, we used three categories of chronic diseases: "Chronic Disease", "Cardiometabolic Chronic Disease" and "Chronic Episodic Disease".

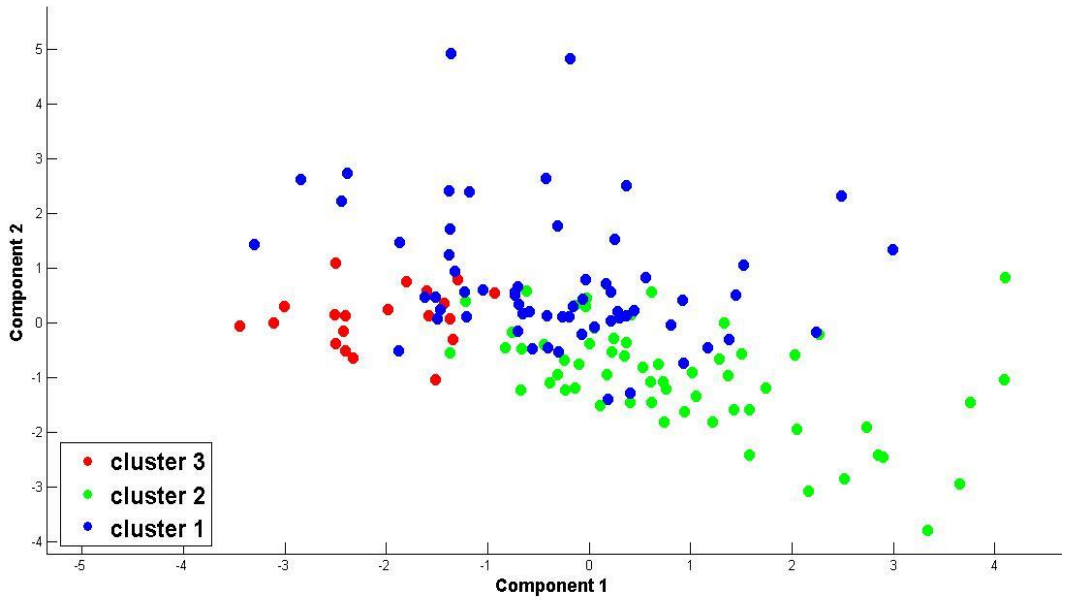

Figure 4*: Scatter plot of the first and second PCA components show reasonable separation of the three PCP clusters.*

Clustering Results as Groupings of Physician's Behavior:

Based on HCA analyses, the PCPs were clustered into 3 groups with distinct documentation behaviors. We explored several values for the possible number of clusters (k), however k=3 produced more natural groupings obtained from the HCA dendrogram (Figure 3).

Analyzing the Behaviors of Physicians in Three Clusters:

We performed a principal component analysis (PCA) of the FHFSS features to assess the discriminatory power of the three clusters as shown in Figure 4. The first and second PCA components are represented by the horizontal and vertical axis, respectively. The three clusters have reasonable separation in the lower dimensional feature space.

We also analysed the average frequencies of FHFSS mentions in the three clusters. In Figure 5, we plot the contribution of each FHFSS to the each of the three clusters by taking the mean of count frequencies of all samples belonging to each cluster. Note that this figure is normalized by z-score for each FHFSS. All FHFSS except assertion of ICVPressure varied among the three clusters to some extent. Overall, cluster 1 has the highest FHFSS counts whereas cluster 3 has the lowest. Cluster 2 contains medium counts of FHFSS.

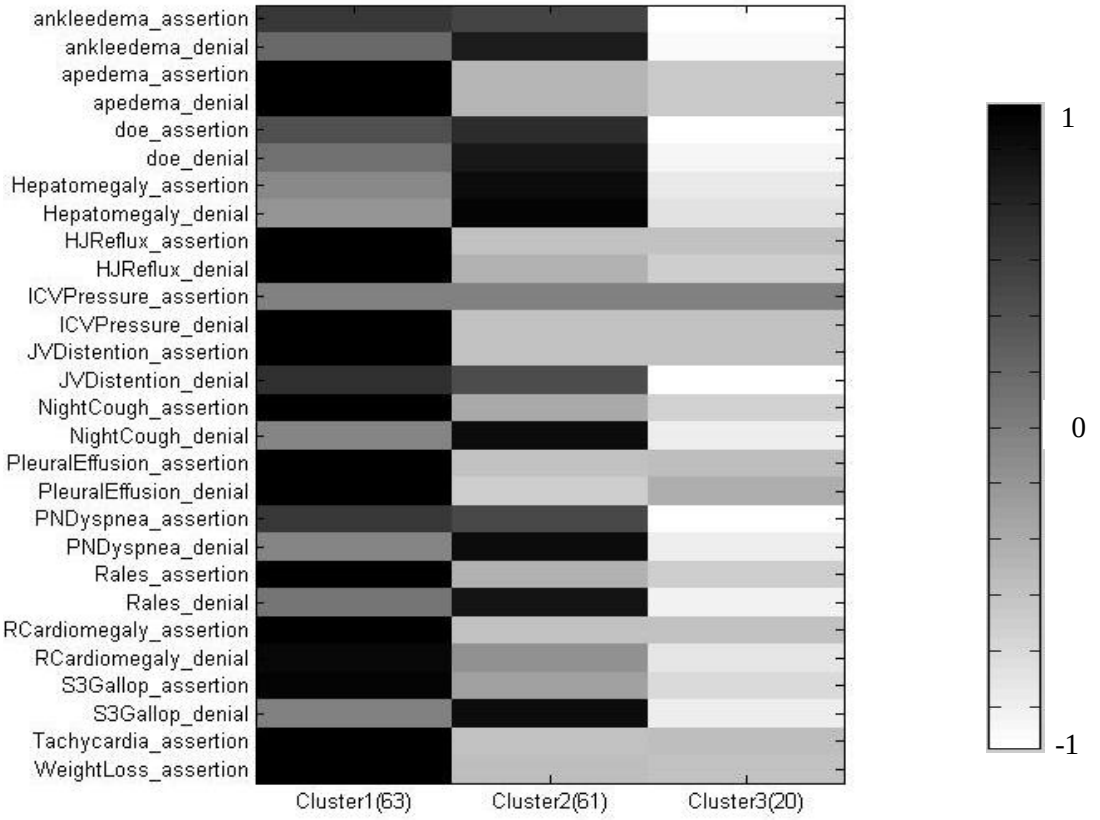

Figure 5. *Mean FHFSS counts of each of the 28 FHFSS symptoms for the three clusters. (Darker shade means higher counts of FHFSS in the corresponding cluster).*

Contrasting the Behaviors of Physicians among Three Clusters:

We also looked for the specific practice variations of the physicians of each cluster by comparing them with other clusters in terms of the FHFSS frequencies. Figure 6 contains the individual fraction of visits for each of the 28 FHFSS. In addition, the right three columns of each of the two panels contain the pairwise comparison of the three clusters to assess whether there is a significant difference between the fractions of visits of the two clusters under consideration based on t-statistics. Only the FHFSS with p-value < 0.05 are marked in the last three columns. Group 1 PCPs (n=63) documented 10 out of 15 assertions, and 11 out of 13 denials of FHFSS significantly more frequently than Group 3 (n=20); while Group 2 PCPs (n=61) have significantly more frequent denial documentation behaviors than the other two (see Figure 6)

Discussion and Conclusion

EHR data contains information about both patient's health characteristics such as the histopathological factors, demographics, treatment history and environmental effects as well as physician's behavior such as treatment plans and orders and documentation behaviors of patient's signs and symptoms. The availability of large-scale multi-source health data presents new opportunities and challenges for research that aims to effectively use these data to discover new knowledge to improve current health-care systems [18]. Such useful knowledge will not only help in personalizing healthcare for each patient with more accurate diagnoses, treatments and prevention plans, but also help reduce the unsustainable growth in healthcare cost.

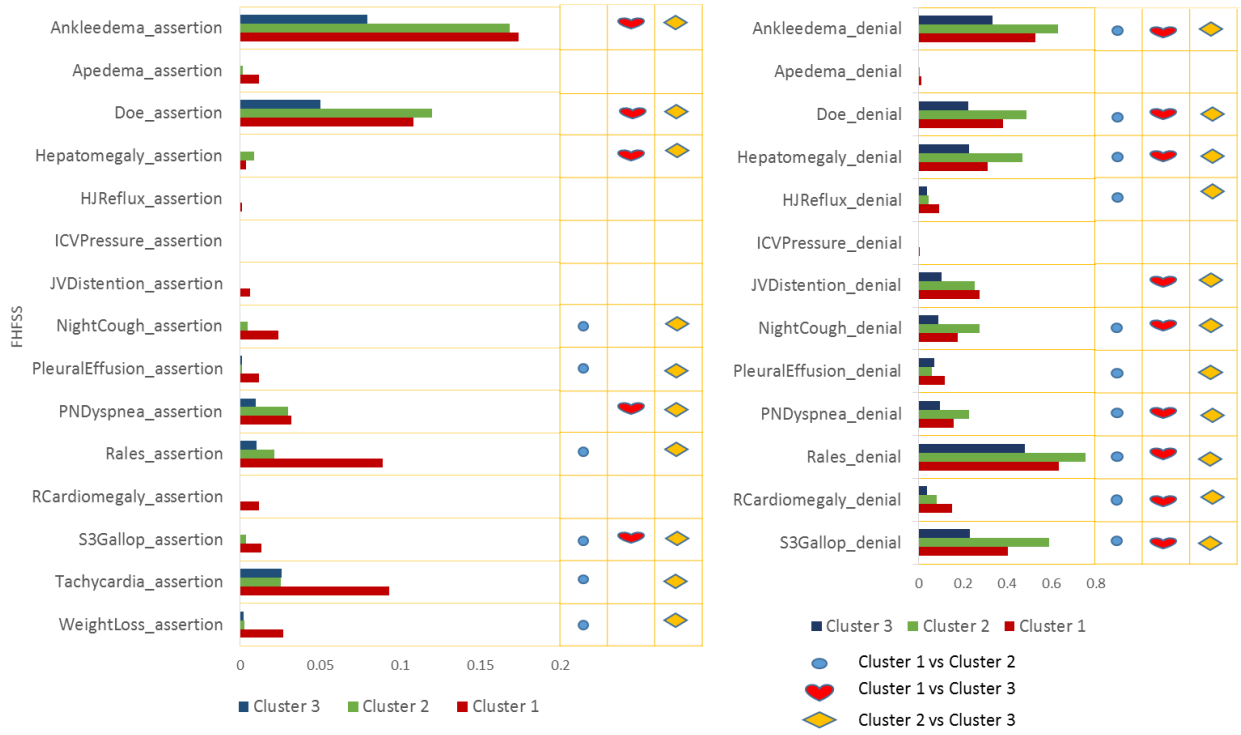

Figure 6: *Comparisons of mean documentation frequencies of FHFSS assertions and denials in 3 PCP groups clustered by HCA. The symbols indicate significant differences in pairwise comparisons by t-test ($p<0.05$). The horizontal axis shows the documentation frequency defined as the percentage of office visit encounters with FHFSS assertions/denials during the 2-year period. The vertical axis shows the assertion and denial FHFSS measure labels.*

Finding patterns for a particular disease requires the secondary analysis of the EHR data collected from multiple sources. The recent growth of machine learning and data mining techniques offer a great help in analyzing large-scale healthcare data with new possibilities of developing predictive modeling for early detection of disease. However, traditional data mining and machine learning techniques often cannot be applied directly to EHR data because they are collected retrospectively in time and therefore can contain a lot of underlying bias and noise factors completely unrelated to disease burden [19, 20]. Unlike prospective studies such as randomized control trials (RCTs) which are designed to avoid such sources of noise and experimental biases, EHR data requires more careful analysis strategies to remove the effect of such noise and biases [21].

This study investigated physician behaviors as a source of bias in EHR data and a potential source of confounders for predictive modeling. PCPs were characterized by their documentation profiles of FHFSS. Distinct groups were identified for each of the profiles using hierarchical clustering analysis. Significant differences among the physicians' practice in the three clusters were observed, which were associated with different documentation behaviors of FHFSS. Most (27 out of 28) of the FHFSS varied among these three clusters.

In terms of future work, we plan to investigate how to incorporate physician behaviors into predictive modeling and to quantify how much value (in terms of prediction performance) is added by eliminating this confounding factor in early detection of heart failure. Another interesting direction will be to further analyze the obtained groups of PCPs for finding their relationships with other potential causal factors such as physicians' expertise, their training, and the geographic variations of healthcare providers as well as any other characteristics of patients corresponding to each group of PCPs.

Acknowledgement

This study was supported by the National Institute of Health (NIH grant No. R01HL116832). We are grateful to Zahra Daar, Heather Law, Elise Blaese, Harry Stavropoulos and Satish Mudiganti for a variety of contributions to this work, including project coordination, data preparation and database management.

References

1. Charles, D., et al., *Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2012*. 2013: Office of the National Coordinator for Health Information Technology.
2. Dey, S., et al., *Predictive Models for Integrating Clinical and Genomic Data.* Healthcare Data Analytics, 2015. **36**: p. 433.
3. Amarasingham, R., et al., *An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data.* Medical care, 2010. **48**(11): p. 981-988.
4. Wang, Y., et al. *Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records*. in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. 2015. IEEE.
5. ICTs, O., *the Health Sector–Towards Smarter Health and Wellness Models*. 2013, OECD Publishing.
6. Weber, G.M. and I.S. Kohane, *Extracting physician group intelligence from electronic health records to support evidence based medicine.* PloS one, 2013. **8**(5): p. e64933.
7. Song, Y., et al., *Regional variations in diagnostic practices.* New England Journal of Medicine, 2010. **363**(1): p. 45-53.
8. Sirovich, B.E., et al., *The association between residency training and internists' ability to practice conservatively.* JAMA internal medicine, 2014. **174**(10): p. 1640-1648.
9. Davis, D., et al., *Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes?* Jama, 1999. **282**(9): p. 867-874.
10. Stewart, W.F., et al., *Patterns of health care utilization for low back pain.* Journal of pain research, 2015. **8**: p. 523.
11. King, M., J. Kingery, and M. BARETTA CASEY, *Diagnosis and evaluation of heart failure.* heart failure, 2012. **100**(21): p. 23.
12. Byrd, R.J., et al., *Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records.* International journal of medical informatics, 2014. **83**(12): p. 983-992.

13. Wu, J., J. Roy, and W.F. Stewart, *Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches.* Medical care, 2010. **48**(6): p. S106-S113.
14. Vijayakrishnan, R., et al., *Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record.* Journal of cardiac failure, 2014. **20**(7): p. 459-464.
15. McKee, P.A., et al., *The natural history of congestive heart failure: the Framingham study.* New England Journal of Medicine, 1971. **285**(26): p. 1441-1446.
16. Tan, P.-N., M. Steinbach, and V. Kumar, *Introduction to data mining*. Vol. 1. 2006: Pearson Addison Wesley Boston.
17. Jolliffe, I., *Principal component analysis*. 2002: Wiley Online Library.
18. Dey, S., et al., *Integration of clinical and genomic data: a methodological survey*. 2013, Technical Report, Department of Computer Science and Engineering University of Minnesota.
19. Dey, S., et al., *Mining Patterns Associated With Mobility Outcomes in Home Healthcare.* Nursing research, 2015. **64**(4): p. 235-245.
20. Yadav, P., et al., *Mining Electronic Health Records (EHR): A Survey.* Technical Report, 2015.
21. Austin, P.C., *An introduction to propensity score methods for reducing the effects of confounding in observational studies.* Multivariate behavioral research, 2011. **46**(3): p. 399-424.