9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Death Predictions:
Holistically and Precisely with Explainable AI

Keywords: Machine Learning, SHAP, Predicting Ceiling, Seed Variability, Social Determinants of Health

Extended Abstract

In both sociological and biological research, the majority of the work which attempts to understand death at the individual level concentrates on a single-topic perspective: it is surprisingly understudied in a holistic approach that integrates information from different disciplines. Pragmatically, we know virtually nothing about how predictable death is. We devote to exploring the existing frontier of the field with data from large longitudinal study of ageing, the Health and Retirement study in the US (HRS), which specifically look at the ageing population who are over 50 years old.

We also develop new methodological concepts useful for the integration of explanation and prediction in the social sciences. Specifically, we construct a predictive research design which allows us to accurately consider how predictable death is across multiple domains which we uniquely re-engineer. We extract information on 61 risk factors from seven health-related domains including Adulthood Psychological Diathesis, Adulthood Socioeconomic, Childhood Adversity, Adulthood Adverse Experience, Health Behaviours, Social Connections and Demography. Risk factors are selected on the ground of sociological theories such as fetal conditions (Barker, 1995) and fundamental causes (Link and Phelan, 1995), contending the strength of social determinants of health. In addition to HRS, multiple ageing studies with similar infrastructure in different regions, such as the Survey of Health, Ageing and Retirement in Europe (SHARE) and Japanese Study of Aging and Retirement(JSTAR) are under consideration, which enrich information and further allow cross-country comparisons.

Prediction accuracy is powered and guaranteed by advanced models like eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), with the potential expansion to SuperLearner, all of which are art-of-state and potent predicting algorithms. We then 'unravel the black box' of prediction via the use of Shapley values, allowing us to better 'surface' variables at the single risk factor level. The classical Shapley Value for a feature i is calculated as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right] \tag{1}$$

where N is the set of all input features (the 61 risk factors in our study, or 66 when incorporating the biological domain), S is a non-zero subset of those input features, M is the number of input features (61 or 66) and $f_x(S)$ is the prediction for feature in set S. By this definition, the Shapley Value is the weighted and summed contribution over all possible feature value combinations. Besides, we devise a domain-level contribution computing algorithm to facilitate the understanding of risk factors' relative importance.

Apart from the explainable cross-discipline research design which contributes to a comprehensive understanding of death predictability, we also devote to understanding the prediction precision, the more methodological and technical aspect of prediction, through three ap-

proaches. Firstly, We develop the foundations of tools which allow us to consider the 'asymptotics of predictive power', estimating the impact of varying training dataset size on predictability. Secondly, via two types of time windows-rolling window and recursive window, we compare the predictive accuracy with different temporal infrastructures, which reveals the death predictivity momentum trajectory of risk factors and provides the 'optimal' predictive timestamp in our study. Finally, we introduce a new heuristic – the median of a large number of trials of seeds – which allows us to account for the randomness inherent in manifold works. Our conclusion is that death is surprisingly predictable with minimal feature engineering, and that with larger datasets we very well could eliminate reducible error and reach the 'predictive ceiling'.

Result 1: Even with an unbalanced dataset, death is highly predictable with a comprehensive predicting framework, where the ROC-AUC, PR-AUC and F1 scores in the best model (with LightGBM) are 0.825,0.687 and 0.585 respectively. Competing with all 61 risk factors, the top eight influential risk factors (with mean absolute SHAP values bigger than 0.1) are Age, Male, History of Smoking, Low/no Vigorous Activity, Income, Trait Anxiety, Wealth, and Lower Occupational Status. Results are illustrated in Figure 1.

Result 2: The top three influential domains are identical and consistent across all performance evaluation metrics where Biomarkers are substantiated with the highest contribution, followed by Health Behaviours and Socioeconomic.

Result 3: In the temporal structure, only the recursive window design presents consistent evidence that the model performance steadily increases with the growing window size, which can be imputed to the enlarged true label group, as deaths spread over the whole time horizon. Although there is no clear conclusion from the rolling window design, we can still infer that the predictivity decreases along with the 'data freshness', which is the time gap between prediction and data collection.

Result 4: Persistent and consistent evidence of the effect of dataset size on model performance is observed across all evaluation metrics: models trained with larger training datasets outperform other models with smaller datasets when predicting the same out-of-sample set, facilitating the understanding of reducible error in machine learning predictions.

Result 5: Prediction accuracy varies greatly with different choices of seed in the train-test set split. Using LightGBM, the prediction accuracy for each evaluating metric spreads over about three standard deviations away from the mean value to each side (i.e. metric span is circa six times s.d). For example, the mean, max, min and std of F1 score across 10000 different seeds in train-test splitting is 0.560, 0.599, 0.512 and 0.011 respectively.

Albeit a normal-like distribution, we call for attention to the span as the choice of seed is completely subjectively random and the resulting single outcome will be an arbitrary point within the span and thus affecting the prediction precision substantially, rather than the accuracy.

References

Barker, D. J. (1995). Fetal origins of coronary heart disease. *Bmj*, *311*(6998), 171–174.

Link, B. G., & Phelan, J. (1995). Social conditions as fundamental causes of disease. *Journal of health and social behavior*, 80–94.

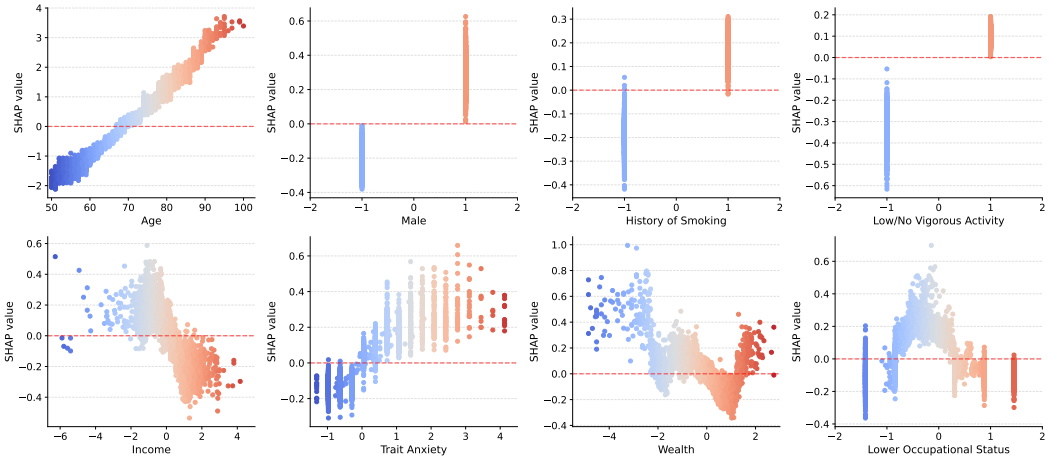

Figure 1: SHAP for the top eight influential predictors