Dissertation Project Protocol

Author: Zhengyang (Amy) Dong
zID: z5396150
Supervisors: A/Prof Katja Hanewald
Dr Andres Villegas Ramirez

Project background

Many countries have undergone an epidemiological transition in which the leading underlying causes of death have transited from infectious to chronic diseases. At the same time, insurance companies and government authorities raise more attention to predicting life expectancy and morbidity prevalence to adjust their premises and fund activities on account of population ageing and the surging demand for long-term care (Okma & Gusmano, 2020).

A growing literature focuses on mortality prediction using traditional statistical methods and machine learning techniques such as neural networks. Some widely used methods, e.g., the Lee-Carter model (Lee & Carter, 1992), have been extended to complex models with explanatory variables (Booth & Tickle, 2008; Hunt & Blake, 2021). Hong et al. (2021) and Wang et al. (2022) introduced the Lee-Carter-Artificial Neural Network and the Generalised Linear Model-Neural Networks to predict mortality rates and health transition intensities.

The multiple causes of death data, which reports all diseases listed on the death certificates, has been used for providing a comprehensive analysis of the burden of individual diseases (Barco et al., 2021) and enhancing drug use surveillance (Nordstrom et al., 2013). However, it has not yet been fully used in mortality prediction. Wall et al. (2005) argued that underlying causes of death were insufficient to represent morbidity and were biased due to demographic features, socioeconomic status, and deceased places. To address the challenges, the World Health Organization (WHO) recommended that countries unify death certificates and generate systematic methods to define the underlying causes and contributing causes (Becker et al., 2006). Medical partitioners, including medical certifiers, nurses and coroners, were trained to report the related diseases on death certificates using International Classification Diseases coding (ICD). Cause-specific mortality explains more epidemiological details related to socioeconomic status (Alai et al., 2018), increases the forecast accuracy (Li et al., 2019), and identifies the mortality drivers (Villegas et al., 2021). Moreno-Betancur et al. (2017) assigned

different weights to the Cox regression model and extrapolated those multiple causes of death in the fitting model to outperform the underlying causes.

Aim

This project aims to explore the methods of using multiple causes of death to predict mortality. The project will focus on data processing, exploratory data analysis and visualisation to demonstrate the validity of using multiple causes of death.

This project will quantify and describe mortality using both underlying causes of death and multiple causes of death and examine how disease prevalence has changed over the past decades. We will explore the disease prevalence trend across age, race, and gender over a half-century. We will track the patterns of the most common associated diseases (those recorded in death certificates rather than underlying causes). We will also develop a data visualisation website using the Shiny app (Version 1.7.4.9002). Mortality prediction will be left for future research due to time limitations.

Methods

We obtained the U.S. mortality data from 1968 to 2020 from the National Bureau of Economic Research (NBER) and will explore trends in mortality rates, underlying causes of death and multiple causes of death by gender, race, and age using heatmaps and parallel coordinates plots. More specifically, we use records axis codes in the datasets, which describe the overall conditions in the mortality dataset (Wall et al., 2005). Considering different versions of ICD codes, we will map the codes into disease categories and perform analysis based on that. As mentioned, the causes of death are coded according to an international standard that undergoes revision and establishment in various revisions.

ICD Revision	ICD coming into effect	ICD used in the U.S. death certificate
8th Revision	1 Jan 1968	1968~1978
9th Revision	1 Jan 1979	1979~1998
10th Revision	1 Jan 1993	1999~2020

Previous research has highlighted concerns that using the new version of ICD codes led to discontinuous time series data due to updating the ICD coding system worldwide (Hsu et al., 2021). This can bias mortality and disease prevalence estimates, particularly for conditions such as diabetes, cardiovascular disease, and cerebrovascular diseases (Joyner-Grantham et al., 2010). We find that previously proposed mapping methods between different versions of ICD codes, including Clinical Classification Software (CCS) and General Equivalent Mappings (GEMs), are not suitable for this project because they failed to map among ICD 8^{th}, 9^{th} and 10^{th} versions (Fung et al., 2016; Hamad et al., 2021). To address this issue, we group ICD codes into 20 broad groupings and then extend to 92 diseases based on the Society of Actuaries matching different versions of ICD codes to specific diseases of the Human Mortality Database (Barbieri, 2017).

Additionally, we will tabulate whether other factors, such as autopsy, education level and place of death, influence the reporting of diseases on death certificates. To describe the prevalence of morbidity, we will include age-standardised rates and use the ratio (diseases mentioned in death certificates or recognised as underlying causes of death) to identify conditions that are likely to be underestimated or neglected (Australian Institute of Health and Welfare, 2017).

Data Management Plan

Data sources

This project uses Multiple Cause-of-Death Mortality data from 1968 to 2020 from the National Bureau of Economic Research (NBER). The dataset includes demographic features, causes of death, place of death and the autopsy occurrence.

Analysis

To clean the data and select variables, we will use R studio (Version 2022.12.0) and packages including ggplot2 (Version 3.4.1), tidyverse (Version 1.3.2), dplyr (Version 1.1.0). We will also develop an online interactive website based on Shiny (Version 1.7.4), providing data visualisation of the results and descriptive analysis. Users will be able to manipulate the data for specific diseases and periods.

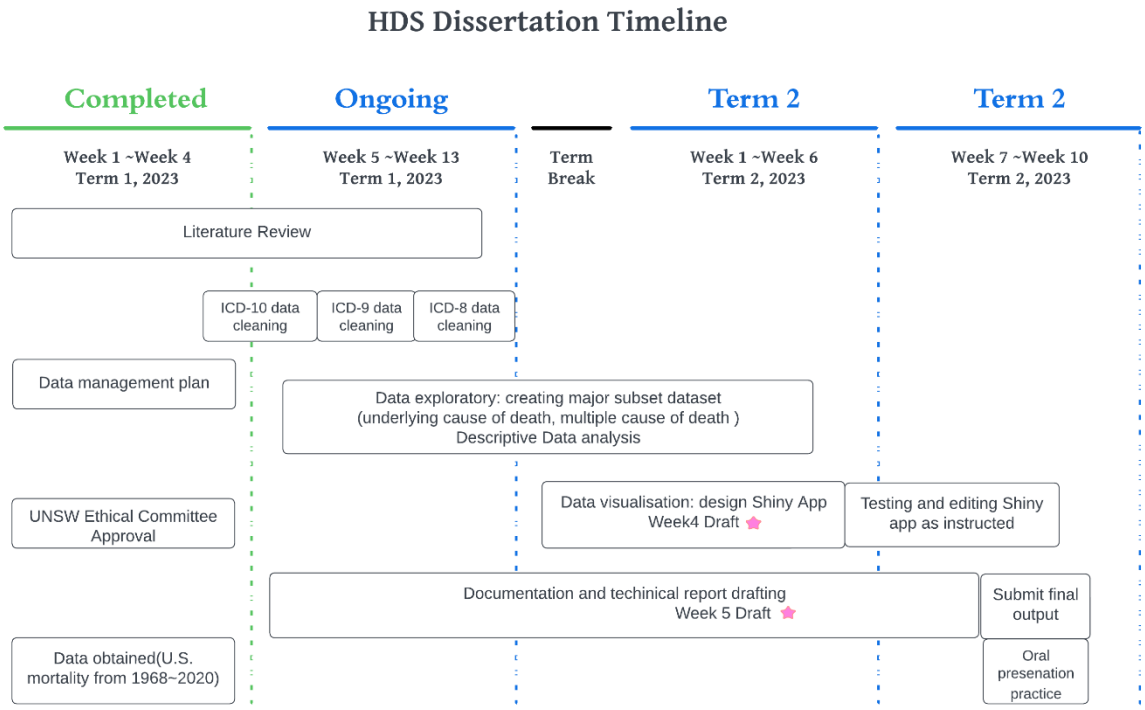
Ethics approval

The ethics clearance was approved as the negligible risk project by UNSW HREAP Executive Panel (HC2100948).

Data storage and retention

The processed data will be stored on the author's desktop PC and UNSW OneDrive Cloud. The raw data will not upload to Github due to the excessive data volume. The raw datasets are available at the NBER website (https://www.nber.org/research/data/mortality-data-vital-statistics-nchs-multiple-cause-death-data). Shiny app codes, and data processing codes will be made available on GitHub.

Timeline



Reference List

Adair, T., Temple, J., Anstey, K. J., & Lopez, A. D. (2022). Is the Rise in Reported Dementia Mortality Real? Analysis of Multiple-Cause-of-Death Data for Australia and the United States. American Journal of Epidemiology, 191(7), 1270–1279.

Alai, D. H., Arnold (Gaille), S., Bajekal, M., & Villegas, A. M. (2018). Mind the Gap: A Study of Cause-Specific Mortality by Socioeconomic Circumstances. *North American Actuarial Journal*, *22*(2), 161–181.

Australian Institute of Health and Welfare. (2017) Multiple causes of death in Australia: an analysis of all natural and selected chronic disease causes of death 1997-2007. https://www.aihw.gov.au/getmedia/2d1815a5-9171-4cf4-8007-c78dbc16b45b/14157.pdf.aspx?inline=true

Barbieri, M. (2017). Expanding the Human Mortality Database to include cause-of-death information. *Society of Actuaries*. https://www.soa.org/globalassets/assets/Files/Research/Projects/2017-hmd-cause-of-death.pdf

Barco, S., Valerio, L., Ageno, W., Cohen, A. T., Goldhaber, S. Z., Hunt, B. J., ... & Konstantinides, S. V. (2021). Age-sex specific pulmonary embolism-related mortality in the USA and Canada, 2000–18: an analysis of the WHO Mortality Database and of the CDC Multiple Cause of Death database. *The Lancet Respiratory Medicine*, *9*(1), 33-42.

Becker, R., Silvi, J., Ma Fat, D., L'Hours, A., & Laurenti, R. (2006). A method for deriving leading causes of death. Bulletin of the World Health Organization, 84(4), 297-304.

Booth, H., & Tickle, L. (2008). Mortality Modelling and Forecasting: a Review of Methods. *Annals of Actuarial Science*, *3*(1–2), 3–43.

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2022). _shiny: Web Application Framework for R_. R package version 1.7.4, https://CRAN.R-project.org/package=shiny

Fung, K. W., Richesson, R., Smerek, M., Pereira, K. C., Green, B. B., Patkar, A., Clowse, M., Bauck, A., & Bodenreider, O. (2016). Preparing for the ICD-10-CM Transition: Automated Methods for Translating ICD Codes in Clinical Phenotype Definitions. *EGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *4*(1), 4.

Hamad, A. F., Vasylkiv, V., Yan, L., Sanusi, R., Ayilara, O., Delaney, J. A., Wall-Wieler, E., Jozani, M. J., Hu, P., Banerji, S., & Lix, L. M. (2021). Mapping three versions of the international classification of diseases to categories of chronic conditions. *International Journal of Population Data Science*, *6*(1).

Hsu, M. C., Wang, C. C., Huang, L. Y., Lin, C. Y., Lin, F. J., & Toh, S. (2021). Effect of ICD-9-CM to ICD-10-CM coding system transition on identification of common

conditions: An interrupted time series analysis. *Pharmacoepidemiology and Drug Safety*, *30*(12), 1653–1674.

Hunt, A., & Blake, D. (2021). On the Structure and Classification of Mortality Models. *North American Actuarial Journal*, *25*(S1), S215–S234.

Joyner-Grantham, J. N., Simmons, D. R., Moore, M. A., & Ferrario, C. M. (2010). The impact of changing ICD code on hypertension-related mortality in the southeastern United States from 1994-2005. *Journal of Clinical Hypertension*, *12*(3), 213–222.

Li, H., Li, H., Lu, Y., & Panagiotelis, A. (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance: Mathematics and Economics*, *86*, 122–133.

Moreno-Betancur, M., Sadaoui, H., Piffaretti, C., & Rey, G. (2017). Survival analysis with multiple causes of death. *Epidemiology*, *28*(1), 12–19.

Natioanl Bureau of Economic Research. (n.d.). Mortality Data - Vital Statistics NCHS Multiple Cause of Death Data. https://www.nber.org/research/data/mortality-data-vital-statistics-nchs-multiple-cause-death-data

Nordstrom, D. L., Yokoi-Shelton, M. L., & Zosel, A. (2013). Using multiple cause-of-death data to improve surveillance of drug-related mortality. *Journal of public health management and practice: JPHMP*, *19*(5), 402.

Okma, K., & Gusmano, M. K. (2020). Aging, pensions and long-term care: What, why, who, how?: Comment on “financing long-term care: Lessons from Japan.” In *International Journal of Health Policy and Management* (Vol. 9, Issue 5, pp. 218–221). Kerman University of Medical Sciences.

Villegas, A.M., Bajekal, M., Haberman, S., & Zhou, L. (2021). Analysis of Historical U.S. Population Mortality Improvement Drivers 1959–2016. *Society of Actuaries.* https://www.soa.org/globalassets/assets/files/resources/research-report/2021/2021-historical-us- population-mortality-improvement-drivers.pdf