[[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 in societies with large populations. These data are often available in large quantities across U.S. states and require Big Data techniques to uncover complex patterns that are hidden in them. We design two different classes of models suitable for large-scale analysis of mortality data, a Hadoop based ensemble of random forests trained over N-grams, and the DeepDeath, a deep classifier based on recurrent neural networks paradigm. We apply both classes to the mortality data provided by the National Center for Health Statistics and show that while both perform significantly better than the random classifier, the deep model that utilizes long short-term memory networks (LSTMs), surpasses the N-gram based models and is capable of learning the temporal aspect of the data without a need for building ad-hoc, expert-driven features.

# 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, that remains to be extensively explored by researchers, 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 to investigate the feasibility of this emerging field in learning hidden complex patterns that are available in the haystack of mortality datasets.

Multiple causes of death data provide a valuable source of information that have been used to analyze death trends in chronic disease such as HIV [3, 4] and lung disease [5], to identify problems with the process of coding/recording cause of death information [6], and can be potentially used in analysis of disease diffusion for controlling plague and other epidemics as well as providing a better understanding of multi-morbid associations between conditions leading to death, As such, designing advanced analytics pipelines for discovering descriptive statistics and trajectories is highly crucial. While the sheer amount of available data collected from the registered death certificates, makes it amenable to Big Data analysis, it poses some key challenges at the same time. In particular, the mortality multiple-cause data are unstructured and are often inaccurate and noisy. Moreover, the high number of ICD-9/10 mortality codes makes analysis of multiple-cause associations even more challenging. These altogether, call for advanced techniques for mining large datasets of unstructured, high dimensional, and noisy structure.

Despite the importance of the subject, only a handful of researches have so far conducted studies in which they sought to relate multiple causes of death to each other or to other factors. These studies are often limited to classical statistical methods (measures) that do not scale up efficiently and are put into four major 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 abovementioned categories, and utilizing advanced machine learning approaches. More specifically, we propose two different classes of models for large-scale analysis of data, namely, shallow learners to learn uni/bi-gram features derived from the multiple-cause data that run over Hadoop framework using the MapReduce programming model, and a deep recurrent neural network that learns the temporal dynamics of the event chains efficiently. The rest of the paper is organized as follows. In section II we detail the data format as well as the challenges that we face when dealing with it. Then we describe the shallow learners that are designed to work over the Hadoop framework. We also present our proposed deep model in the same section and our motivation to resort to deep learning. Next, in section III, we compare the accuracy of each model when applied to a large dataset show that our deep pipeline outperforms the baselines we design by utilizing its ability to learn the temporal aspect of the data and finally in section IV, we conclude the paper and shed light on future directions that we would like to pursue.

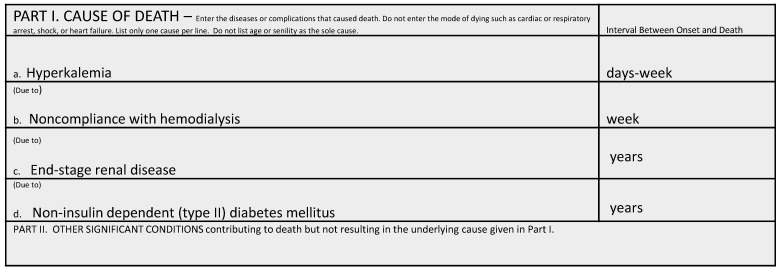
# Materials and Methods

## Data source

Civil registration systems collect death information from deceased persons in form of death certificates, based on a standard format that 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 sufficient clinical expertise and judgement on the occurred death, however, in case the manner of death is unnatural or unplanned, a medical examiner or a coroner can also fill in the death certificate. The cause-of-death section is divided into two parts (see Figure 1). Part I lists the causal chain of conditions directly leading to the death in reverse chronological order, and part II includes the conditions that contributes but not directly leads to death [9]. The conditions in part II are not ordered in time.

We used 2015 mortality data published by the United States National Center for Health Statistics (NCHS) which consists of over 2.7 million deaths recorded in U.S. during the year 2015. Because we were interested in the temporal information that is available in the multiple causes of death, we filtered out the conditions listed in part II as well as the underlying causes if listed in this part. We also removed that records with attributed unnatural underlying causes of death such as suicide. Based on the 113-recode of the underlying cause, these are assigned to codes 111-113. Moreover, if an underlying cause appears less than 1000 times throughout the whole dataset, we exclude that from our training and test sets leaving 67 recoded classes of the underlying causes. Our goal in this study was to predict the class attributed to each case given the multiple cause conditions listed for him/her. We divided the resulting set into a training set and a test set comprising 70% and 20% of the samples, respectively.

1. An screenshot of the multiple cause section in a death certificate [1]

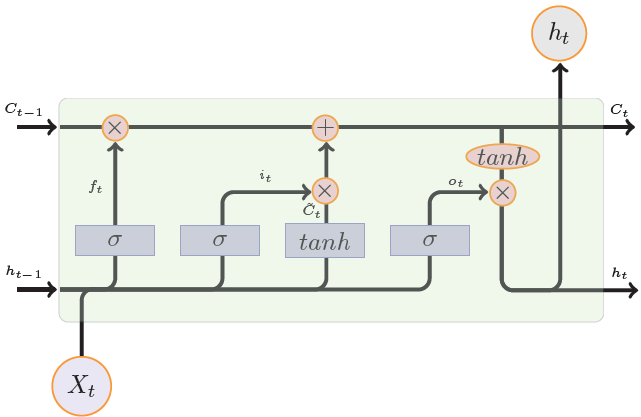


## Using n-Gram Models to Learn Associations Between Multiple Causes

To construct our baseline models, we derived the n-gram features where an n-gram is a defined as an n-tuple consisting of *n* consecutive tokens within sequential data [10] (in our case, causes listed in part I of the death certificates). N-gram based models have been widely used in natural language processing [11-13] and bioinformatics [14, 15] due to their performance and ease of implementation. In this study, we only use uni-gram features and bi-gram features. Despite the fact that higher order n-grams (e.g. tri-grams) can provide more expressiveness and capture more context from the data, they make the models prone to overfitting due to an exponential increase in the number of possible features, which also makes training the resulting model computationally infeasible, therefore, in this study, we only included uni-gram and bi-gram features.

Once the n-gram features derived, we used random forest (RF) to train over the extremely large and sparse matrix of features. RF is an efficient model for dealing with sparse features, however, it does not properly scale up to fit the size of NCHS datasets, therefore, we implemented our model over the Hadoop framework. Hadoop is a distributed storage and computing framework inspired by Google File System (GFS) and Google MapReduce [16]. It uses a reliable storage mechanism, the so called Hadoop Distributed File System (HDFS) for fast storage and retrieval of large size datasets. Hadoop can efficiently utilize commodity compute nodes, that are distributed in a network, to split the tasks into smaller sub-tasks and perform the analysis on the smaller chunks of data through MapReduce, an efficient model that runs on HDFS file system. To train our baseline models over the complete dataset, we used Hadoop streaming and used the python Scikit-learn library to train multiple random forests on subsampled data provided by the mapping stage. Once multiple random forests were trained by the reducer jobs, we aggregate them through majority voting using the scores that each model predict on the test set.

1. Internal Design of an LSTM Module

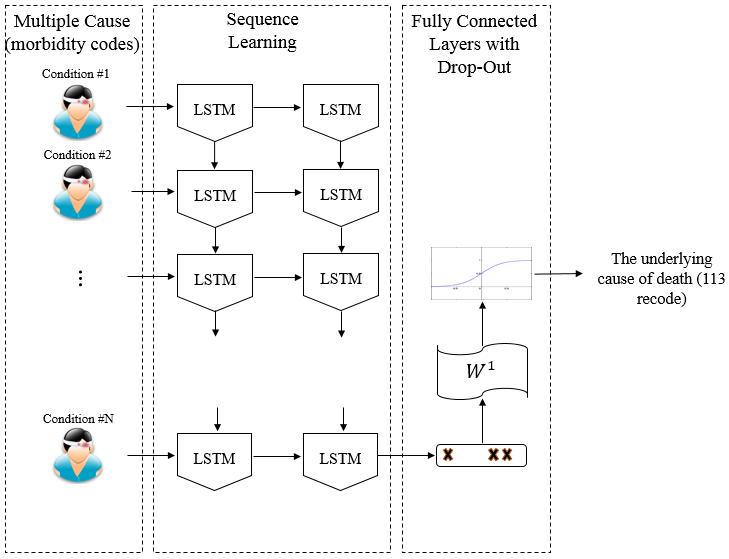


## Deep Recurrent Neural Networks

Deep learning is an emerging technology that is now being deployed in a wide range of domains including Biomedical areas [17-19] due its success in improving the previously recorded state-of-the-art performance measures. Deep learning is now becoming an indispensable part of any wining model in today’s complex computational challenges. Long short-term memory networks (LSTMs) [20] are an important class of deep architectures that are able to capture the temporal dynamics of sequential data. LSTM networks have recently proved to outperform surprisingly well in many traditional sequence learning algorithms, such as the hidden Markov models (HMMs) and conditional random fields (CRF), in supervised settings where significant amount of labeled data is available. As such, LSTMs fit the framework of the task we would like to solve, namely, using the timed causes of death events that contributed to the final event, death, to predict the underlying cause that initiated these events in the first place/xxxxxxx. 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. **Error! Reference source not found.** shows the internal structure of an LSTM block.

## DeepDeath

1. The block diagram of 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 [21], 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.

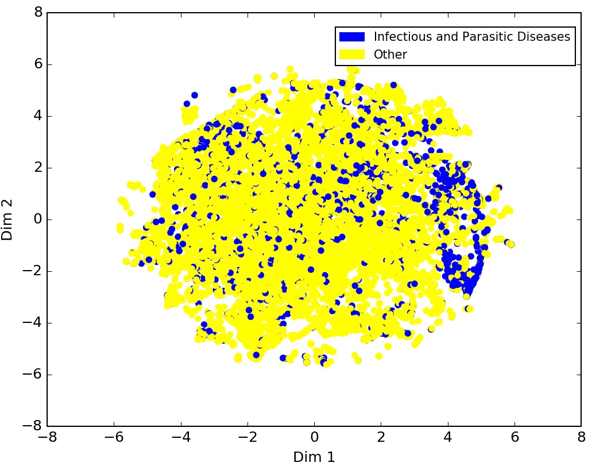
1. xxxxxx 

Table 1: Performance comparision between baselines and the DeepDeath

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Random Forest on Hadoop | | | DeepDeath |
| UniGram  (5000) | BiGram  (5000) | Uni+BiGram  (10000 |
| ACC  (67 classes) | 31.92% | 22.11% | 34.48% | 36.98% |

# Conclusion

Concluded that deep is the power!.

Our approach is the first step in designing more advanced decision support systems that can be deployed in hospitals for providing useful hints to physicians and practitioners in filling out the death certificates properly.

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

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] E. G. Brooks and K. D. Reed, "Principles and pitfalls: a guide to death certification," *Clinical medicine & research,* vol. 13, pp. 74-82, 2015.

[2] P. Jha, "Counting the dead is one of the world’s best investments to reduce premature mortality," *Hypothesis,* vol. 10, p. e3, 2012.

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] R. A. Israel, H. M. Rosenberg, and L. R. Curtin, "Analytical potential for multiple cause-of-death data," *American journal of epidemiology,* vol. 124, pp. 161-81, 1986.

[10] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics,* vol. 18, pp. 467-479, 1992.

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

[12] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, R. A. Dudley, and J. Boscardin, "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,* vol. 21, pp. 871-875, 2014.

[13] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the conference pacific association for computational linguistics, PACLING*, 2003, pp. 255-264.

[14] A. Tomović, P. Janičić, and V. Kešelj, "n-Gram-based classification and unsupervised hierarchical clustering of genome sequences," *Computer methods and programs in biomedicine,* vol. 81, pp. 137-153, 2006.

[15] B. Wajid and E. Serpedin, "Review of general algorithmic features for genome assemblers for next generation sequencers," *Genomics, proteomics & bioinformatics,* vol. 10, pp. 58-73, 2012.

[16] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM,* vol. 51, pp. 107-113, 2008.

[17] Y. Park and M. Kellis, "Deep learning for regulatory genomics," *Nat Biotechnol,* vol. 33, pp. 825-6, 2015.

[18] H. R. Hassanzadeh and M. D. Wang, "DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins," presented at the Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on, 2016.

[19] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning," *Nature biotechnology,* vol. 33, pp. 831-838, 2015.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, pp. 1735-1780, 1997.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research,* vol. 9, pp. 2579-2605, 2008.

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)