[[1]](#footnote-1)

DeepDeath: Learning to Predict the Underlying Cause of Death with Big Data \*

Hamid Reza Hassanzadeh, *Student Member, IEEE*, Ying Sha, Paula Burns, Centers for the Disease Control (CDC), and May D. Wang, Senior Member, IEEE

*Abstract*— Multiple cause of death data Learning to to the best of our knowledge this is the first deep learning pipeline. We use unstructured data.

# INTRODUCTION

Many of the scientific discussions and studies in biomedical and healthcare domains address tasks whose end goal is to prevent death or diseases. Since the emergence of the Big Data science, numerous machine learning based techniques and technologies have been proposed and applied to improve human health by solving different computational challenges that we face today. A less obvious question to ask from the big-data communities is whether they can contribute to our understanding of factors leading to death or diseases. In fact, there has been a number of studies [some references] that have proven to offer profound impacts on our understanding of the major causes of death by statistical analysis of death data. In light of that, we were interested to explore the feasibility of big data science, and deep learning in particular, for conducting predictive analytics on multiple cause of death data.

Multiple cause of death data provides a valuable source of information that can be used to analyze death trends in chronic disease such as HIV [1, 2] and lung disease [3], disease diffusion [mention some reference], to provide better understanding of multi-morbid associations between conditions leading to death [reference], [mention other applications and references], and to identify problems with the process of coding and recording cause of death information [4]. As such, designing advanced analytics pipelines for discovering descriptive statistics and trajectories is highly [some verb!]. The sheer amount of available data from recorded death certificate data, makes them suitable for big-data analysis techniques but at the same time pose some key challenges. In particular, the mortality multiple-cause data is unstructured and can be inaccurate due to several reasons [a good reference that describes the reasons for inaccurate entries]. Moreover, the high number of ICD-9/10 mortality codes make analysis of multiple-cause associations even more challenging. This altogether, calls for advanced techniques for mining in large datasets of unstructured, high dimensional, and noisy environment.

Despite the importance of the subject, only a handful of researches have so far conducted studies in which the sought to relate multiple causes of death to other factors. These studies are often restricted to classical statistical methods that can be put into four categories [5]: 1) Univariate measures, consisting of counts and frequencies, 2) cross-tabular measures, which incorporate variables that identify the roles (e.g. contributory, non-contributory, complication and underlying) associated with multiple death causes, 3) measures of association, in which some measure of multiple mentions of a cause is related to some measure of mentions of the underlying cause; and finally, 4) derived measures, where univariate measures such as multiple-cause rates are integrated to build higher order models.

In this study, we present an exploratory analysis that is well positioned in a fifth group, by building upon both the third and the fourth categories above, and utilizing advanced machine learning approaches. Specifically, we propose two different categories models, namely, shallow learners to learn mono/bi-gram features derived from the multiple-cause data which we trained over Hadoop using the MapReduce programming model as well as a deep recurrent neural network that learns the dynamics of the morbidity chains efficiently. The rest of the paper is organized as follows. In section II we detail the describe the data format as well as the challenges that we face when dealing with it. Then we detail the shallow learners that we train over the Hadoop framework. We also present our deep model in the same section and our motivation to resort to deep learning. Next, in section III, we compare and contrast the aforementioned models through different experiments and show that our deep model can model the data more efficiently and finally in section IV, we conclude the paper and shed light on future directions we would like to pursue.

# Materials and Methods

xxxx

## Data source

We used mortality data published by the United States National Center for Health Statistics (NCHS) which is available to public.

## A. Explanation of the first non-deep approach

TBF out

## B. DeepDeath

Figure 1: The block diagram of DeepDeath



TBF out

# Results

One paragraph about the data and the filtering criteria.

First explain the NCHS data.

1. Sample Table

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

a. Sample of a Table footnote. (Table footnote)

Example of a figure caption. *(figure caption)*

# Conclusion

Concluded that deep is the power!.

Acknowledgment

Would like to thank Dr. Baurenstein and CDC folks.

References

[1] T. L. Chorba, R. C. Holman, M. J. Clarke, and B. L. Evatt, "Effects of HIV infection on age and cause of death for persons with hemophilia A in the United States," *American journal of hematology,* vol. 66, pp. 229-240, 2001.

[2] W. C. Hooper, R. C. Holman, M. Clarke, and T. L. Chorba, "Trends in non‐hodgkin lymphoma (NHL) and HIV‐associated NHL deaths in the United States," *American journal of hematology,* vol. 66, pp. 159-166, 2001.

[3] D. M. Mannino, C. Brown, and G. A. Giovino, "Obstructive lung disease deaths in the United States from 1979 through 1993: an analysis using multiple-cause mortality data," *American journal of respiratory and critical care medicine,* vol. 156, pp. 814-818, 1997.

[4] C. Gordon, "Australian Bureau of Statistics, Multiple cause of death analysis. Publication 3319.0. 55.001," ed, 2003.

[5] S. Bah and M. M. Rahman, "Measures of multiple-cause mortality: a synthesis and a notational framework," *Genus,* vol. 65, pp. 29-43, 2009.

1. \*This work was supported by the grants from the Centers for Disease Prevention and Control (CDC) and the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

   H. R. Hassanzadeh is with the Department of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. (email: [hassanzadeh@gatech.edu](mailto:hassanzadeh@gatech.edu)).

   Y. Sha is with the Department of Biology, Georgia Institute of Technology, Atlanta, GA 30332 USA. (email: [ysha8@gatech.edu](mailto:ysha8@gatech.edu))

   P. Burns is with the Center for Diseases Control and Prevention (CDC), Atlanta, GA 30333 USA. (email: [pax1@cdc.gov](mailto:pax1@cdc.gov))

   M. D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University and the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (corresponding author, phone: 404-385-2954; e-mail: [maywang@bme.gatech.edu](mailto:maywang@bme.gatech.edu)). [↑](#footnote-ref-1)