Phenotyping Hypotensive Patients in Critical Care Using Hospital Discharge Summaries

Yang Dai, Sharukh Lokhandwala, William Long, Roger Mark, Li-wei H. Lehman†
Institute for Medical Engineering and Science,
Massachusetts Institute of Technology, Cambridge, MA

***Abstract*— Among critically-ill patients, hypotension represents a failure in compensatory mechanisms and may lead to organ hypoperfusion and failure. In this work, we adopt a data-driven approach for phenotype discovery and visualization of patient similarity and cohort structure in the intensive care unit (ICU). We used Hierarchical Dirichlet Process (HDP) as a non-parametric topic modeling technique to automatically learn a d-dimensional feature representation of patients that captures the latent "topic" structure of diseases, symptoms, medications, and findings documented in hospital discharge summaries. We then used the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to convert the d-dimensional latent structure learned from HDP into a matrix of pairwise similarities for visualizing patient similarity and cohort structure. Using discharge summaries of a large patient cohort from the MIMIC II database, we evaluated the clinical utility of the discovered topic structure in phenotyping critically-ill patients who experienced hypotensive episodes. Our results indicate that the approach is able to reveal clinically interpretable clustering structure within our cohort and may potentially provide valuable insights to better understand the association between disease phenotypes and outcomes.**

I. Introduction

Text-based electronic health records contain useful information about the disease progression and treatment plans of patients and provide an important resource to better understand the associations between complex disease processes and patient outcomes. Hospital discharge summaries, in particular, document detailed information about a patient's entire course of stay, including problems, treatments, symptoms, diagnoses, test results, procedures, as well as the condition of patients at discharge.

In this work, we propose a data-driven approach to automatically discover phenotypic patterns and visualization of patient cohort structure using clinical concepts from hospital discharge summaries of patients who underwent hypotensive episodes in an intensive care unit (ICU). Hypotension, defined as a decrease in either systolic or mean blood pressure, is a manifestation of circulatory failure, or shock. While there are numerous etiologies of shock, including distributive, cardiogenic, hypovolemic, and obstructive shock, all forms if left untreated may lead to organ failure and/or death [1]. Frequently, clinicians at the bedside are tasked with both determining the etiology of and treatment for hypotensive patients. Characterizing the clinical manifestation and patterns of care of shock in the intensive care units can potentially provide insights into new risk stratification and treatment strategies.

Corresponding Author: † Li-wei H. Lehman, Massachusetts Institute of Technology, 45 Carleton Street, Cambridge, MA 02142, USA. Email: lilehman@mit.edu.

We used Hierarchical Dirichlet Process (HDP) [2] as a topic modeling technique to model the latent structure of diseases, symptoms, and findings documented in hospital discharge summaries. Probabilistic topic models, such as HDP, are Bayesian modeling techniques for finding patterns and uncovering the hidden thematic structure in a collection of documents [3]. We represent the diseases, symptoms and findings extracted from patients' discharge summaries as an un-ordered set of Unified Medical Language System (UMLS) codes [4]. HDP was used to infer "topics" as a collection of co-occurring UMLS clinical concepts. The learned "topics" capture the shared latent structure in the diseases, symptoms, medication, and findings documented in patients' hospital discharge summaries. For visualization of the latent cohort structure learned from HDP, we used the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm [5], which learns a low-dimensional embedding that preserves the distances between data points in a higher-dimensional space.

Several recent works have studied the problem of computational phenotyping using a variety of data sources, including longitudinal clinical time series, high-resolution physiological signals, and other data types from electronic health records [6], [7], [8]. The focus of the current work is to investigate a data-driven approach for phenotype discovery from text-based electronic health records using probabilistic topic modeling. In our previous works [9], [10], we demonstrated that features extracted from clinical text can be used to better stratify mortality risks of patients in a critical care setting. In this work, we evaluate the clinical utility of the discovered latent structure in phenotyping hypotensive patients in a critical care setting using a large cohort of patients who underwent hypotensive episodes in ICUs.

II. Methodology

A. Data Preparation

Data for this study was obtained from the Multi-Parameter Intelligent Monitoring in Intensive Care (MIMIC-II) database [11] available from PhysioNet [12]. The database contains records from 24,581 ICU patients admitted to Bostons Beth Israel Deaconess Medical Center between 2001-2008. The patient cohort consists of adult patients from the MIMIC-II

database who experienced at least one hypotensive episode, defined as mean arterial pressure (MAP) less than 60 mmHg for at least two consecutive measurements (recorded every 5 to 60 minutes). For each patient, we extracted the discharge summary. A total of 9,889 patients were included in the study. The text of discharge summaries was parsed into an unordered set of UMLS codewords using natural language processing techniques. UMLS codewords labeled "Diseases", "Symptoms", "Findings", "Medications", and "Procedures" were used. Codewords that occur fewer than 5 times in the entire corpus and codewords designated as stopwords were eliminated because they added little information about the patient. Stopwords included terms such as "medication" or "Hospital admission" which appear in nearly all patients' discharge summaries and thus add little meaning. Stopwords were found using the term frequency-inverse document frequency (tf-idf) of the codewords. The 300 codewords with the lowest tf-idf, meaning that they add the least amount of information, were manually examined to generate the stopword list. Acronyms (e.g., "ED", "LAD") with ambiguous meanings, and were five characters long or less, were not converted to UMLS codes; instead we used the original text as the dictionary codeword.

B. Latent Topic Discovery Using HDP

HDP uses a non-parametric prior to enable mixture models to share components [2]. The number of topics is assumed to be unknown a priori, and is inferred from the data. A topic is a multinomial distribution over words from a finite, known vocabulary. The HDP models documents with multiple Dirichlet Processes (DP), one for each document, to enable document-specific mixing proportions. For HDP parameter settings, we used the same notations as in [2]. A two-level hierarchical Dirichlet process implementation was used to build our topic models. We used a symmetric Dirichlet distribution with parameters of 0.2 for the prior H over topic distributions. We used fixed concentration parameters 0.1 and 1 for γ and α respectively. Results presented are output of the model after 1000 iterations of Gibbs sampling. Running HDP results in a topic model containing an inferred number of discovered topics. Each topic is a distribution over words (UMLS codewords). Using the word to topic assignments, we constructed a topic proportion vector for each patient, which is a d-dimensional vector that contains the proportion of words belonging to each of the d topic. The topic proportion vectors were then aggregated into a matrix of size Nxd, where N is the number of patients and d is the number of topics.

C. Evaluation and Statistical Methods

For visualization of the latent cohort structure learned from HDP, we used t-SNE to convert the d-dimensional latent structure into a two-dimensional embedding, where d is the number of topics inferred by HDP from hospital discharge summaries of the entire patient cohort. Specifically, we used t-SNE to project the Nxd multi-dimensional topic proportion matrix into a two-dimensional space. Rare topics were pruned, and the remaining topics were used in constructing the t-SNE embedding. Next, we perform K-means clustering on the patients using the output of t-SNE to find clusters of "similar" patients. The most optimal value of K, the number of clusters, was determined using silhouette evaluation [13].

Clinical relevance of the discovered topics was assessed both qualitatively based on clinician review and ICD-9 codes as well as quantitatively using one-year post hospital discharge mortality as an outcome measure. We report characteristics of each of the K clusters, including the most common topic for each cluster, as determined by the sum of topic proportions for each patient in the cluster, and the most frequently occurring ICD-9 (International Classification of Disease) code. ICD-9 codes are standardized specifications of the patient's condition used to bill the insurance company; we used the primary ICD-9 code for each patient, which represents the main cause of hospitalization.

III. RESULTS

Running HDP on the parsed discharge summaries resulted in 47 topics. We represented each of the 9868 hypotensive patients as a d-dimensional topic proportion feature vector, where d is the top most common topics after pruning away rare topics with less than 200 UMLS codewords assigned, resulting in 27 most common topics. We used t-SNE to embed the Nxd (N = 9868, d=27) matrix in a two-dimension space, and performed K-means clustering on the resulting Nx2 dimensional matrix to form K patient clusters (K=26 as determined by the silhouette function).

The cluster size ranges from approximately 200 to 600 patients, hospital mortality rate from 0.3% to 81%, and one-year mortality from 3.4% to 38%. Patients with high in-hospital mortality were grouped together; the highest mortality group had a 81% one-year mortality rate whereas the maximum mortality rate among the other clusters was 28.8%.

Figure 1 shows the two-dimensional t-SNE embedding of the 9,868 adult hypotensive patients as well as the cluster assignment from K-means. Each point represents a patient; distance between the points represents how similar the patients are to one another as measured by their topic proportion. The color reflects the cluster assignment. For readability and clarity, we annotated only a representative sub-set of clusters in Figure 1. For each annotated cluster, we show its cluster size, in-hospital mortality and one-year post hospital discharge mortality of patients in the cluster; the most prevalent primary ICD-9 code, and the top five UMLS descriptions from the most dominant topic for patients within that cluster. Table I lists the patient characteristics and the dominant topic for the remaining clusters. Noted acronyms in Figure1 and Table I include: LAD - left anterior descending; CABG - Aortocoronary bypass for heart revascularization; CHF - congestive heart failure; A-Fib - atrial Fibrillation; ED - Emergency Department.

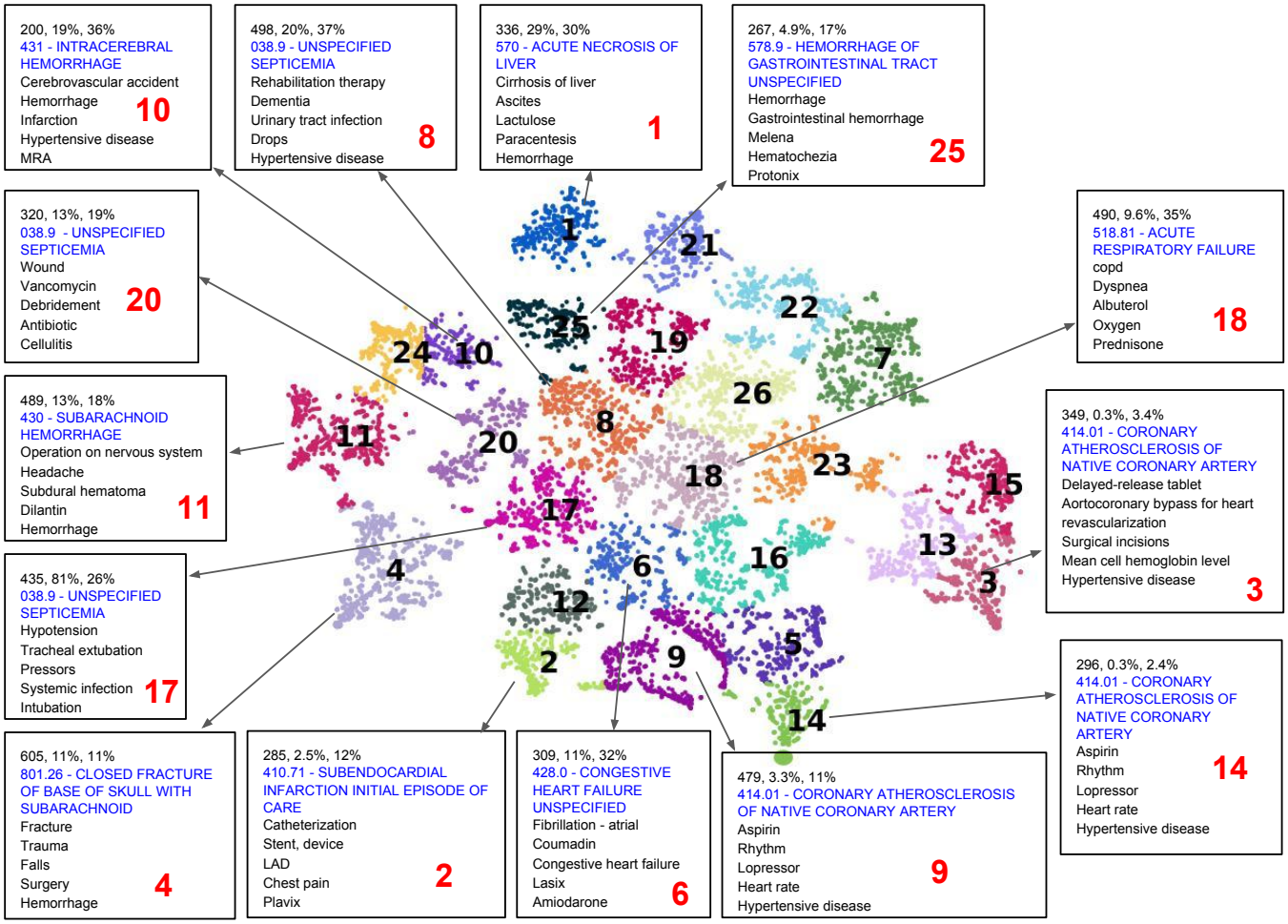

Fig. 1. Clustering of hypotensive patients based on latent topic structure inferred from patients' discharge summaries. A selected subset of clusters were annotated in the Figure, noting its cluster size, in-hospital mortality, one-year post hospital discharge mortality, the most prevalent primary ICD-9 code (in blue text), and the top five UMLS descriptions of the most dominant topic for patients within each cluster.

IV. Discussion

Our approach was able to cluster patients into clinically relevant subgroups. In our cohort of hypotensive, critically-ill patients, our methodology discovered distinct clusters of patients with coronary artery disease, congestive heart failure, gastrointestinal hemorrhage, trauma, cerebrovascular accidents, sepsis and respiratory failure.

While this is an exploratory analysis, our findings raise numerous interesting questions. Among patients in whom the most frequent UMLS codes related to cardiovascular disease (clusters 2, 3, 5, 6, 9, 13, 14, 15, 16) those with the highest mortality were those who more commonly have atrial fibrillation and congestive heart failure (cluster 6, 1-year mortality = 32%). This agrees with the existing body of literature regarding congestive heart failure [14].

In addition, our clustering technique placed patients with liver disease (cluster 1) in close proximity to patients with gastrointestinal hemorrhage (cluster 25), implying that perhaps a large proportion of patients with liver disease who subsequently become hypotensive may potentially have pathophysiological mechanisms related to blood loss. With regards to patients with cerebrovascular accidents and/or intracerebral hemorrhage, there exists a fair amount of dissonance regarding the differences between our clusters (clusters 10, 11, 24). Specifically, one wonders whether there exists specific reasons for the differing in-hospital and one-year mortality rates. Further work will be necessary to determine these patient level differences. Our clustering technique identified three unique clusters (8, 17, 20) related to sepsis. Interestingly, the in-hospital (20%, 81%, and 13%) and one-year (37%, 26%, and 19%) mortality varies significantly between the three clusters. Finally, we note that t-SNE aims to preserve small distances and local structure at the expense of large ones; as such, one should be cautioned against drawing conclusions based on the relative positions of the clusters at the macro scale.

V. Conclusion

In this paper, we used Hierarchical Dirichlet Process mixture models to discover latent topic structure to reveal phenotypic patterns of diseases and symptoms shared across subgroups of a hypotensive patient cohort. We used t-SNE to automatically construct a low-dimensional embedding

TABLE I
PATIENT CHARACTERISTICS OF SELECTED CLUSTERS

Cluster	N	Hosp. Mort.	1-year Mort.	Topic	UMLS Description	ICD-9
5	369	1.4%	6.5%	42	Aspirin,Rhythm,Lopressor,Heart rate,Hypertension	414.01 - CORONARY ATHEROSCLEROSIS OF NATIVE CORONARY ARTERY
7	539	2.8%	19.3%	35	Pain,Surgery,Drainage,Surgical incisions,Wound	151.0 - MALIGNANT NEOPLASM OF CARDIA
12	278	21.2%	28.4%	32	Catheterization,Stent,LAD,Chest pain,Plavix	410.71 - SUBENDOCARDIAL INFARCTION INITIAL EPISODE OF CARE
13	363	1.7%	8.0%	12	Delayed-release tablet,CABG,Surgical Incision, Mean cell hemoglobin level,hypertension	414.01 - CORONARY ATHEROSCLEROSIS OF NATIVE CORONARY
15	330	2.7%	8.2%	12	Delayed-release tablet,CABG,Surgical Incision, Mean cell hemoglobin level,hypertension	414.01 - CORONARY ATHEROSCLEROSIS OF NATIVE CORONARY
16	438	12.3%	20.8%	42	Aspirin,Rhythm,Lopressor,Heart rate,Hypertension	414.01 - CORONARY ATHEROSCLEROSIS OF NATIVE CORONARY
19	394	8.1%	18.8%	28	Mean cell hemoglobin level,ED,Glucose, Hypertension,Delayed-release tablet	042 - HUMAN IMMUNODEFICIENCY VIRUS (HIV) DISEASE
21	316	4.8%	8.9%	10	Seizure,Ativan,Dilantin,Depression,Alcohol abuse	969.4 - POISONING BY BENZODIAZEPINE-BASED TRANQUILIZERS
22	385	8.1%	16.6%	44	Abdominal pain,Pancreatitis,Lipase, Amylase preparation,Flagy	577.0 - ACUTE PANCREATITIS
23	418	6.5%	30.9%	45	Rehab,Vancomycin,Tube feeding,Aspiration,Tube	441.01 - DISSECTION OF AORTA THORACIC
24	266	27.8%	18.8%	41	Cerebrovascular accident,Hemorrhage, Infarction,Hypertension,MRA	431 - INTRACEREBRAL HEMORRHAGE
26	414	19.8%	37.7%	23	Neoplasms-malignant,Neoplasms,Chemotherapy, Neoplasm,secondary,Lesion	518.81 - ACUTE RESPIRATORY FAILURE

of patients based on similarity defined in the learned d-dimensional latent structure. Our results indicate that our approach is able to reveal clinically interpretable local and global latent structure within our cohort and may potentially be used to generate insights in the association between disease phenotypes and outcomes. As part of our on-going and future work, we plan to incorporate other data types, e.g. clinical time series [15], in the HDP framework for disease phenotyping.

ACKNOWLEDGMENT

This work was supported by the National Institutes of Health (NIH) grants R01-EB017205, R01-EB001659 and R01GM104987. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIBIB or the NIH.

REFERENCES

[1] Vincent J, De Backer D, Circulatory Shock. N Engl J Med, October 2013.
[2] Teh Y, Jordan M, Beal M J., Blei D, Hierarchical Dirichlet Processes, Journal of the American Statistical Association, 101, 1566-1581, 2006.
[3] Blei D, Carin L, Dunson D. Probabilistic Topic Models, IEEE Signal Processing Magazine, Nov. 2010, 5565.
[4] Long W. Extracting Diagnoses from Discharge Summaries, AMIA 2005 Symposium Proceedings, 2005, 470474.
[5] Maaten Laurens van der, Hinton Geoffrey, Visualizing Data using t-SNE. Journal of Machine Learning Research, 9 (2008), 2579-2605.
[6] Lasko T, Denny J, Levy M: Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data, PLOS ONE, 2013.
[7] Lehman LH, Johnson M, Nemati S, Adams RP, Mark RG, Bayesian nonparametric learning of switching dynamics in cohort physiological time series: application in critical care patient monitoring, Chapter 11 in Advanced State Space Methods for Neural and Clinical Data, ed. by Chen Z., Cambridge University Press, 2015, 257-282.
[8] Che Z, Kale D, Li W, Bahadori MT, and Liu Y, Deep Computational Phenotyping, Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD), 2015.
[9] Lehman LH, Saeed M, Long W, Lee J, Mark RG, Risk stratification of ICU patients using topic models inferred from unstructured progress notes. Proceedings of the AMIA Annual Symposium, 505-511, Nov. 2012.
[10] Lehman LH, Long W, Saeed M, Mark RG, Latent Topic Discovery of Clinical Concepts from Hospital Discharge Summaries of a Heterogeneous Patient Cohort. Proceedings of the 36th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Chicago, August 2014.
[11] Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LH, Moody G, Heldt T, Kyaw TH, Moody B, and Mark RG, Multiparameter intelligent monitoring in intensive care (MIMIC II): a public access intensive care unit database. CritCare Med, vol. 39, no. 5, pp. 952960, May 2011.
[12] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, and Stanley HE, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation, vol. 101,no. 23, pp. e215e220, 2000.
[13] Ng AY, Jordan MI, Weiss Y, et al. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems 2002;2:849856.
[14] Solomon SD, Dobson J, Pocock S, et al. Influence of nonfatal hospitalization for heart failure on subsequent mortality in patients with chronic heart failure. Circulation. 2007;116(13):1482-7.
[15] Zalewski A, Long W, Johnson AE, Mark RG, Lehman LH, Patient Risk Stratification Using Latent Topics Inferred from Clinical Time Series and Text, Proceedings of IEEE International Conference on Biomedical and Health Informatics (BHI), February 2017, Orlando FL.