Latent Topic Discovery of Clinical Concepts from Hospital Discharge Summaries of A Heterogeneous Patient Cohort

Li-wei Lehman[1*], PhD, William Long[2], PhD, Mohammed Saeed[3], MD, PhD, Roger Mark[1], MD, PhD

***Abstract*— Patients in critical care often exhibit complex disease patterns. A fundamental challenge in clinical research is to identify clinical features that may be characteristic of adverse patient outcomes. In this work, we propose a data-driven approach for phenotype discovery of patients in critical care. We used Hierarchical Dirichlet Process (HDP) as a non-parametric topic modeling technique to automatically discover the latent "topic" structure of diseases, symptoms, and findings documented in hospital discharge summaries. We show that the latent topic structure can be used to reveal phenotypic patterns of diseases and symptoms shared across subgroups of a patient cohort, and may contain prognostic value in stratifying patients' post hospital discharge mortality risks. Using discharge summaries of a large patient cohort from the MIMIC II database, we evaluate the clinical utility of the discovered topic structure in identifying patients who are at high risk of mortality within one year post hospital discharge. We demonstrate that the learned topic structure has statistically significant associations with mortality post hospital discharge, and may provide valuable insights in defining new feature sets for predicting patient outcomes.**

I. INTRODUCTION

Large-scale clinical databases and electronic health records provide an opportunity to better understand the associations between complex disease processes and patient outcomes. In this work, we propose a data-driven approach to automatically discover prognostic phenotypic patterns of clinical concepts from discharge summaries of a large, heterogeneous patient cohort. We used Hierarchical Dirichlet Process (HDP) [1] as a topic modeling technique to model the latent structure of diseases, symptoms, and findings documented in hospital discharge summaries.

Probabilistic topic models are Bayesian modeling techniques for finding patterns and uncovering the hidden thematic structure in a collection of documents [2], [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, and used HDP to infer "topics" as a collection of co-occurring UMLS clinical concepts.

Using discharge summaries from the MIMIC II database [4], we demonstrate that our approach automatically extracts disease phenotypes as subgroups of clinically related pathologies and symptoms that tend to co-occur. We evaluate the utility of the discovered topic structure in identifying patients who are at high risk of mortality within one year post hospital discharge.

In our previous work [5], we combined the learned "topic" structure of UMLS clinical concepts extracted from the first 24-hour ICU nursing notes with physiologic data (from SAPS I) for risk stratification of in-hospital mortality. Our current work focuses on discovering clinical topics in discharge summaries that are predictive of post hospital discharge mortality.

II. METHODOLOGY

In topic models, documents are represented by un-ordered sets of words. A *topic* is thus a set of words that tend to co-occur, and represents word co-occurrence patterns that are shared across multiple documents in the corpus[2], [3]. To apply topic modeling in unstructured clinical text, we represent patients using un-ordered sets of UMLS codes extracted from their hospital discharge summaries. Hospital discharge summary of each patient is modeled as a separate "document". A "word" in our case is thus a UMLS code, representing either a disease, symptom, or finding from the patient's discharge summary. Topics are sets of UMLS codes that tend to co-occur in a collection of discharge summaries.

A. Data Preparation

UMLS clinical concepts were extracted from discharge summaries using a previously described [6] natural language processing (NLP) technique. We used three types of UMLS codes in our model: Diseases, Symptoms, and Findings. UMLS codes that occur less than 5 times in the entire corpus were eliminated; some UMLS code words (findings) that were considered stop words (e.g., concept such as "Past Medical History") were also removed. For patients with multiple hospital admissions, the discharge summary from the first hospital visit for each patient was included.

Hospital discharge summaries for 21,053 patients from the MIMIC II database were included to fit the model. The number of unique UMLS terms in this corpus was 9,152. The total number of UMLS terms across the corpus was 1,332,141 (average 63 UMLS terms per discharge summary). Of the 21,053 patients, 17,948 were adults; neonates were excluded from the mortality analysis in this study. The hospital mortality rate of the adult patients was 11%. One-year post hospital discharge mortality was 16%. Among the 17,948 adult patients, 15,962 patients survived beyond hospital discharge and 15,310 survived beyond 28 days post

* Corresponding Author: lilehman@mit.edu. [1]L. H. Lehman and R. G. Mark are with the Institute for Medical Engineering & Science, Massachusetts Institute of Technology, 45 Carleton Street, Cambridge, MA 02142, USA. [2]William Long is with the Computer Science and Artificial Intelligence Laboratory, MIT. [3]Mohammed Saeed is with University of Michigan, Ann Arbor. Manuscript received April 7th, 2014.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

discharge. Performance for one year mortality prediction was based on 14,203 adult patients who survived beyond hospital discharge and who also had SAPS I and comorbidity variables. Performance for one year mortality excluding patients who died within 28 days from discharge, was based on 13,612 patients who survived beyond 28 days after hospital discharge and who also had SAPS I and comorbidity variables.

B. Topic Modeling with HDP

HDP uses a non-parametric prior to enable mixture models to share components [1], [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 [1]. A two-level hierarchical Dirichlet process implementation [7] 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.

C. Evaluation and Statistical Methods

Clinical relevance of the discovered topics was assessed both qualitatively based on clinician annotations and ICD-9 codes and quantitatively using one-year post hospital discharge mortality as outcome measures.

We evaluated the clinical interpretability of the topics that had at least 5,000 words in their word frequency counts. These topics were reviewed by a clinician, and each topic was assigned a clinical category that best described the "topic" based on a review of the top 10 words in each topic. For each topic, we report the top five most common ICD9 codes for patients with at least 10% of the topic of interest.

To evaluate the utility of the discovered topic structure in identifying patients who are at high risk of post discharge mortality, we used the inferred topic proportion of each discharge summary (defined as the proportion of words assigned to each inferred topic) as input to logistic regression for post hospital discharge mortality prediction. Forward search was used for feature selection. We report the median AUC (with interquartile range) from 10-fold cross-validations. SAPS I [8] was used as a baseline for comparison.

Uni-variate logistic regression was performed to find the association between each topic (proportion) variable and one-year mortality. For each topic variable, we computed its p-value and odds ratio (OR, with 95% confidence interval). Odds ratios for post hospital discharge mortality are defined per 10% increase in topic proportions. Multivariate logistic regression was performed to find the association between each topic variable and one-year mortality, after adjusting for SAPS I and the 30 comorbidity variables.

III. RESULTS

A. Clinical Interpretation of the Topics

The model used to present our main results in this paper contained a total of 44 topics; the topics from this model were typical across multiple runs of the algorithm. We evaluated the clinical interpretability of the 28 topics that had at least 5,000 words in their word frequency counts.[1] These 28 topics were reviewed by a clinician, and each topic was assigned a clinical category that best described the "topic" based on a review of the top 10 words in each topic.

The clinical interpretation corresponding to the 28 topics were: sepsis, cardiovascular diseases, acute coronary syndromes, pulmonary disorders, GI bleeding, stroke and head trauma, hemorrhage and bleeding, chronic heart and diabetes problems, congenital heart disease, and trauma, ob/gyn disease findings, GI/liver disorders, cancer, heart/abdomen/pulmonary physical exams, cardiovascular/pulmonary disorders, mental health disorders, skin lesions and infections, lung and mediastinal disorders, upper airway and oropharyngeal disorders, stroke, cardiovascular diseases, hematological malignancies, physical exam findings, disorders of spines, thrombo-embolic disease, causes or associated findings with dementia/delirium, seizures, and infection.

B. Evaluating the Predictive Value of Topics

In this section, we evaluate the utility of the discovered topic structure in identifying patients who are at high risk of mortality within one year post hospital discharge. Topic proportions alone achieved a median AUC of 0.76 (0.75, 0.77) in predicting one year post hospital discharge mortality. After removing patients who died within 28-days after hospital discharge, topic proportions achieved a median AUC of 0.75 (0.73, 0.75). SAPS I alone achieved a median AUC of 0.60 (0.59, 0.63).

C. Discovering Predictive Topics for One-Year Mortality Post Hospital Discharge

Univariate and multivariate logistic regressions were performed to find the association between each of the topic variables and one-year post hospital discharge mortality (excluding patients who died within 28 days from hospital discharge). Our results based on uni-variate logistic regression indicate significant associations (p values ≤ 0.05) between 17 learned topics and one-year post hospital discharge mortality.

In particular, ten "high-risk" topics, corresponding to "sepsis", "cancer", "pulmonary disorders", "cardiovascular disease", "dementia/delirium", "Chronic heart and diabetes problems", "GI bleed", and "GI problems", were significantly associated with increased post discharge mortality (with odds ratios greater than one), indicating that increasing proportions of these topics were associated with an increased chance of one-year mortality. Table I shows a selected subset of these high-risk topics.

[1] Sixteen topics with word frequency less than 5,000 words were not considered, as they represent topics exhibited in only a small subset of patients.

TABLE I

HIGH-RISK TOPICS ASSOCIATED WITH ONE YEAR MORTALITY.

Topic Label by Clinician	Top UMLS Descriptions	Odds Ratio (95% CI)	N, Age	ICD-9	
Sepsis	Hypotension Systemic infection Respiratory failure Pneumonia Acute renal failure Sedated state	7.38 (4.91 11.10)	2591 70 (55, 80)	518.81 038.9 428.0 584.9 507.0	RESPIRATORY FAILURE SEPTICEMIA NOS CONGESTIVE HEART FAILURE ACUTE RENAL FAILURE NOS FOOD/VOMIT PNEUMONITIS
Cancer related	Secondaries Lesion Neoplasms Pain Nodule Breast cancer	27.64 (17.96 42.53)	1292 65 (55, 76)	518.81 197.7 198.5 197.0 162.8	RESPIRATORY FAILURE SECOND MALIG NEO LIVER SECONDARY MALIG NEO BONE SECONDARY MALIG NEO LUNG MAL NEO BRONCH/LUNG NEC
Pulmonary disorders	Shortness of breath Pneumonia Hypoxia Hypertensive disease Congestive heart failure Fever	3.47 (2.60 4.63)	2875 68 (55, 79)	518.81 428.0 486 507.0 491.21	RESPIRATORY FAILURE CONGESTIVE HEART FAILURE PNEUMONIA, ORGANISM NOS FOOD/VOMIT PNEUMONITIS OBSTRUCTIVE CHRONIC BRON
Cardiovascular disease	Fibrillation - atrial Congestive heart failure Coronary artery disease Shortness of breath Mitral valve insufficiency Hypertensive disease	4.38 (3.30 5.80)	3284 76 (66, 84)	428.0 427.31 414.01 410.71 518.81	CONGESTIVE HEART FAILURE ATRIAL FIBRILLATION CORON ATHEROSCLER NATIVE AMI, SUBENDOCARD INFARCT RESPIRATORY FAILURE
Causes or associated findings with dementia/ delirium	Urinary tract infection Dementia Hypertensive disease Fever Confusion Pneumonia	21.38 (15.06 30.36)	1892 80 (69, 86)	518.81 507.0 428.0 599.0 038.9	RESPIRATORY FAILURE FOOD/VOMIT PNEUMONITIS CONGESTIVE HEART FAILURE URIN TRACT INFECTION NOS SEPTICEMIA NOS
Hematological malignancies, Cancer	Fever Acute myelocytic leukemia Lymphoma Thrombocytopenia Multiple myeloma Febrile neutropenia	26.29 (11.76 58.78)	451 60 (48, 72)	205.00 518.81 486 584.9 996.85	ACUTE MYELOID LEUK W/O R RESPIRATORY FAILURE PNEUMONIA, ORGANISM NOS ACUTE RENAL FAILURE NOS COMPLIC BONE MARROW TRAN
Pulmonary, heart problems	Pericardial effusion Pleural effusion Effusion Atelectasis Mitral valve insufficiency Bilateral pleural effusion	7.47 (4.46 12.52)	1464 68 (54, 79)	428.0 518.81 427.31 038.9 584.9	CONGESTIVE HEART FAILURE RESPIRATORY FAILURE ATRIAL FIBRILLATION SEPTICEMIA NOS ACUTE RENAL FAILURE NOS
GI bleed	Hemorrhage Gastrointestinal hemorrhage Cirrhosis of liver Melena Ascites Hematochezia	2.04 (1.50 2.78)	1925 64 (51, 77)	285.1 428.0 571.2 578.9 584.9	AC POSTHEMORRHAG ANEMIA CONGESTIVE HEART FAILURE ALCOHOL CIRRHOSIS LIVER GASTROINTEST HEMORR NOS ACUTE RENAL FAILURE NOS

Seven "low-risk' topics, corresponding to "Heart physical exams", "Trauma", "Acute coronary syndromes", "Mental health disorders", "Stroke and Head Trauma", and "ob/gyn disease findings", were significantly associated with decreased one-year mortality (with odds ratios less than one), indicating that increasing proportions of these topics were associated with a decreased chance of one-year mortality. Although the "Stroke and Head Trauma" topic was associated with a high in-hospital mortality rate, it was associated with a low one-year mortality rate, among those patients who survived beyond 28 days post hospital discharge. Table II shows a selected subset of these low-risk topics.

For each of the "high-risk" topics in Table I and the "low-risk" topics in Table II, we report the top six UMLS terms associated with each topic, as well as the clinician assigned labels. For each topic, we report its odds ratio (OR, with 95% confidence interval) for post hospital discharge one-year mortality. To define a patient subgroup for each topic, we included patients with topic proportions $\geq 10\%$ for that topic to form a subgroup. Age shown is median with

TABLE II

LOW-RISK TOPICS ASSOCIATED WITH ONE YEAR MORTALITY.

Topic Label by Clinician	Top UMLS Descriptions	Odds Ratio	N, Age (95% CI)	ICD-9	
Cardiovascular disease	Coronary artery disease Hypertensive disease Hypersensitivity showering Mitral valve insufficiency Aortic valve stenosis	0.10 (0.06 0.16)	2010 65 (55, 76)	414.01 427.31 428.0 424.1 424.0	CORON ATHEROSCLER NATIVE ATRIAL FIBRILLATION CONGESTIVE HEART FAILURE AORTIC VALVE DISORDER MITRAL VALVE DISORDER
Trauma	Fracture Trauma Injury Hemorrhage Pain Falls	0.19 (0.13 0.28)	1799 50 (32, 73)	860.0 518.5 427.31 861.21 305.00	TRAUM PNEUMOTHORAX-CLOSE POST TRAUM PULM INSUFFIC ATRIAL FIBRILLATION LUNG CONTUSION-CLOSED ALCOHOL ABUSE-UNSPEC
Acute coronary syndromes	Chest pain Coronary artery disease Stenosis Lesion Myocardial infarction Hypertensive disease	0.30 (0.22 0.42)	2750 68 (57, 79)	414.01 410.71 428.0 410.41 410.11	CORON ATHEROSCLER NATIVE AMI, SUBENDOCARD INFARCT CONGESTIVE HEART FAILURE AMI OTHER INFER WALL INI AMI OTHER ANT WALL INIT
Mental health disorders	Alcohol abuse Depression NOS (disorder) Drug overdose Suicide attempt Hypersensitivity Agitation	0.16 (0.08 0.31)	1172 47 (37, 57)	518.81 507.0 584.9 780.39 291.81	RESPIRATORY FAILURE FOOD/VOMIT PNEUMONITIS ACUTE RENAL FAILURE NOS OTHER CONVULSIONS ALCOHOL WITHDRAWAL
Stroke and head trauma	Hemorrhage Subdural hematoma Headache Aneurysm Subarachnoid hemorrhage Pain	0.51 (0.35 0.75)	1541 63 (47, 78)	431 430 401.9 427.31 780.39	INTRACEREBRAL HEMORRHAGE SUBARACHNOID HEMORRHAGE HYPERTENSION NOS ATRIAL FIBRILLATION OTHER CONVULSIONS

interquartile range. All topics in Tables I and II remained statistically significantly associated with one year mortality after adjusting for SAPS I and co-morbidity variables.

IV. DISCUSSIONS AND CONCLUSIONS

In this paper, we used Hierarchical Dirichlet Process mixture models to automatically discover clinically coherent groups, or "topics", of co-occurring diseases and symptoms, represented as UMLS concepts in a large corpus of hospital discharge summaries. The inference was performed in a completely un-supervised manner; no prior medical knowledge in disease associations was used. We demonstrated that the learned topic structure of diseases and symptoms contain prognostic values in stratifying patients for one-year post hospital discharge mortality risks. As part of our future work, we plan to conduct a more comprehensive analysis to evaluate the clinical utility of our approach in identifying disease subtypes and variants associated with poor prognosis. We aim to combine latent topic structure from unstructured text as well as time series data [9] for improved patient prognosis, and to investigate whether the uncovered disease phenotypes can be used to predict the long-term health and quality-of-life of patients post hospital discharge.

V. ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH) grant R01-EB001659 and R01GM104987 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB).

REFERENCES

[1] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101(476), pp. 1566-1581, 2006.

[2] D. Blei and J. Lafferty, "Topic Models," in *Text Mining: Classification, Clustering, and Applications,* A. Srivastava and M. Sahami, editors, 2009.

[3] D.M. Blei, L. Carin, and D. Dunson, "Probabilistic Topic Models," *IEEE Signal Processing Magazine*, pp. 55-65, Nov. 2010.

[4] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.H. Lehman, G. Moody, T. Heldt, TH Kyaw, B. Moody, and R.G. Mark, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access ICU database," *Crit Care Med*, 39(5), 2011.

[5] L.H. Lehman, M. Saeed, W. Long, R.G. Mark, "Risk stratification of ICU patients using topic models inferred from unstructured progress notes," Proc. AMIA Annu Symposium, pp. 505-511. Nov. 2012.

[6] W. Long, "Extracting Diagnoses from Discharge Summaries," Proc. AMIA 2005 Symposium, pp. 470-474, 2005.

[7] C. Wang, http://www.cs.princeton.edu/∼chongw/.

[8] J.R. Le Gall, P. Loirat, A. Alperovitch, et al. "A simplified acute physiology score for ICU patients. Critical Care Medicine," 12(11), 975-977, 1984.

[9] L.H. Lehman, S. Nemati, R.P. Adams, R.G. Mark, "Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU," Proc. IEEE Eng Med Biol Soc (EMBC), pp. 5939-42, 2012.