Co-occurrence of Medical Conditions: Exposing Patterns Through Probabilistic Topic Modeling[1]

Moumita Bhattacharya†
Computer and Information
Sciences Dept.
University of Delaware
Newark, DE
moumitab@udel.edu

Claudine Jurkovitz, MD,MPH
Value Institute
Christiana Care Health System
Wilmington, DE
CJurkovitz@christianacare.org

Hagit Shatkay, PhD
Computer and Information
Sciences Dept.
University of Delaware
Newark, DE
shatkay@udel.edu

1 Introduction

Multiple adverse health conditions co-occurring in a patient are typically associated with poor prognosis and increased office or hospital visits. Developing methods to identify patterns of co-occurring conditions can assist in diagnosis. This study aims to identify and characterize patterns of co-occurring medical conditions in patients employing a probabilistic framework. Specifically, we apply topic modeling [1] in a non-traditional way to find associations across SNOMED-CT codes assigned and recorded in the EHRs of *>13,000* patients diagnosed with kidney disease. Unlike most prior work on topic modeling, we apply the method to codes rather than to natural language. Moreover, we quantitatively evaluate the topics, assessing their tightness and distinctiveness, and also assess the medical validity of our results. Our experiments show that each topic is succinctly characterized by a few highly probable and unique disease codes, indicating that the topics are tight. Furthermore, inter-topic distance between each pair of topics is typically high, illustrating distinctiveness. Last, we show that conditions that are highly probable to be associated with the same topic, indeed tend to co-occur in patients. Notably, our results uncover a few indirect associations among conditions that have hitherto not been reported as correlated in the medical literature.

2 Methods

We employ a probabilistic topic modeling method, Latent Dirichlet Allocation (LDA)[1], 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 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*) [2], which measures how well-separated topics are from one another. The JSD values range from 0 to *ln(2)*, where 0 indicates topics whose distributions are identical, while *ln(2)* indicates orthogonal distributions.

3 Results and Discussion

We ran multiple experiments varying the number of topics, and focus here on results obtained when using *20* 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 *15* or fewer codes, illustrates the tightness of the topics. The inter-topic distance among all *20* topics has high mean and median JSD values (close to the upper bound of ln(2)), indicating that the majority of topic pairs are indeed distinct. Our results 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 one topic, are not directly associated; however, treating the former with depot-steroid injections increases the risk of the latter.

4 Conclusion

We show that our approach indeed identifies tight, distinct topics of co-occurring conditions that are clinically relevant, and thus has the potential to support clinical decision making.

REFERENCES

[1] Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. Journal of machine Learning research. 2003; 3: 993-1022.

[2] Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory. 1991; 37(1):145-51.

[1] *ACM-BCB'18*, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233723