[[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 provides a valuable source of information that can be used to enhance health standards by predicting health related trajectories of large populations of a society. We propose a deep pipeline that can learn the temporal and non-temporal aspects of the multiple-cause mortality data. We apply our model to mortality data provided by the national center for health statistics and compare our results to the multiple classical models that are based on N-gram features and show that our proposed pipeline outperforms the proposed baselines without a need for enforcing ad-hoc expert driven rules.

# 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 for researchers that remains to be extensively explored, is whether Big Data science can contribute to our understanding of factors leading to death or diseases via analysis of multiple-cause mortality data. In fact it is widely believed that counting the dead is a significant investment to reduce the premature mortality [2]. There has been a number of studies that have proven to offer profound impacts on our understanding of the major causes of death using the statistical analysis of recorded death data. In light of these studies, we were interested the feasibility of big data science, and deep learning in particular, in learning hidden complex patterns available in the haystack of mortality datasets.

Multiple cause of death data provide a valuable source of information that can be used to analyze death trends in chronic disease such as HIV [3, 4] and lung disease [5], disease diffusion for controlling plague and other epidemics, to provide better understanding of multi-morbid associations between conditions leading to death, and to identify problems with the process of coding/recording cause of death information [6]. 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 [7]: 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.

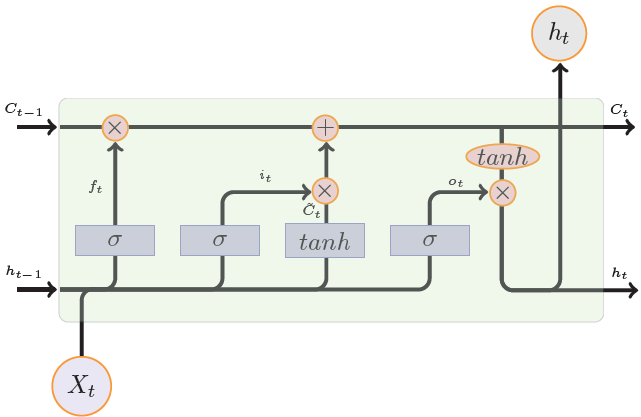
# Materials and Methods

## Data source

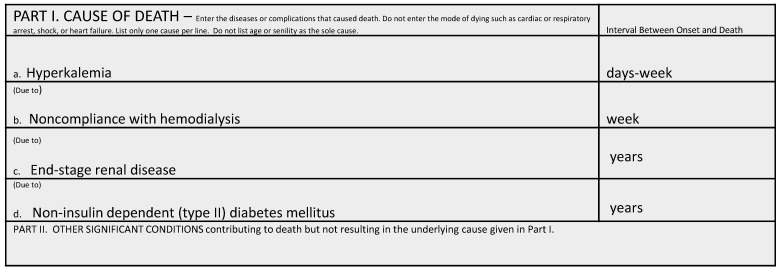
We used mortality data published by the United States National Center for Health Statistics (NCHS) [put a reference] which is available to public. Specifically, we used the 2015 reported

We briefly introduce the typical process of completing a death certificate here. Civil registration systems collects death information of individuals in the form of a death certificate, the standard format of which is designed by the World Health Organization (WHO) [8]. The section that is of most interest to public health researchers is the cause-of-death section, which has to be completed by a medical certifier. An ideal person to complete a death certificate is the attending physician, who has most sufficient clinical expertise and judgement. However, if the manner of death is unnatural or unplanned, a medical examiner or a coroner can also fill in a death certificate. The cause-of-death section is divided into two parts (figure 1), where part I lists the causal chain of conditions directly leading to the death in reverse chronical order, and part II includes the conditions that contributes but not directly leads to death [9].

1. Figure 2: Internal Design of an LSTM Module



1. An example of cause-of-death certification [1]



One paragraph about the data and the filtering criteria.

## Using n-Gram Models to Learn [Blah Blah]

To construct our baseline models, we use the n-gram model to extract features from data. N-gram is a contiguous sequence of n tokens from text or sequential data [10], which has been widely applied in the area of natural language processing [11-13] and bioinformatics [14, 15]. The n of n-gram can be any positive integers, where we generate unigrams if we choose n to be one, and bigrams if we choose n to be two. Longer n-grams are able to encode more context information. Since the majority of conditions on death certificates follow sequential order, we can construct n-grams as features from death certificates.

The exact process of constructing n-gram features consists of two steps, selecting n-grams and vectorization. First we process all samples in our data, and generate a dictionary of n grams generated from data. We select the top K frequent n-grams as features and then represent the data by a MK matrix, where M is the number of samples. Each row of the matrix denotes a sample death certificate characterized by the counts of chosen K n-grams. For our data, we prepare three sets of features. We construct the first set of features with all unigrams, the second set with top 5,000 bigrams, and the third set using the top 10,000 frequent unigrams and bigrams. After constructing the features, we will feed them into k nearest neighbors [16] and random forests [17].

* Discuss scalability issue and why we need to run on hadoop

## Deep Recurrent Neural Networks

Deep learning is an emerging analytics technology that is now being deployed in a wide range of domains including Biomedical areas [18-20] due its success in improving the previously recorded state-of-the-art performance measures. Deep learn is now becoming an indispensable part of any wining model in today’s complex computational challenges. Recurrent neural networks, and long short-term memory networks [21], in particular, are an important class of deep architectures that are able to capture the temporal dynamics sequential data and as such, they fit the task of learning the time ordered chain of morbidity events that eventually lead to death. These models have recently got a significant attention from the researchers due to more powerful hardware, such as graphics processing units (GPUs), due to their massively parallel architecture that makes learning on large sequential datasets (such as NCHS multiple-cause mortality data) feasible. Figure 1 shows the internal structure of an LSTM block.

## DeepDeath

# Results

* Bigrams capture some termporal information by comparing unigrams to bigrams.

We represent XXX of YYY in a two-dimensional space by t-SNE [28], as shown in figure X. We divide the samples into two groups based on whether the underlying cause belong to infectious and parasitic diseases (ICD-10 code starting with “A” or “B”) or not. We believe that the pathologies and disease progression patterns of infectious and parasitic diseases are significantly different from other diseases. From figure X, we observe that the dots representing infectious and parasitic diseases primarily form unique structures on the right of the figure, which suggests that our deep learning model have the potential to distinguish samples with clinically interpretable meanings.

Table 1: Performance comparision between baselines and the DeepDeath

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Random Forest on Hadoop | | | DeepDeath |
| UniGram  (5000) | BiGram  (5000) | Uni+BiGram  (10000 |
| SN | 31.92% | 22.11% | 34.48% | 36.96% |

1. Figure 3: The block diagram of DeepDeath

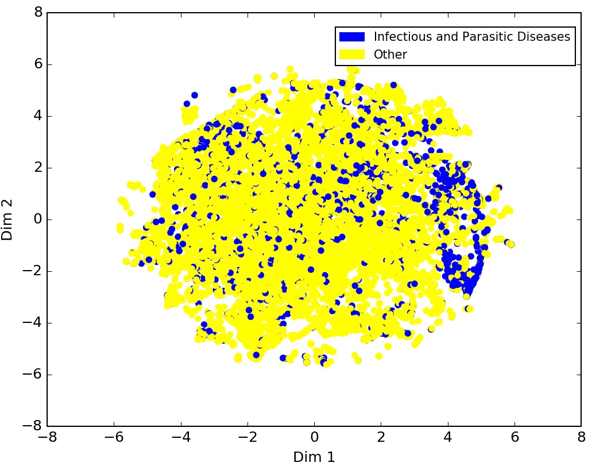


# Conclusion

Concluded that deep is the power!.

Future work to use other available features in recorded moratlity data such as demographic, socioeconomic

Also to use the trained model to derive clusters of features

1. xxxxxx 

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

1. Brooks, E.G. and K.D. Reed, *Principles and pitfalls: a guide to death certification.* Clinical medicine & research, 2015. **13**(2): p. 74-82.

2. Jha, P., *Counting the dead is one of the world’s best investments to reduce premature mortality.* Hypothesis, 2012. **10**(1): p. e3.

3. Chorba, T.L., et al., *Effects of HIV infection on age and cause of death for persons with hemophilia A in the United States.* American journal of hematology, 2001. **66**(4): p. 229-240.

4. Hooper, W.C., et al., *Trends in non‐hodgkin lymphoma (NHL) and HIV‐associated NHL deaths in the United States.* American journal of hematology, 2001. **66**(3): p. 159-166.

5. Mannino, D.M., 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, 1997. **156**(3): p. 814-818.

6. Gordon, C., *Australian Bureau of Statistics, Multiple cause of death analysis. Publication 3319.0. 55.001*. 2003.

7. Bah, S. and M.M. Rahman, *Measures of multiple-cause mortality: a synthesis and a notational framework.* Genus, 2009. **65**(2): p. 29-43.

8. Organization, W.H., *International statistical classification of diseases and related health problems*. Vol. 1. 2004: World Health Organization.

9. Israel, R.A., H.M. Rosenberg, and L.R. Curtin, *Analytical potential for multiple cause-of-death data.* American journal of epidemiology, 1986. **124**(2): p. 161-81.

10. Brown, P.F., et al., *Class-based n-gram models of natural language.* Computational linguistics, 1992. **18**(4): p. 467-479.

11. Pak, A. and P. Paroubek. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. in *LREc*. 2010.

12. Marafino, B.J., et al., *N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit.* Journal of the American Medical Informatics Association, 2014. **21**(5): p. 871-875.

13. Kešelj, V., et al. *N-gram-based author profiles for authorship attribution*. in *Proceedings of the conference pacific association for computational linguistics, PACLING*. 2003.

14. Tomović, A., P. Janičić, and V. Kešelj, *n-Gram-based classification and unsupervised hierarchical clustering of genome sequences.* Computer methods and programs in biomedicine, 2006. **81**(2): p. 137-153.

15. Wajid, B. and E. Serpedin, *Review of general algorithmic features for genome assemblers for next generation sequencers.* Genomics, proteomics & bioinformatics, 2012. **10**(2): p. 58-73.

16. Keller, J.M., M.R. Gray, and J.A. Givens, *A fuzzy k-nearest neighbor algorithm.* IEEE transactions on systems, man, and cybernetics, 1985(4): p. 580-585.

17. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

18. Park, Y. and M. Kellis, *Deep learning for regulatory genomics.* Nat Biotechnol, 2015. **33**(8): p. 825-6.

19. Hassanzadeh, H.R. and M.D. Wang, *DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins*, in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. 2016, IEEE. p. 178-183.

20. Alipanahi, B., et al., *Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning.* Nature biotechnology, 2015. **33**(8): p. 831-838.

21. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

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)