Identifying Patterns of Co-occurring Medical Conditions through Topic Models of Electronic Health Records

Moumita Bhattacharya[1], Claudine Jurkovitz, MD, MPH[2] and Hagit Shatkay, PhD[1]
[1]Computational Biomedicine Lab, Computer Sciences, University of Delaware, Newark, DE, USA; [2]Value Institute, Christiana Care Health System, Wilmington, DE, USA

Introduction

Multiple adverse health conditions co-occurring in a patient are typically associated with poor prognosis and increased office or hospital visits[1]. Developing methods to identify patterns of co-occurring conditions can assist in diagnosis by suggesting conditions that may co-occur with a patient's current condition. We thus aim to identify patterns of association among diagnosed conditions by applying *topic modeling* [2], to *Electronic Health Records* (*EHRs*) to identify latent *topics*, each characterized by a distribution over conditions. The dataset we use to train/test such models consists of EHRs of 13,111 patients showing evidence of decrease in kidney function. We specifically use the *diagnosed conditions* attribute in the EHR dataset, listed as *SNOMED-CT* codes[3]. We show that diagnosed conditions that are highly probable to be associated with the same topic, indeed tend to co-occur in patients.

Methods

We employ a well-established topic modeling technique, *Latent Dirichlet Allocation* (*LDA*)[2], to model patient records as though they were generated as a mixture of *K* underlying topics, where a topic is a multinomial distribution over all SNOMED-CT codes. By inferring the probability distributions associated with the topics, we characterize patient records as multinomial distributions over codes. We evaluate the performance of our method in two ways: (*1*) We assess the *medical validity* of our results examining whether the conditions that show a high probability to be associated with the same topic are known to co-occur according to the medical literature; (*2*) We also *quantitatively assess* the topics obtained from our model by measuring their *tightness* and *distinctiveness*. To assess the *tightness* of topics we examine whether each topic can be specified by a small number of coded conditions. The *distinctiveness* is assessed by calculating the *inter-topic distance* using *Jensen-Shannon divergence (JSD)*[4], which measures how well-separated topics are from one another. The *JSD* values range from *0* to $ln(2)$ (~ 0.69), where *0* indicates topics whose distributions are identical, while $ln(2)$ indicates non-overlapping distributions between topics.

Results

We ran multiple experiments varying the number of topics and focus here on results obtained when using 20 topics. Table 1 shows examples of four characteristic topics. Conditions showing a high probability to be associated with the same topic indeed tend to co-occur, as validated by the clinical literature. An inspection of the topics reveals that more than 0.9 of the cumulative probability mass for each topic can be attributed to 10 or fewer codes, which indicates that 10 conditions or fewer can succinctly characterize a topic. The inter-topic distance, as measured by the *JSD*, among all 20 topics has a mean, median, and minimum values of 0.666, 0.692 and 0.483 respectively.

Discussion

Some of the associations among conditions shown within the same topic in Table 1 are well-known. For instance, many of the conditions grouped together in Topic A (leftmost column) are related to *Metabolic bone disease* such as *Limb* or *Joint pain. Metabolic bone disease* is a common complication of advanced *Kidney* disease, which explains the high probability of *Chronic Renal failure*, *Limb*- and *Joint-pain* to all be associated with the same topic[5].

Our results also uncover some *indirect associations* among conditions, which are supported by evidence in the medical literature. For instance, *Allergic Rhinitis* and *Osteoporosis*, two conditions grouped together under Topic B, are not directly associated; however, treating the former with *depot-steroid injections* increases the risk of the latter[6].

Furthermore, the observation that 10 or fewer conditions are sufficient for characterizing a topic illustrates the *tightness* of the topics, while the high mean and median *JSD* values (close to the upper bound of $ln(2)$), indicate that the majority of topic pairs are indeed *distinct*.

Conclusion

In this study, we show that our data-driven approach indeed identifies *tight*, *distinct* topics of co-occurring conditions that are *clinically relevant*. Our approach can suggest conditions that may co-occur with a patient's current diagnosed conditions, and thus has the potential to support clinical decision making.

Table 1. Examples of four characteristic topics from the twenty identified by our model; each column lists ten diagnosed conditions that have the highest probabilities to be associated with the respective topic, along with their probabilities

Topic A	Prob	Topic B	Prob	Topic C	Prob	Topic D	Prob
Pain in limb	.195	Allergic rhinitis	.361	Asthma	.200	Chronic obstructive lung disease	.196
Arthralgia of the lower leg	.166	Osteoporosis	.186	Cough	.164	Tobacco dependence syndrome	.189
Low back pain	.144	Acute sinusitis	.111	Benign neoplasm of colon	.145	Benign essential hypertension	.141
Shoulder joint pain	.128	Benign essential hypertension	.110	Acute bronchitis	.117	Abnormal glucose level	.138
Chronic renal failure	.107	Female sexual arousal disorder	.076	Disorder of lung	.091	Chronic kidney disease stage 3	.132
Arthralgia of the pelvic region and thigh	.092	Chronic sinusitis	.061	Impotence of organic origin	.086	Carpal tunnel syndrome	.102
Thoracic radiculitis	.071	Acute upper respiratory infection	.046	Pneumonia	.076	Heart murmur	.056
Joint pain	.036	Disease of liver	.043	Overweight	.044	Human immunodeficiency virus infection	.044
Acute upper respiratory infection	.030	Acute bronchitis	.007	Cholelithiasis without obstruction	.040	Diarrhea	.000
Chronic rhinitis	.029	Impacted cerumen	.000	Acute upper respiratory infection	.034	Goiter	.000

References

1. Centers for Disease Control and Prevention. Multiple chronic conditions. http://www.cdc.gov/chronicdisease/-about/multiple-chronic.htm, last accessed 10/01/16.
2. Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. Journal of machine Learning research. 2003; 3: 993-1022.
3. NIH U.S. National Library of Medicine. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/, last accessed 10/01/16.
4. Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory. 1991; 37(1):145-51.
5. Margolis DJ, Hofstad O and Feldman HI. Association between renal failure and foot ulcer or lower-extremity amputation in patients with diabetes. Diabetes care. 2008; 31(7):1331-6.
6. Aasbjerg K, Torp-Pedersen C, Vaag A and Backer V. Treating allergic rhinitis with depot-steroid injections increase risk of osteoporosis and diabetes. Respiratory medicine. 2013;107(12):1852-8.