A Novel Approach for Longitudinal Modeling of Aging Health and Predicting Mortality Rates

Hannah Guan
Harvard College
hguan000@gmail.com

Abstract

Aging is a complex stochastic process that affects healthy functioning through various pathways. In contrast to the more commonly used cross-sectional methods, our research focuses on longitudinal modeling of aging, a less explored but crucial area. We have developed a Stochastic Differential Equation (SDE) model, at the forefront of aging research, designed to accurately forecast the health trajectories and survival rates of individuals. This model adeptly delineates the connections between different health indicators and provides clear, interpretable results. Our approach utilizes the SDE framework to encapsulate the inherent uncertainty in the aging process. Moreover, it incorporates a Recurrent Neural Network (RNN) to integrate past health data into future health projections. We plan to train and test our model using a comprehensive dataset tailored for aging studies. This model is not only computationally cost-effective but also highly relevant in assessing health risks in older populations, particularly for those at high risk. It can serve as an essential tool in anticipating and preparing for challenges like infectious disease outbreaks. Overall, our research aims to improve health equity and global health security significantly, offering substantial benefits to public health and deepening our understanding of the aging process.

Introduction

Aging stands as the primary risk factor for various diseases, including cardiovascular ailments, cancer, type 2 diabetes, hypertension, and Alzheimer's disease (Guan 2021, 2022; Guan et al. 2022). Several factors influence the aging process, such as cholesterol levels, walking capabilities, grip strength of the dominant hand, and hemoglobin levels (Miller et al. 2017). Relationships have been observed between aging and factors like diuretic use, sodium levels, and walking ability (Miller et al. 2017). Additionally, different age groups show varying associations between HDL Cholesterol, Triglycerides, Glucose levels, and Alzheimer’s Disease (Zhang et al. 2022). The complexity of aging is evident in the multitude of variables affecting health states, underscoring its high-dimensional nature (Farrell et al. 2022).

Our study introduces a Stochastic Differential Equation (SDE) model to accurately depict an individual's health progression and likelihood of survival at any given time. Current models, which employ either weight vectors (Lopez-Pacheco et al. 2022; Somers et al. 2009) or pairwise interaction matrices (Somers et al. 2009), tend to suffer from low predictive accuracy. The SDE approach is a burgeoning field in aging research. While existing frameworks (Gladyshev et al. 2016; Farrell et al. 2020) have made strides, they are limited to modeling a small number of health variables or are restricted to binary values. Our model's discrete system architecture allows each component to contribute to the loss function, differing from existing models that rely solely on hidden weights. This approach offers more precise control over the loss function, enhancing the training process and allowing for more accurate predictions.

Background

Neural networks are frequently used to model complex, non-linear relationships, like aging, due to their capacity to handle numerous health variables as discussed in this study (Somers et al. 200; Buhrmester et al. 2021). However, they often lack clear interpretability, functioning as 'black boxes' (Buhrmester et al. 2021; Zhang et al. 2021). Enhancing their interpretability could lead to identifying biases in predictions (Tan 2018), uncovering key features (Yap et al. 2021), detecting erroneous correlations, and developing more dependable models.

In aging research, stochastic models have been used, and joint models suggested, but these typically focus on limited variables (Stukalin et al. 2013; Gladyshev et al. 2016; Zhbannikov et al. 2017). The Weighted Network Model (WNM) creates trajectories for 10 'deficits' and forecasts survival (Zhbannikov et al. 2017), yet it simplifies these 'deficits' to binary variables, reducing its complexity. The Joint Model (JM) presents a combined framework for analyzing both longitudinal and survival data dynamics (Lopez-Pacheco et al. 2022). Additionally, the Stochastic Process Model (SPM) employs a Stochastic Differential Equation

(SDE) for dynamic analysis (Somers et al. 2009). The Dynamic Joint Interpretable Network (DJIN) model (Farrell 2022) aims to explore pairwise interactions among various health variables, furthering this line of research.

Approach

Our approach is centered around a comprehensive longitudinal analysis conducted over multiple follow-up periods within a large-scale cohort study. In this longitudinal framework, we observe the same set of participants across several intervals, allowing for an in-depth examination and continuous observation of each individual's health state over an extended period. This approach is particularly advantageous compared to cross-sectional studies, which provide a snapshot of health data from a population at a single point in time. By contrast, longitudinal data enhances statistical power, facilitates more accurate estimation of temporal changes in health, and allows for a more nuanced understanding of health trajectories.

To interpret the complexities inherent in this data, our model utilizes a sophisticated three-dimensional interaction network. This network effectively captures and illustrates the intricate relationships and strengths of connections between various health variables. To address the inherent uncertainty and variability in the aging process, we have employed a Stochastic Differential Equation (SDE). Additionally, we incorporate a Recurrent Neural Network (RNN) to model mortality, allowing the historical health data of individuals to inform and influence future health predictions. To manage instances of incomplete health data, a common challenge in longitudinal studies, we apply a missing mask to the observed health variables and use a Variational Autoencoder for the imputation of missing data, ensuring a comprehensive and accurate dataset.

For the empirical aspect of our study, we have chosen the English Longitudinal Study of Aging (ELSA) dataset as our primary data source. ELSA provides a rich longitudinal dataset encompassing nine waves of data collection over a 20-year period from 1998 to 2020. The initial baseline data was sourced from the Health Survey for England conducted between 1998 and 2001. Comprising a total of 27,365 individuals, this dataset effectively represents the English population, particularly those over the age of 50, making it highly suitable for our aging study. ELSA collects self-reported health information every two years and nurse-reported data approximately every four years, as evidenced in waves 2, 4, 6, 8, and 9. Leveraging this extensive dataset, our goal is to predict the health states and survival rates of individuals over time, providing valuable insights into the aging process and its implications for public health.

Evaluation

Our model's effectiveness will be assessed using various statistical measures. The C-Index (Giunchiglia et al. 2018), a metric indicating the likelihood of the model accurately predicting which individual among a pair will live longer, will be a key tool. A C-index of 1 signals perfect predictions, while 0.5 suggests random accuracy. We'll assess the model's predictions across all ages and specific age groups to ensure it's not overly reliant on age as a predictive factor. The Brier Score (Heller et al. 2021) will measure the accuracy of our survival curve predictions, with lower scores indicating higher accuracy and a score of zero representing perfect alignment with the actual survival curve. D-Calibration (Haider et al. 2020) will evaluate the distribution of survival predictions, aiming for a uniform distribution as a sign of model calibration. This will involve dividing predictions into ten equal segments from 0 to 1 and comparing the distribution of these predictions to an ideal scenario where each segment comprises 10% of total predictions. We'll employ Pearson's $\chi 2$ test to further assess the distribution quality, with lower $\chi 2$ and higher p-values indicating better calibration. Finally, we will calculate and plot Relative Root Mean Square Error (RMSE) scores (Chai and Draxler 2014) for each health variable up to six years' post-baseline, and the total relative RMSE over time, confirming the model's short-term predictive accuracy and its sustained accuracy for up to 14 years' post-baseline.

Discussion

One possible limitation of our model comes from selective attrition in the ELSA dataset. Participants would drop out of the survey, shrinking the sample size and decreasing the amount of data collected. Therefore, the final group may not reflect the original representative sample, and attrition may affect the experiment's validity. Other limitations of this research included validating the model in a clinical setting. Due to resource limitations, all tests will be based on the public dataset. Broader users (such as doctors and clinical practitioners) and software developers may have different views toward conceptual frameworks and concepts of modeling and simulation. This difference may significantly affect the design of the model and the interpretation of the result. In the future, we hope to collaborate with the clinical and industrial sectors on software's successful development and validation.

Conclusion

Doctors can use our model with basic, easy-to-access lab measures to predict the onset of age-related diseases like Alzheimer's and cardiovascular disease in their patients, which would significantly increase the quality and longevity of life across the grid. Further investigation into the neural network's inner processes would enable us to better understand the root causes of the aging process. We could use this knowledge to develop anti-aging medications and cures for different age-related diseases.

References

Buhrmester, V.; Münch, D.; Arens, M. 2021. Analysis of explainers of black box deep neural networks for computer vision: a survey. *Machine Learning and Knowledge Extraction* 3(4): 966-989.

Chai, T.; Draxler, R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? - arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7: 1247–1250.

Farrell, S.; Mitnitski, A.; Rockwood, K.; et al. 2020. Generating synthetic aging trajectories with a weighted network model using cross-sectional data. Scientific Reports 10: 19833.

Farrell, S.; Mitnitski, A.; Rockwood, K.; Rutenberg, A.D. 2022. Interpretable machine learning for high-dimensional trajectories of aging health. *PLoS Computational Biology* 18(1): e1009746.

Giunchiglia, E.; Nemchenko, A.; van der Schaar, M. 2018. RNN-SURV: a deep recurrent model for survival analysis. In: Kurkov´a, V.; Manolopoulos, Y.; Hammer, B.; Iliadis, L.; Maglogiannis, I. (eds) *Artificial Neural Networks and Machine Learning, Lecture Notes in Computer Science*, 11141.

Gladyshev, V.N. 2016. Aging: progressive decline in fitness due to the rising deleteriome adjusted by genetic, environmental, and stochastic processes. *Aging Cell* 15(4): 594-602.

Guan, H. 2021. Genetic prediction of biological age: exploring the relationship between epigenetic markers and all-cause mortality, *Biomedical Journal of Scientific & Technical Research* 34(2): 26546-26552.

Guan, H. 2021. The genetics of human aging: predicting age and age-related diseases by deep mining high dimensional biomarker data, *12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* 79.

Guan, H. 2021, Effect of fish oil on heart health: a meta-analysis, *Biomedical Journal of Scientific & Technical Research* 34(3): 26701-26705.

Guan, H.; Zhang, C. 2022. Predicting diabetes in imbalanced datasets using neural networks, *13th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 57: 1-10.

Haider, H.; Hoehn, B.; Davis, S.; Greiner, R. 2020. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research* 21:1–63.

Heller, G.; 2021. The added value of new covariates to the brier score in cox survival models. *Lifetime Data Analysis* 27, 1–14.

Li, L.; Zhang, C.; Liu, S.; Guan, H.; Zhang, Y.; 2022. Age prediction by DNA methylation in neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19(3): 1393-1402.

Li, L.; Zhang, C.; Guan, H.; Zhang, Y.; 2022. Application of correlation pre-filtering neural network to DNA methylation data: biological aging prediction, *Epigenome-Wide Association Studies: Methods and Protocols*, *Springer Nature* 2432: 201-210.

Lopez-Pacheco, M.; Yu, W.; 2022. Complex valued deep neural networks for nonlinear system modeling. *Neural Processing Letters* 54: 559–580.

Miller, A.J.; Theou, O.; McMillan, M.; Howlett, S.E.; Tennankore, K.K.; Rockwood, K. 2017. Dysnatremia in relation to frailty and age in community-dwelling adults in the national health and nutrition examination survey. Journals of Gerontology Series A 72(3): 376–381.

Somers, M.J.; Casal, J.C. 2009. Using artificial neural networks to model nonlinearity: the case of the job satisfaction-job performance relationship. *Organizational Research Methods* 12(3): 403-417.

Stukalin, E. B.; Aifuwa, I.; Kim, J. S.; Wirtz, D.; Sun, S. X. 2013. Age dependent stochastic models for understanding population fluctuations in continuously cultured cells, *Journal of The Royal Society Interface* 10: 20130325.

Tan, S. 2018. Interpretable approaches to detect bias in black-box models. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* 382-383.

Yap, M.; Johnston, R.L.; Foley, H.; MacDonald, S.; Kondrashova, O.; Tran, K.A.; Nones, K.; Koufariotis, L.T.; Bean, C.; Pearson, J.V.; et al. 2021. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Scientific Reports* 11(1): 2641.

Zhang, X.; Tong, T.; Chang, A.; et al. 2022. Midlife lipid and glucose levels are associated with Alzheimer's disease. *Alzheimer's Demetiaa* 1-13.

Zhang, Y.; Tino, P.; Leonardis, A.; Tang, K. 2021. A survey on neural network interpretability, *IEEE Transactions on Emerging Topics in Computational Intelligence* 5(5): 726-742.

Zhbannikov, I.Y.; Arbeev, K.; Akushevich, I.; Stallard, E.; Yashin, A.I. 2017. stpm: An R Package for Stochastic Process Model. *BMC Bioinformatics* 18(125).