Trends, Patterns, and Predictions for Disease Mortality in the US

Jason Lacson Banaag, Min Zhong, Mingyu Yan, Stephanie M Deen, Xiaoming Wang

1. Introduction and Objectives. Disease specific death is one of the most fundamental indicators of population health (Mboera et al., 2018). Trends and patterns in disease specific death can inform decision makers about what policies might be working, where progress lags behind, and the emergence of new health challenges (GBD 2017 Causes of Death Collaborators, 2018). The recent health challenge, COVID-19, affects health-system decision making. Professionals must decide on an effective control strategy, provide massive funding to develop new vaccines and drugs, guide health care providers for testing and treatment, and even suggest travel and gathering policies (Fischhoff, 2020). Understanding possible future trends in mortality and their causes is crucial to guiding investments and policy implementation (Foreman et al., 2018). The objective of this project is to develop an interactive mapping tool which allows users to visualize trends, patterns, and predictions of disease mortality over time at the state level. We expect our product could be used by three groups of professionals: public-health officials; healthcare providers; funding agencies

2. Problem Definition. Our project aims to address the following research problems: What will be the leading causes of death in the next decade? How does disease mortality change with time and space?

3. Literature Survey. The Centers for Disease Control and Prevention (CDC) publishes an annual data brief that reviews year to year trends in mortality data broken down by factors such as age, race, and cause of death (Kochanek et al., 2020). This is a paper with several bar charts, but no interactive visualization for better understanding of the data. A study by Dwyer-Lindgren et al. analyzed the major causes of death at the county level of the United States and produced a visualization that is much more useful for users (Dwyer-Lindgren et al., 2016), however the study and data only include the time period 1980-2014.

The Lee-Carter Method is widely used for projecting mortality estimates, and considered to be highly accurate (Booth & Tickle, 2008). Lee discusses limitations to the model and extensions that have been done including disaggregation by sex, geography, cause of death, and alterations to the algorithm (Lee, 2000). Tuljapurkar and Boe reviewed various considerations when predicting mortality data, including comparing mortality in the young vs old and between genders. They also reviewed different methods for forecasting mortality, including the Lee-Carter model (Tuljapurkar & Boe, 1998). In 2002, Brouhns et al. proposed a Poisson distribution for estimating Lee-Carter that is widely used today (Brouhns et al., 2002) and the traditional Lee-Carter uses Auto Regressive Integrated Moving Average (ARIMA) for parameter estimation.

Recently, neural networks have been used to predict mortality rates (Petneházi & Gáll, 2019) and have been used to predict parameters of the Lee-Carter model (Lindholm & Palmborg, n.d.; Marino et al., 2021). Long Short Term Memory (LSTM) neural networks are a type of recurrent neural network (RNN) that is particularly well suited for predicting sequential data. RNNs suffer from the so-called vanishing gradient problem, in which the weights can become exponentially large or get extremely close to zero. LSTMs solve the vanishing gradient problem by using an input gate, an output gate, and a forget gate which control the flow of information through the cell, including how much weight to give to previous cell states (Bhagwat et al., 2019).

4. Proposed Method

4.1 Intuition. Traditionally, visualizations about the subjects only utilize historical data (Dwyer-Lindgren et al., 2016) and published work on forecasts is classically illustrated with static charts and tables that users are not able to interact with (Lilienfeld & Perl, 1993); (Murray & Lopez, 1997); (Elkins & Claiborne Johnston, 2003)). The innovation of our method includes: 1) compare and implement three state-of-the-art forecasting models in predicting disease mortality; 2) combine and visualize both historical and forecasted disease mortality at the state level.

4.2 Data Source and Preprocess

Data Source: We obtain annual disease mortality data (1979-2019) from the CDC WONDER database.
Data Preprocess: The mortality data prior to 1999 uses ICD-9 coding system for the cause of death, whereas post 1999 data uses ICD-10 coding system. We match the two coding systems to make the data consistent in terms of underlying causes. The time series mortality data is split into two parts: an initial fit period (80%) in which a model is trained, and a subsequent (temporally) testing period (20%) held out for estimating the performance of the model. We drop missing data without imputation and no standardization is performed except for LSTM.

4.3 Forecasting Modeling Algorithms

We use three models to predict mortality rate (death per 100K population): ARIMA, LSTM, and Poisson Lee-Carter model. ARIMA and LSTM can be used for various time series forecasting, including mortality prediction, while the Poisson Lee-Carter model is specifically designed for mortality. The Poisson Lee-Carter model requires additional time-series modeling for its time-dependent parameter and we will use either ARIMA or LSTM to model this parameter (details described in 4.3.3).

4.3.1 ARIMA is a class of widely used models that predict time series data based on its own past values including its own lags and the lagged forecast errors. An ARIMA model is characterized by 3 terms: p, d, q where p is the order of the AR term, q is the order of the MA term, and d is the number of differencing required to make the time series stationary. A general forecasting equation for ARIMA is:

$$\hat{y}_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j}$$

y_{t-i} represents p AR terms, e_{t-j} is q MA term, Φ and θ are the parameters for AR and MA, respectively. One of the challenges for our project is to predict mortalities of different causes at state level and for various age and gender groups.
4.3.2 LSTM In the current work, we will use an LSTM implementation from the Keras Python library (Keras Team, n.d.). We will prepare the data by splitting it into training and testing sets,

scaling the data using scikit learn MinMaxScaler (Pedregosa et al., 2011), and remove trends by differencing the data (subtracting the previous time step's data from the current time step's). We will use mean squared error as the cost function and Adam stochastic gradient descent as the optimization algorithm. In keeping with the consensus from other work, we will have one hidden layer. Parameters we will tune include: the activation function (the tangent hyperbolic function is the default), the number of input neurons (to range from one to the number of test samples in increments of three), and the number of epochs.

4.3.3 Poisson Lee-Carter In the Poisson Lee-Carter model (Brouhns et al., 2002), we assume the number of deaths ($D_{x,t}$) in the age group x and year t follows the Poisson distribution with:

$$D_{x,t} \sim Poisson\,(E_{x,t}\, m_{x,t}) \quad (1)$$

where $E_{x,t}$ is the number of population, $m_{x,t}$ is the mortality rate and can be described as:

$$m_{x,t} = exp(\alpha_x + \beta_x k_t), \quad (2)$$

where α_x is the average age-specific mortality, k_t tracks the time trend of mortality, $\beta_x$represents how each age group changes with k_t change. The model parameters are subject to two constraints: $\sum_x \beta_x = 1$ and $\sum_t k_t = 0$. To estimate the three model parameters,α_x,$\beta_x$and k_t, we use a maximum likelihood estimation (MLE) technique, which is given by:

$$L(\alpha_x, \beta_x, k_t) = \sum_{x,t}\left(D_{x,t}(\alpha_x + \beta_x k_t) - E_{x,t} exp\,(\alpha_x + \beta_x k_t)\right) + constant \quad (3)$$

Due to the presence of the bilinear term $\beta_x k_t$, it is not possible to derive a closed-form solution for Eq.(3). We create python functions to search for parameter values within 10,000 times and constrain the total absolute differences of three parameters between iterations less than 1e(-10). To accelerate the searching process, we use the singular value decomposition (SVD) of Eq.(2) to get initial values of the three parameters. We forecast future mortality using Eq. (2) with α_x and $\beta_x$obtained from Eq. (3) and k_t in the future years. The future k_t is modeled and predicted by ARIMA or LSTM based on k_t that is obtained from previous years using Eq.(3).

4.4 User Interface The Tableau interactive dashboard is overlaid on the US states map. The dashboard has filters that allow users to select the year, age group, causes of death, and gender to show the change of mortality rate in each category over time at the national level. When users hover the mouse over a state on the map, the state level historical and forecasted trend line will show up.

5. Experiments / Evaluation The upcoming experiments are designed to address the following question: what is the better approach in predicting $k_t$and which model performs better in predicting mortality?

Experiment 1: Compare ARIMA and LSTM in predicting k_t of the Possion Lee-Carter

The value of $k_t$in the Possion Lee-Carter model is a function of year. To predict k_t, we compare two approaches, ARIMA and LSTM. We fit 80% of historical k_t data and test the performance using the remaining 20% data. We then calculate four error metrics: root mean square error (RMSE), mean absolute error (MAE), mean percentage error (MPE), and mean absolute percentage error (MAPE) using predicted and observed k_t. Table 1 shows the calculated four errors using ARIMA and LSTM for all states, genders, ages and all causes. Figure 1 displays predicted k_t as a function of year for both training and test datasets in California, male, and all causes. RMSE of LSTM is smaller than that of ARIMA, although the other three errors are similar. We select LSTM to forecast k_t in the Possion Lee-Carter model.

Table1. Comparison of error metrics in predicting all k_t using ARMIA and LSTM for test dataset

	RMSE	MAE	MAPE	MPE
ARIMA	0.71	0.46	0.22	-0.10
LSTM	0.58	0.40	0.22	-0.13

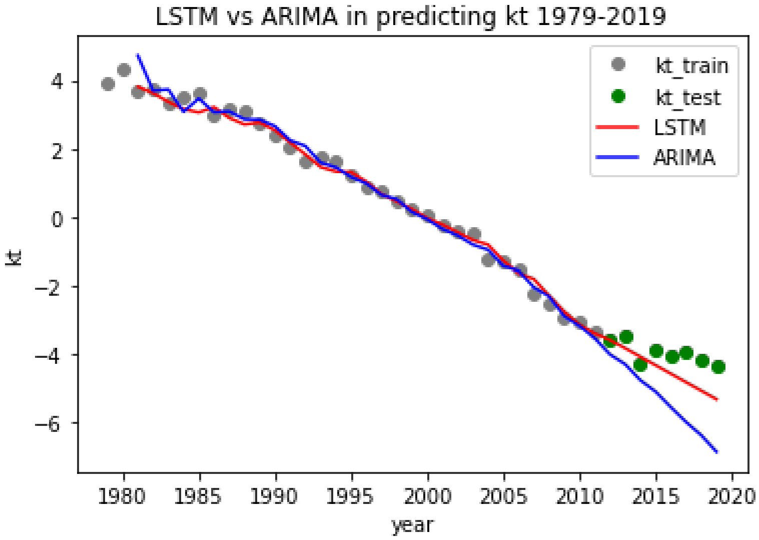


Figure 1. Comparison of lstm vs arima in predicting kt of lee-cart for mortality in California, male and all causes.

Experiment 2: Compare ARIMA, LSTM, and Poisson Lee-Carter in predicting mortality

In addition to predicting the $k_t$value of the Poisson Lee-Carter model, ARIMA and LSTM can be used for mortality prediction. After selecting LSTM for the Poisson Lee-Carter model from Experiment 1, we aim to compare it to ARIMA and LSTM models in predicting mortality rates. Similar to Experiment 1, we fit and test these models using mortality rate and compare their performance with the four error metrics. Table 2 shows LSTM has the smallest RMSE, MAE, and MAPE, while Poisson Lee Carter gives the lowest MPE. The errors from ARIMA fall between these two models.

The combination of state, gender, age, and disease gives 11526 groups. Each model needs to be trained and tested for thousands of groups. We want to select a computational efficient model which gives low errors. We therefore compare computational time for each model. LSTM needs 35 hours, about 14 times longer than ARIMA and 8 times longer than Poisson Lee Carter. Considering both error metrics and computational time, we select ARIMA as our preferred model which takes the shortest time while still giving relatively low errors.

Table 2. Comparison of error metrics and computational time in predicting mortality rate (per 100K population) of all causes using three models for test dataset

	RMSE	MAE	MAPE	MPE	Computational time
ARIMA	285.72	110.03	0.196	-0.042	2.5 h
LSTM	237.74	79.61	0.16	-0.12	35 h
Poisson Lee Carter with LSTM	308.27	127.03	0.20	0.005	4.6 h

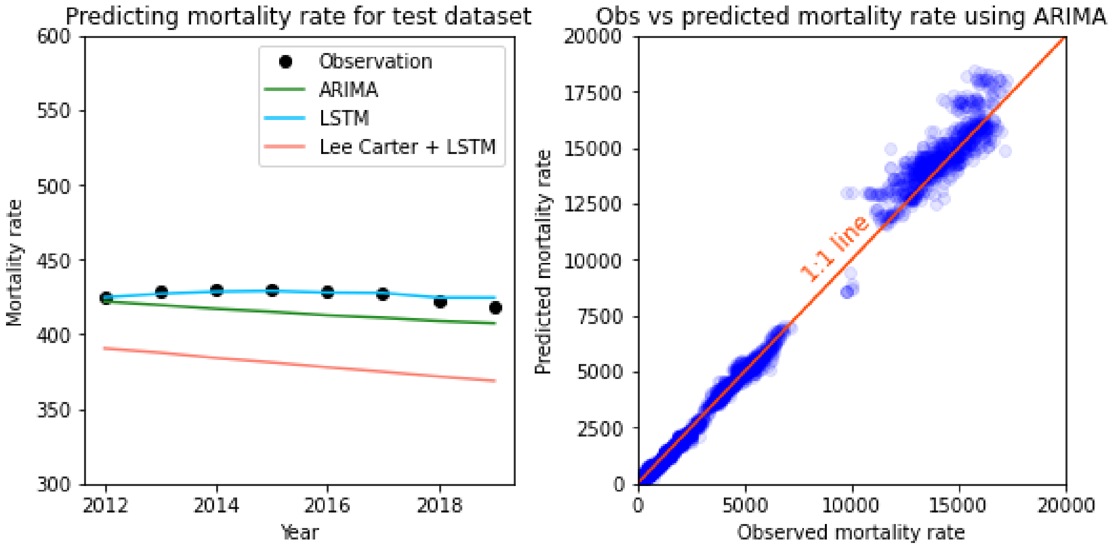
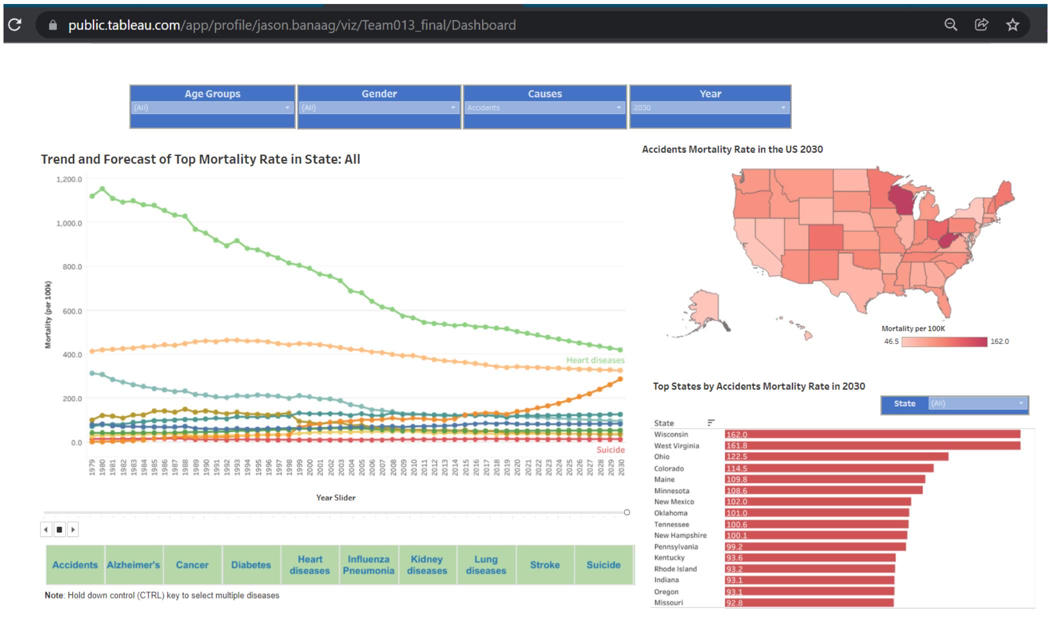


Figure 2. Left: Comparison of three models in predicting the mortality rate in all states, both gender and all causes for age group 45-54 years. Right: Observed and predicted mortality rates using ARIMA for all test datasets.

6. Discussion and Conclusions

A Tableau interactive dashboard was created and published here. Users could explore the trends, patterns and predictions of disease mortality rates from 1979-2030 at the state level (Fig. 3).



public.tableau.com/app/profile/jason.banaag/viz/Team013_final/Dashboard

Age Groups: (All)
Gender: (All)
Causes: Accidents
Year: 2030
State: (All)

Fig. 3 User Interface of the mortality dashboard.

At the national level, heart disease is the leading cause of death in the past 40 years, followed by cancer. The mortality rates of most major causes keep decreasing with time, except Alzheimer’s disease. Starting from 2019, Alzheimer’s disease emerged as the third leading cause of death and its mortality rate keeps increasing. By 2030, although cancer is still the second leading cause, Alzheimer’s disease is likely to replace cancer within a few years after 2030.

At the state level, the trend of each leading cause is similar to the national level and heart disease and cancer are the top two leading causes for many states in most years. However, in Minnesota, South Dakota, North Dakota, South Carolina, Arkansas, and Mississippi, Alzheimer’s disease will become the No.1 leading cause by 2030. In California, Colorado, and Alabama, this disease will replace cancer to be the second leading cause of death.

Our analysis implies the current policy in disease control and prevention is effective especially for major causes such as heart disease and cancer. More efforts are needed to prevent, cure or slow Alzheimer's disease and make it the next public health success story.

7. Distribution of Team Member Effort Stephanie Deen researched LSTM models and how it has been used in conjunction with the Lee-Carter model in other work. She performed coding, tuning, and evaluating the LSTM model used for the project. Min Zhong wrote python scripts and functions to implement the Poisson Lee-Carter model. She evaluated and improved the Lee-Carter model and forecast future mortality. Jason Banaag built Tableau visualization for continental US, Alaska, and Hawaii. He updated state population projection and researched animation of time series data. Mingyu Yan downloaded and wrote Python scripts to clean the 1979-1998 (ICD-9) mortality data and researched how to add Python scripts to Tableau. He also identified issues with the raw data and prepared the final processed raw data. Xiaoming Wang obtained the 1999-2019 (ICD-10) mortality data, wrote python scripts for grouped/hierarchical ARIMA prediction and performed model evaluation.

References

Bhagwat, R., Abdolahnejad, M., & Moocarme, M. (2019). *Applied Deep Learning with Keras: Solve complex real-life problems with the simplicity of Keras*. Packt Publishing Ltd.

Booth, H., & Tickle, L. (2008). Mortality Modelling and Forecasting: a Review of Methods. In *Annals of Actuarial Science* (Vol. 3, Issues 1-2, pp. 3–43). https://doi.org/10.1017/s1748499500000440

Brouhns, N., Denuit, M., & Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. In *Insurance: Mathematics and Economics* (Vol. 31, Issue 3, pp. 373–393). https://doi.org/10.1016/s0167-6687(02)00185-3

Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Kutz, M. J., Huynh, C., Barber, R. M., Shackelford, K. A., Mackenbach, J. P., van Lenthe, F. J., Flaxman, A. D., Naghavi, M., Mokdad, A. H., & Murray, C. J. L. (2016). US County-Level Trends in Mortality Rates for Major Causes of Death, 1980-2014. *JAMA: The Journal of the American Medical Association*, *316*(22), 2385–2401.

Elkins, J. S., & Claiborne Johnston, S. (2003). Thirty-Year Projections for Deaths From Ischemic Stroke in the United States. In *Stroke* (Vol. 34, Issue 9, pp. 2109–2112). https://doi.org/10.1161/01.str.0000085829.60324.de

Fischhoff, B. (2020). Making Decisions in a COVID-19 World. In *JAMA* (Vol. 324, Issue 2, p. 139). https://doi.org/10.1001/jama.2020.10178

Foreman, K. J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., Pletcher, M. A., Smith, A. E., Tang, K., Yuan, C.-W., Brown, J. C., Friedman, J., He, J., Heuton, K. R., Holmberg, M., Patel, D. J., Reidy, P., Carter, A., Cercy, K., … Murray, C. J. L. (2018). Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. *The Lancet*, *392*(10159), 2052–2090.

GBD 2017 Causes of Death Collaborators. (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1736–1788.

Kenneth D. Kochanek, Jiaquan Xu, Elizabeth Arias. (2020). *Mortality in the United States, 2019* (No. 395). Centers for Disease Control. https://www.cdc.gov/nchs/data/databriefs/db395-H.pdf

Keras Team. (n.d.). *Keras: the Python deep learning API*. Retrieved November 5, 2021, from https://keras.io

Lee, R. (2000). The Lee-Carter Method for Forecasting Mortality, with Various Extensions and Applications. In *North American Actuarial Journal* (Vol. 4, Issue 1, pp. 80–91). https://doi.org/10.1080/10920277.2000.10595882

Lilienfeld, D. E., & Perl, D. P. (1993). Projected Neurodegenerative Disease Mortality in the United States, 1990–2040. In *Neuroepidemiology* (Vol. 12, Issue 4, pp. 219–228). https://doi.org/10.1159/000110320

Lindholm, M., & Palmborg, L. (n.d.). Efficient Use of Data for LSTM Mortality Forecasting. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3805843

Marino, M., Levantesi, S., & Nigri, A. (2021). Deepening Lee-Carter for longevity projections with uncertainty estimation. In *arXiv [stat.AP]*. arXiv. http://arxiv.org/abs/2103.10535

Mboera, L. E. G., Rumisha, S. F., Lyimo, E. P., Chiduo, M. G., Mangu, C. D., Mremi, I. R., Kumalija, C. J., Joachim, C., Kishamawe, C., Massawe, I. S., Matemba, L. E., Kimario, E., Bwana, V. M., & Mkwashapi, D. M. (2018). Cause-specific mortality patterns among hospital deaths in Tanzania, 2006-2015. In *PLOS ONE* (Vol. 13, Issue 10, p. e0205833). https://doi.org/10.1371/journal.pone.0205833

Murray, C. J. L., & Lopez, A. D. (1997). Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. In *The Lancet* (Vol. 349, Issue 9064, pp. 1498–1504). https://doi.org/10.1016/s0140-6736(96)07492-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Petneházi, G., & Gáll, J. (2019). Mortality rate forecasting: can recurrent neural networks beat the Lee-Carter model? In *arXiv [q-fin.RM]*. arXiv. http://arxiv.org/abs/1909.05501

Tuljapurkar, S., & Boe, C. (1998). Mortality Change and Forecasting. In *North American Actuarial Journal* (Vol. 2, Issue 4, pp. 13–47). https://doi.org/10.1080/10920277.1998.10595752