HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2020 June 23.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2017 February ; 2017: . doi:10.1109/bhi.2017.7897234.

Causes of death in the United States, 1999 to 2014

Hanyu Jiang,

Dept. of Biomedical Engineering, Peking University, Beijing, 100871, China

Hang Wu [Student Member],

Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA

May Dongmei Wang, Ph.D. [Senior Member, IEEE]

Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA

Abstract

Statistical methods have been widely used in studies of public health. Although useful in clinical research and public health policy making, these methods could not find correlation among health conditions automatically, or capture the temporal evolution of causes of death correctly. To cope with two challenges above, we implement an unsupervised machine learning model, termed topic models, to investigate the mortality data of the United States. Our model successfully groups morbidities based on their correlation, and reveals the temporal evolution of these groups from 1999 to 2014, which are also validated by existing literature. This work could provide a novel view for clinical practitioners to provide more accurate healthcare service, and for public health policymakers to make better policy.

I. INTRODUCTION

The identification and trend analysis of the leading causes of death have been challenging from the aspect of public health. Death certificates, which record demographic information and causes of death, have provided researchers and policymakers an abundant resource of retrospective analysis. For years, traditional analysis uses statistical methods, such as ANOVA (analysis of variance) and chi-square test, on the public health data [1][2][3]. Some progress has been made in this field and public health policymakers have been benefited from these statistical results. Judith Swan et al. statistically analyzed data on cancer screening practices, to help policymaker plan public health strategies to reduce or eliminate health disparities [4]. However, traditional methods treat each cause of death as independent and may hence may overlook the hidden correlation between the different health conditions. Moreover, they may fail to capture the dynamics of these health conditions and thus fail to provide deeper insight for clinical practitioners and public health policymakers.

Given the huge collection of death certificates in the U.S. for several decades, how can we capture the hidden relationships among various causes of death and the temporal evolution of them? In this study, we presented an alternative perspective and novel analysis on the mortality data of the United States that provides a description of the large collection of death certificates using probabilistic generative models. Our contribution is two-fold. First, we use topic models to describe causes of death and find their correlation. Second, we analyze the prevalence of different topics from Year 1999 to 2014, which was the evolution of each topic.

Specifically, we assume health conditions can be grouped into different clusters, termed *topic* in the literature, and each death is a mixture of all these topics (Fig.1). Based on our formulation, each topic captures the hidden correlation among all health conditions, and the combination of topics explain each death case. The prevalence of each topic and their changes through years can thus characterize the temporal evolution of causes of death in the U.S. Our hypothesis is reasonable and leads to a probabilistic model called topic models [5] in the machine learning community.

The rest of our paper is organized as follows: we will present the data and our model in Sec. II and III respectively. We then analyze our results in two different approaches: Sec. IV shows the convergence properties and how we use it to guide our parameter selection, while Sec. V illustrates the algorithm output and validates our model using statistical results in other studies, confirming that ours was convincing and reliable. Besides, we briefly discuss relevant literature in Sec. VI.

II. RELATED WORK

A. Causes of Death

An enormous amount of studies has been carried out, which includes causes of death, and how death rates vary among different patient groups. Dombrovskiy et al. studied the death resulting from severe sepsis, and concluded that mortality rate for females with severe sepsis were greater than that for males [1]. Seshasai et al. stated that the diabetes patients were more likely to die of cancer, vascular causes and some other causes, and the death rates were also associated with gender information [2].

B. Trend Analysis

The trend analysis is also an important issue for public health. Jemal et al. analyzed the causes of death in the United States from 1970 to 2002, claiming that the mortality rates had decreased from stroke (62%), heart disease (52%), and accidents (41%) [3]. Dombrovskiy et al. concluded that although for severe sepsis, case fatality rate showed a decreasing trend from 1993 to 2003, the overall mortality still increased because of the increase of total cases [1].

C. Topic Models

Topic models were originally proposed as Latent Dirichlet Allocation (LDA) [5] and has since gained much popularity in text mining and other applications. Dynamic topic modeling

[6], which studies the evolution of topics, and LDA with hierarchical Dirichlet process [7], which automatically infers the number of topics, are among successful variants of the original LDA model.

III. DATASET DESCRIPTION

Our study uses the multiple causes-of-death mortality data from the National Center for Health Statistics [8]. Each record of this dataset contains information of death certificates from vital statistics offices of each State and District of Columbia in the United States. Each record includes demographic information (e.g.: age, sex, education level, when people died, and where people died), and causes of death for the deceased (up to 20). The causes of death, coded in International Statistical Classification of Diseases (ICD), are selected from the deceased's health records by physicians and temporally ordered.

Standards of ICD codes have been updated several times over time, and the ICD-10 code has been adapted since 1999. To make the study more coherent, we only used the data from 1999 to 2014 in the database mentioned above. An example of ICD-10 code is shown in Table 1.

IV. BITERM TOPIC MODEL

A. Topic Models

Topic modeling is an unsupervised machine learning algorithm for discovering the abstract “topics” that occur in a collection of documents, i.e. corpus. Classic topic-model algorithms, such as LDA [5], treat every topic as a specific statistical distribution over words and treat every document as a mixture of topics. According to the frequency of words in a document, the algorithms can infer the proportions of different topics in a document, to decide which topics the document belongs to.

Formally, let $z \in [1, K]$ be a topic indicator variable, then for the corpus, a K -dimensional multinomial distribution $\vec{\theta}$, with $\theta_k = P(z = k)$ and $\sum_{k=1}^K \theta_k = 1$, is used to represent the prevalence of topics. The distribution of words in each topic is represented as $\phi_{k,w} = P(w|z = k)$ with $\sum_{w=1}^W \phi_{k,w} = 1$. LDA uses symmetric Dirichlet prior for $\vec{\theta}$ with a single-valued hyperparameter α . For a document, the generative process of LDA is described as follows:

1. Draw a prior topic distribution $\vec{\theta} \sim \text{Dirichlet}(\alpha)$
2. For each word in the document
 - a. Draw a topic for this word $z \sim \text{Multinomial}(\vec{\theta})$
 - b. Draw the word $w \sim \text{Multinomial}(\vec{\phi}_z)$ from the dictionary

B. Biterm Topic Model

When dealing with short documents, traditional topic model algorithms suffer serious challenges of sparsity. To overcome the sparsity problem, Cheng et al. propose a novel algorithm called biterm topic model (BTM) [9]. The idea behind BTM is to treat the combination of two words as the fundamental elements (biterm), and to model how the paired words co-occur in the documents. If two words, w_1 and w_2 , co-occur more frequently, the two words are more likely to belong to a same topic.

One major difference between BTM and conventional topic models is that BTM aggregates all word co-occurrence patterns in the corpus rather than by documents. Assume that the corpus contains N_B biterms $B = \{b_i\}$ with $b_i = (w_{i,1}, w_{i,2})$. The generative process of BTM is similar to that of LDA, except that the step 2.b in LDA generative process is modified as drawing w_1 and w_2 from the same topic in BTM. The values of parameters in BTM are estimated by Gibbs sampling.

To address the challenge of massive data, Cheng et al. develop two algorithms for online learning, online BTM algorithm (oBTM) and incremental BTM algorithm (iBTM) [7]. These two algorithms are able to deal with time-sliced dataset and to analyze how topics evolve over time.

In our study, we treat each cause of death as a word in the topic model, and all temporally ordered causes of death for a deceased person as a short document in BTM. The ICD codes are hierarchically structured: prefix for general category and suffix for detailed classification. To discover more general patterns, we discarded the suffix of every detailed code (i.e. for codes A00.0, A00.1 and A00.9, we all treat them as A00).

V. COHERENCE EVALUATION

A. Coherence Score

Because of the unsupervised characteristic of topic models and the nature of data, no label information is available, so evaluation metrics in supervised learning cannot apply here. Instead, some automatic evaluation metrics have been developed to evaluate discovered topics' quality, including the coherence score, which is proposed by Mimno et al [10]. This metric states that if the co-occurrence of the most probable words in a topic is more frequent, then this topic will be considered more coherent, which is consistent with the idea of BTM. So in our study, we use coherence score as the criteria for the quality of topics discovered by BTM.

The definition of coherence score is as followed. We denote the number of documents that have word w as $D(w)$ and the number of documents that have both w and w' as $D(w, w')$. The coherence score is

$$C(t; W^{(t)}) = \sum_{n=2}^N \sum_{l=1}^{n-1} \log \frac{D(w_n^{(t)}, w_l^{(t)}) + 1}{D(w_l^{(t)})} \quad (1)$$

where $W^{(t)} = (w_1^{(t)}, \dots, w_N^{(t)})$ is defined as the ordered list of the N most probable words in topic t .

B. Experiment

We use the average coherence score to determine how many topic number we choose for BTM, and which online algorithm has better performance, oBTM or iBTM. Average coherence score is defined as

$$\overline{C(K, M)} = \sum_{\text{year} = 1999}^{2014} \frac{\sum_{t=1}^K C_{\text{year}}(t, W^{(t)})}{K} \quad (2)$$

where we denote the topic number as K . The average coherence score is the overall coherence score of all documents divided by the number of topics.

In the experiment, we use the code shared by Yan in GitHub, and the parameter settings are by default [11]. The numbers of topics range from 5 to 15. Both oBTM and iBTM are tested in our experiment.

VI. RESULT

A. Coherence Scores

The average coherence score is not a monotone decreasing function of K (Fig.2). When $M=10$, the score has the trend of decreasing for both oBTM and iBTM. When $M=20$, the score is more stable. We hypothesize such irregularity occurs because we used too many most probable codes, and the low $D(v, v')$ smooths the decreasing trend. The overall performance of oBTM is better than that of iBTM, since oBTM gets higher scores.

Based on average coherence score and the fact that Jemal et al. had studied six leading causes of death [3] (so the number of topics should be greater than 6), we used the result of oBTM algorithm and set the topic number to 12.

B. Topics

Because of space constraints, we display only the topics discovered in the mortality data of 2005 and their corresponding most five probable code titles (Table 2). Results in other years are similar to that in 2005. These topics can be interpreted reasonably: topic 1 as newborn topic, topic 2 as heart topic, topic 8 as malignant neoplasm topic, and topic 12 as respiratory tract topic. The code 5 in topic 2 (the heart topic) is diabetes mellitus, which is consistent with the conclusions in Seshasai et al. that heart disease is a major common cause of death of patients with diabetes [2].

B. Trend Analysis

Using the results of oBTM, we can know the proportion of each topic in every year, thus analyze the topic evolution. The trend is shown in Fig.3. Our results are consistent with the work of Jemal et.al on analysis of causes of death in the U.S. in the following two points [3]. First, our results show that the proportion of heart topic is largest among all 12 topics, and

the study also concluded that the heart diseases were the most severe leading cause of death. Second, the proportion of heart topic decreased from 2000 to 2014, which confirms the conclusion that the mortality rate of heart diseases was decreasing in recent years. The three most probable causes of death were unchanged during the 16 years, as displayed in Table 2.

The evolution of another topic also shows some patterns. The newborn topic decreased from 1999 to 2014, which maybe partially resulted from the fact that American birth rate decreased from 14.2 per 1000 in 1999 to 12.5 per 1000 in 2014 [12]. The trends of the malignant neoplasm and respiratory tract topic also decreased from 1999 to 2014.

VII. CONCLUSION

In this study, we implement the BTM to investigate the mortality data of the United States from 1999 to 2014 by clustering them as “topics”, and use the coherence score to measure the quality of discovered topics. The topics have reasonable interpretation (e.g. newborn topic, heart topic, malignant neoplasm topic, and respiratory tract topic), and our results are consistent with those from traditional statistical methods. Also, our discovery may reveal the novel correlation among medical symptoms. Clinical practitioners and public health policymaker could exploit our results to provide better healthcare service.

However, our results are limited by the nature of death certificate. Some physicians and researchers have doubted the accuracy of death certificates, claiming that physicians could be biased when they fill the death certificates according to their perspectives, and might overlook some hidden causes [13]. So based on the noise introduced during the filling process, our result may be noisy and inaccurate. Also, because of the unsupervised characteristic of topic models, determining the number of topics is difficult, we may not use the optimal topic number for this study.

In our future work, we could group the deceased population by their demographic information, to discover different topic in each group, since statistical results have shown that demographic information such as gender and age is associated with death rates for many diseases [1][2]. Besides, we could also exploit the temporal order of causes of death, and it is likely that there are some temporal patterns hidden in the dataset.

Acknowledgement

The authors thank Paula Braun and Mark Braunstein for their valuable insights on the project, and also thank Janani Venugopalan and Ying Sha for their comments on the manuscript.

Hanyu Jiang thanks the funding from Department of Biomedical Engineering, Peking University.

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health and National Science Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health nor National Science Foundation.

REFERENCES.

- [1]. Dombrovskiy VY, Martin AA, Sunderram J, and Paz HL, “Rapid increase in hospitalization and mortality rates for severe sepsis in the United States: a trend analysis from 1993 to 2003.,” *Critical care medicine*, vol. 35, no. 5 pp. 1244–50, 2007. [PubMed: 17414736]

- [2]. Seshasai SRK et al., "Diabetes Mellitus, Fasting Glucose, and Risk of Cause-Specific Death," N. Engl. J. Med, vol. 364, no. 9, pp. 829–841, 2011. [PubMed: 21366474]
- [3]. Jemal A, Ward E, Hao Y, and Thun MJ, "Trends in the leading causes of death in the United States, 1970–2002," JAMA, vol. 294, no. 10, pp. 1255–1259, 2005. [PubMed: 16160134]
- [4]. Swan J, Breen N, Coates RJ, Rimer BK, and Lee NC, "Progress in cancer screening practices in the United States," Cancer, vol. 97, no. 6 pp. 1528–1540, 2003. [PubMed: 12627518]
- [5]. Blei DM, Ng AY, and Jordan MI, "Latent Dirichlet allocation," J. Mach. Learn. Res, vol. 3 pp. 993–1022, 2003.
- [6]. Blei DM and Lafferty JD, "Dynamic Topic Models," International Conference on Machine Learning, pp. 113–120, 2006.
- [7]. Teh YW, Jordan MI, Beal MJ, and Blei DM, "Hierarchical Dirichlet Processes," J. Am. Stat. Assoc, vol. 101, no. 476, pp. 1566–1581, 2006.
- [8]. Nber.org. (2016). Mortality Data -- Vital Statistics NCHS' Multiple Cause of Death Data, 1959–2014. [online] Available at: <http://www.nber.org/data/vital-statistics-mortality-data-multiple-cause-of-death.html> [Accessed 1 Sep. 2016].
- [9]. Cheng X, Yan X, Lan Y, and Guo J, "BTM: Topic modeling over short texts," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 12 pp. 2928–2941, 2014.
- [10]. Mimno D, Wallach HM, Talley E, Leenders M, and McCallum A, "Optimizing Semantic Coherence in Topic Models," Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, no. 2 pp. 262–272, 2011.
- [11]. GitHub. (2016). xiaohuiyan/OnlineBTM. [online] Available at: <https://github.com/xiaohuiyan/OnlineBTM> [Accessed 22 Jul. 2016].
- [12]. Statista. (2016). U.S. birth rate 1990–2014 | Statistic. [online] Available at: <https://www.statista.com/statistics/195943/birth-rate-in-the-united-states-since-1990/> [Accessed 14 Oct. 2016].
- [13]. Lloyd-Jones DM, "Accuracy of Death Certificates for Coding Coronary Heart Disease as the Cause of Death," Annals of Internal Medicine, vol. 129, no. 12 p. 1020, 1998. [PubMed: 9867756]

Figure 2.
Average coherence scores of oBTM and iBTM results.

Figure 3.
Topics evolution

Table 1.

An Example of Icd-10 Code.

Code	ICD Title
A00	Cholera
A00.0	Cholera due to <i>Vibrio cholerae</i> 01, biovar cholerae
A00.1	Cholera due to <i>Vibrio cholerae</i> 01, biovar eltor
A00.9	Cholera, unspecified

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Some Topics of 2005 Mortality Data

Topic No.	Code 1	Code 2	Code 3	Code 4	Code 5
1	Effects of other external causes	Other respiratory conditions originating in the perinatal period	Cardiovascular disorders originating in the perinatal period	Respiratory distress of newborn	Bacterial sepsis of newborn
2	Chronic ischemic heart disease	Heart failure	Cardiac arrest	Acute myocardial infarction	Unspecified diabetes mellitus
8	Secondary malignant neoplasm of respiratory and digestive organs	Secondary malignant neoplasm of other sites	Malignant neoplasm of bronchus and lung	Cardiac arrest	Malignant neoplasm of breast
12	Pneumonitis due to solids and liquids	Inhalation and ingestion of other objects causing obstruction of respiratory tract	Foreign body in respiratory tract	Stroke, not specified as hemorrhage or infarction	Respiratory failure, not elsewhere classified