Revisiting the social determinants of health with explainable AI: a cross-country perspective

Jiani Yan [*1,2,3,4]

[1]Leverhulme Centre for Demographic Science, University of Oxford
[2]Department of Sociology, University of Oxford
[3]Wolfson College, University of Oxford
[4]Max Plank Institute of Demographic Research

Abstract

In social science and epidemiological research, individual risk factors for mortality are often examined in isolation, while approaches that consider multiple risk factors simultaneously remain less common. Using the Health and Retirement Study in the US, the Survey of Health, Ageing and Retirement in Europe, and the English Longitudinal Study of Ageing in the UK, we explore the predictability of death with machine learning and explainable AI algorithms, which integrate explanation and prediction simultaneously. Specifically, we extract information from all datasets in seven health-related domains, including demographic, socioeconomic, psychology, social connections, childhood adversity, adulthood adversity, and health behaviours. Our self-devised algorithm reveals consistent domain-level patterns across datasets, with demography and socioeconomic factors being the most significant. However, at the individual risk-factor level, notable differences emerge, emphasising the context-specific nature of certain predictors.

Keywords: Social Determinants of Health, Cross-National Comparisons, Machine Learning, Explainable AI, Predictive Precision.

*__For Correspondence:__ Jiani Yan, Email: jiani.yan@sociology.ox.ac.uk. **Code Availability**: Codes can be found at GitHub: https://github.com/vallerrr/Mortality_prediction_cross_countries. Please see the readme.md file within that repository for a data availability statement. **Acknowledgements**: Funding is gratefully acknowledged from the Leverhulme Research Centres Grant (grant RC-2018-003) for the Leverhulme Centre for Demographic Science and the Economic and Social Research Council Grand Union Doctoral Training Partnership is gratefully acknowledged. We are grateful for comments on earlier versions of the work from participants at the International Conference on Computational Social Science, European Population Conference, Population Association of America, and internal feedback received from the Leverhulme Centre for Demographic Science, Office of Population Research, Helsinki Institute for Demography and Population Health and Max Plank Institute of Demographic Research. Special thanks to C. Rahal and R. Kashyap for their invaluable guidance and supervision throughout the development of this work, and to C.F. Breen for their insightful feedback.

Introduction

The contention that death is not merely biologically determined has increasingly been embraced by recent research [1]. Specifically, social epidemiology emphasises the distribution of social determinants of health [2]. Manifold literature substantiates the significance of social determinants of health. Theoretically, studies such as Link and Phelan [3] present the Fundamental Cause Theory, positing that health is associated with multiple social factors including wealth and education. Socioeconomic status, for example, influences health through various mechanisms such as access to nutritious food, adopting healthy lifestyles, and the availability of medical resources. This relationship is complicated and pervasive [4].

Lorant et al. [5] conducted a meta-analysis to assess the magnitude, shape and modifiers of the association between socioeconomic status and depression, finding compelling evidence for socioeconomic inequality in depression. Stafford and Marmot [6] examined the impact of neighbourhood socioeconomic status on individual health outcomes, revealing that poorer individuals who live in deprived neighbourhoods experience the most significant negative health effects. Additional evidence has been accrued regarding other social factors such as social relationships [7], mental health [8], and childhood adversity [9].

Nevertheless, social determinants of health are inherently multidisciplinary, necessitating a holistic study design. For example, the biopsychosocial model proposed by Engel [10] underscores the interplay among social, biological, and psychological factors, where all of these factors dynamically interact with each other and continuously affect health outcomes. Moreover, intersectionality theory specifically highlights that disadvantages affecting life outcomes (including health) accumulate from multiple sources that are not independent of each other [11]. Although the theory mainly focuses on interactions of individual identities, its emphasis on a holistic research framework facilitates understanding of the complicated, multidimensional nature of health inequalities.

While substantial evidence demonstrates the statistically significant association between theoretically proven risk factors and mortality, only a few studies have ventured beyond this explanatory framework to adopt an integrated approach alongside advanced methods. Puterman

et al. [12] compared the contribution of 57 economic, behavioural, social, and psychological factors in predicting death within the US context using multivariate Cox regressions, LASSO, and random forests. Similarly, Breen and Seltzer [13] applied machine learning algorithms to predict life expectancy based on early adulthood socioeconomic and demographic information using US administrative data. In a different context, Vabalas et al. [14] employed Finnish registry data to develop a deep-learning algorithm for modelling one-year mortality, incorporating over 8,000 features from the domains of medical care, socioeconomic status, and demography. While these studies achieved varying levels of predictive performance, they are all limited to modelling within a single country's perspective.

Given that health outcomes reflect the effects of health policy, medical conditions and cultural differences [15], comparing the impact of established risk factors across different countries can provide valuable insights for improving health inequalities regionally. For example, Raghupathi and Raghupathi [16] explored the association between adult education level—specifically upper secondary education—and cancer mortality in OECD countries by continent. While the association indicates that a higher proportion of upper secondary education correlates with fewer cancer deaths across all continents, the strength of this association varies, with North America exhibiting the strongest correlation.

Consequently, we propose a holistic approach to examine death, as one of the most definitive factors linked with health and its most notable downstream consequences. Here, 'holistic' refers to a comprehensive, multidisciplinary research framework that incorporates a wealth of health-related information. However, holism should not be misconstrued as focusing solely on the quantity of risk factors. Instead, it ought to be conceived in terms of the relevance and utility of information to the research topic [17]. Therefore, in our research, we elect to re-examine the risk factors from the work of Puterman et al. [12], which encompasses information from seven different domains: demography, socioeconomics, psychology, social connections, childhood adversity, adulthood adversity, and health behaviours. This approach not only enhances the credibility of the results and potentially improves prediction accuracy, but also offers the opportunity to compare the relative contributions of factors from different domains. To foster

cross-country comparisons, we employ data from three different data sources: the US Health and Retirement Study (HRS), the English Longitudinal Study of Ageing (ELSA), and the Survey of Health, Ageing and Retirement in Europe (SHARE). All of these are sampled from developed countries and are part of the Gateway to Global Ageing Data, which ensures a similar data structure, facilitating meaningful comparisons.

Methodologically, we adopt a predictive framework as most theories are developed based on internal processes, leaving their external predictive validity rarely tested. Additionally, considering the need to bridge explanation and prediction, we propose an innovative research paradigm to estimate the risk factors of death. This involves training a high-accuracy predictive model using advanced machine learning algorithms and deconstructing the results with explainable artificial intelligence (XAI) methods, such as Shapley Values [18], to present an explainable framework. This approach remains comparatively unexplored, and it helps to unpack predictability and expose inequalities and biases residing in social systems by revealing the impact and direction of predictors, allowing further investigation, scientific discovery, and improved policy-making.

Methods

Data

The comprehensive design of the HRS, SHARE, and ELSA fosters a holistic investigation into risk factors influencing death probabilities. We have selected the HRS as the primary dataset for calibration. All risk factors are harmonised and comparable across different datasets (See Figure S1 for details of the sampling procedure and data sources for all datasets and Figure S2 for data description and risk factor availability). The study focuses on individuals aged over 50, with summary information for the three main datasets presented in Table 1 (see Table S1 for combined datasets). See Table S2 and S3 for the coverage of risk factors by dataset and domain.

In this study, we define the outcome variable death as a binary indicator, where 1 denotes

Table 1: Overview of Three Mortality Datasets

Dataset	HRS	SHARE	ELSA
Age range	50-100	50-99	50-97
Female portion	0.592	0.545	0.553
Risk factor number	61	25	25
Sample size	13210	17818	8389
Death prevalence	0.316	0.214	0.169
Prediction window	2008 - 2019	2006 - 2021	2005 - 2012

death and 0 indicates survival until the end of the prediction window. We aim to predict the probability of death within a specified future time frame. Each dataset employs distinct sampling and prediction windows, as illustrated in Figure 1. The 'Survey' windows correspond to the periods during which data collection was conducted, while the 'Death' windows represent the time frames over which mortality outcomes were recorded. For instance, in the HRS dataset, we utilise information collected between 1998 and 2008 to predict deaths occurring from 2008 to 2019. This framework reframes our research question as follows: Given an individual's previously collected data, what is the probability of death within the next X years (where X varies by dataset)?

We define death prevalence as the ratio $\frac{\text{death counts within observation window}}{\text{total observations within the sampling window}}$. For all individuals included in the analysis, either the exact death or a confirmed alive status is available. Figure 2 presents the age distributions by death status across the three datasets (Figure S3 provides age and gender decomposed death prevalence). Among them, ELSA exhibits the lowest death prevalence, which can be attributed to its relatively shorter prediction window.

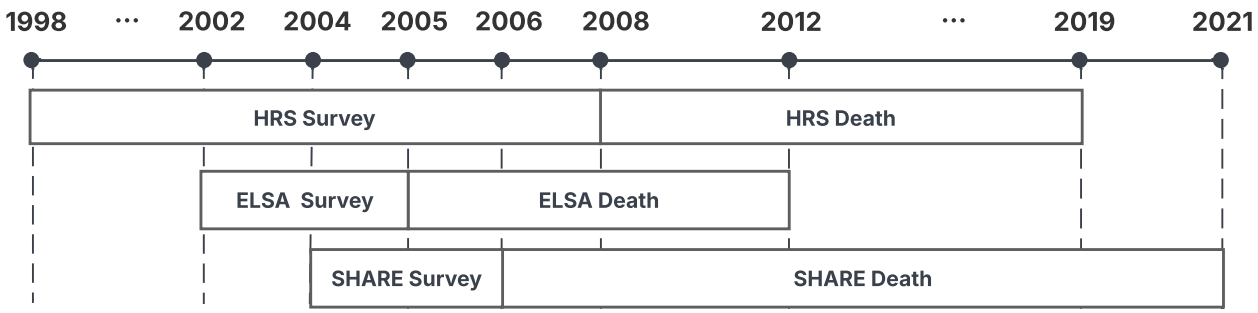

Figure 1: Sampling and Death Window of Three Datasets

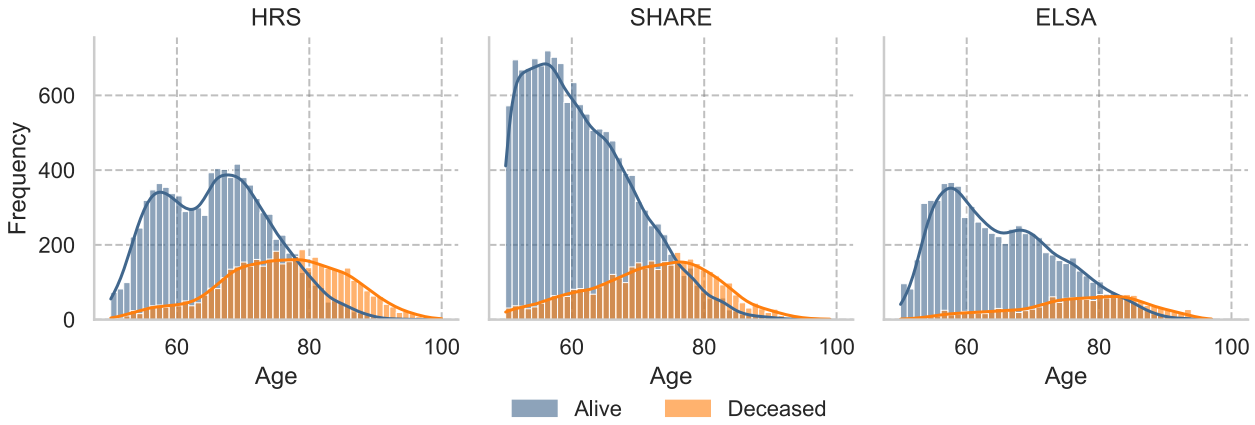

Figure 2: Age Distribution by Death Status of Three Datasets

Statistical Analysis

We estimate the predictability of elderly mortality using each of the three datasets individually as well as combined datasets, employing twelve classification machine learning algorithms. These algorithms include the Stochastic Gradient Descent classifier, K-nearest Neighbours classifier, Logistic regression, Decision Tree classifier, Support Vector Machines, Gaussian Naive Bayes, Adaptive Boosting Classifier, Bagging Classifier, Random Forest Classifier, Extra Trees Classifier, Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting Machine. We also adopt a state-of-the-art ensemble learning method, the Super Learner (SL), to estimate death predictability. Following the guidance of Bengio [19], all models are fitted multiple times with different seeds.

Model performances are assessed using several metrics: ROC-AUC, PR-AUC, pseudo-R^2, and the Inter-Model Vigorish (IMV) score. Higher values across all metrics indicate better performance. Among them, IMV offers a comparative assessment of binary classification performance across different models. ROC-AUC and PR-AUC measure the area under the receiver operating characteristic curve and the precision-recall curve, respectively. In the context of imbalanced datasets, PR-AUC is generally preferred over ROC-AUC, as the latter can be overly optimistic when the positive class is rare. Metric definitions can be found in SI: Evaluation Metrics. All datasets are split into conventional 7:3 train-test subsets, and training sets are used to fit models.

To ensure explainability, we select the algorithm with the best performance to investigate the source of predictability by calculating the Shapley Value for each risk factor. Shapley Values are designed to elucidate the importance of input variables within machine learning algorithms. The Shapley value (SHAP hereafter) of a risk factor i quantifies its marginal contribution to the predicted probability of death, relative to the model's baseline prediction (i.e., the mean predicted probability across all individuals). It captures how much i shifts the individual's predicted probability from this average (see SI: Methodology for technical details). Since SHAP values can be both positive and negative, we primarily present the mean absolute SHAP for interpretation. Given the sensitivity involved in predicting subjective mortality, individual-level SHAP values will not be presented.

In addition to analysing single risk factors, we devise a new leave-one-domain-out algorithm to estimate the domain-level importance in prediction accuracy. This algorithm iterates through all possible domain combinations (127 combinations with seven domains) and calculates the marginal contribution of each specific domain to the evaluation metric. The final domain importance is the mean of all marginal contributions across every possible combination. This approach provides a comprehensive comparison of risk factors at an aggregate level.

Results

Death Predictability

Table 2 presents the average model evaluation metrics across ten different seeds for the three individual datasets and the combined datasets, respectively. The standard errors are reported in parentheses. Higher values for all metrics indicate better out-of-sample predictive performance. BM stands for the benchmark model of Logistic Regression trained with age and gender only, following the practice in Vabalas et al. [14]. As most of our samples are uneven regarding death outcomes—with test set in-sample prevalence ranging from 0.169 in ELSA to 0.315 in HRS—we use the PR-AUC score as the primary machine learning evaluation metric.

Table 2: Average Model Performance for Death Prediction across Ten Random Seeds.

	HRS			SHARE			ELSA		
Metrics	SL	LightGBM	BM	SL	LightGBM	BM	SL	LightGBM	BM
IMV	0.196	0.202	0.175	0.094	0.098	0.093	0.061	0.069	0.067
	(0.003)	(0.003)	(0.046)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
ROC-AUC	0.816	0.820	0.790	0.812	0.815	0.806	0.824	0.832	0.818
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)
PR-AUC	0.695	0.698	0.675	0.575	0.586	0.566	0.508	0.533	0.529
	(0.002)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)	(0.009)	(0.007)	(0.007)
R^2	0.282	0.287	0.246	0.246	0.253	0.239	0.212	0.240	0.231
	(0.002)	(0.002)	(0.002)	(0.004)	(0.005)	(0.004)	(0.009)	(0.007)	(0.006)
Pseudo R^2	0.282	0.287	0.246	0.246	0.253	0.239	0.212	0.241	0.231
	(0.002)	(0.002)	(0.002)	(0.004)	(0.005)	(0.004)	(0.009)	(0.007)	(0.006)
IP	0.315	0.315	0.315	0.214	0.214	0.214	0.169	0.169	0.169

	HRS + SHARE			HRS + ELSA			SHARE + ELSA		
Metrics	SL	LightGBM	BM	SL	LightGBM	BM	SL	LightGBM	BM
IMV	0.136	0.138	0.124	0.139	0.142	0.122	0.082	0.086	0.076
	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
ROC-AUC	0.821	0.823	0.804	0.823	0.825	0.796	0.811	0.816	0.798
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
PR-AUC	0.642	0.647	0.621	0.646	0.650	0.618	0.548	0.561	0.523
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)
R^2	0.277	0.281	0.250	0.276	0.280	0.240	0.233	0.246	0.213
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Pseudo R^2	0.277	0.281	0.250	0.277	0.280	0.240	0.234	0.246	0.213
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
IP	0.258	0.258	0.258	0.259	0.259	0.259	0.200	0.200	0.200

	HRS + SHARE + ELSA		
Metrics	SL	LightGBM	BM
IMV	0.118	0.121	0.108
	(0.001)	(0.001)	(0.001)
ROC-AUC	0.817	0.821	0.801
	(0.001)	(0.001)	(0.001)
PR-AUC	0.616	0.624	0.592
	(0.001)	(0.002)	(0.002)
R^2	0.264	0.271	0.238
	(0.002)	(0.002)	(0.002)
Pseudo R^2	0.264	0.271	0.238
	(0.002)	(0.002)	(0.002)
IP	0.239	0.239	0.239

Note: Standard errors are shown in parentheses.

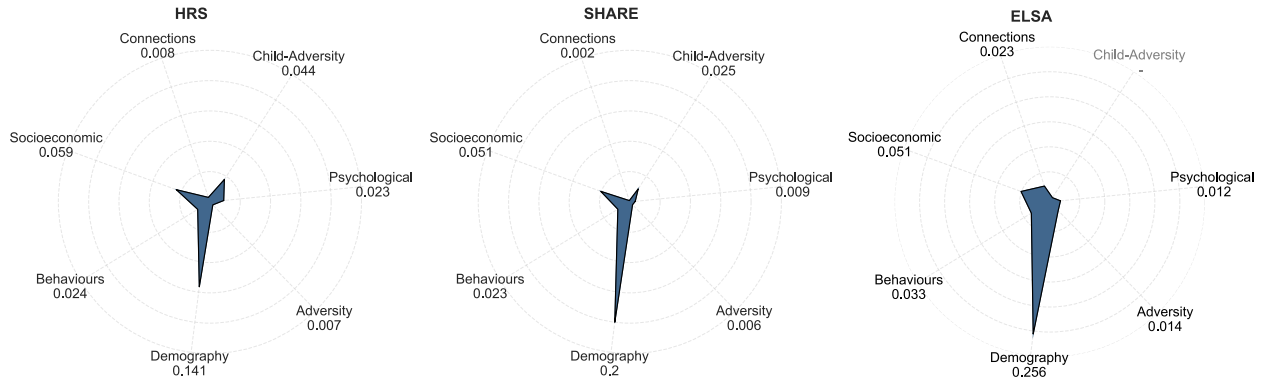

Figure 3: Domain-level Marginal Prediction Contribution of All Datasets. It presents the domain-level risk factors' marginal PR-AUC score contribution. LightGBM is used for all model fitting procedures. Consistent patterns exist in HRS, SHARE and ELSA.

We find that death is generally highly predictable across all models and datasets, as the PR-AUC scores are more than double the in-sample prevalence, which represents the probability of random guessing and serves as the interpretation baseline of PR-AUC. Predictions made using HRS and combined datasets demonstrate better performance, as both dataset sizes and in-sample prevalence are significantly larger compared to SHARE and ELSA. Surprisingly, the ensemble learner SL performs worse than LightGBM in all scenarios, signalling that LightGBM achieves overwhelmingly strong predictive performance and thus is not suitable for averaging with other algorithms in this case.

Cross Sectional Similarities: Domain Contribution

Figure 3 illustrates the domain-level contribution of risk factors in improving the PR-AUC score using the LightGBM model for each dataset separately. The HRS and SHARE datasets encompass predictors from all seven domains, while ELSA includes six domains. The number of predictors within each domain varies across datasets.

A consistent pattern emerges in all datasets, highlighting the demographic and socioeconomic domains as the most influential in enhancing the PR-AUC score, thereby improving the precision and accuracy of predictions. This finding aligns with the Fundamental Cause Theory proposed by Link and Phelan [3], which posits that socioeconomic status exerts a

persistent influence on health through multiple mechanisms, playing a fundamental role in health disparities. The demographic domain, encompassing key variables such as age and gender, consistently exhibits a substantial contribution across all datasets. This demographic-socioeconomic pattern also exists in the combined datasets, as shown in Figure S4.

It is important to acknowledge that the model's predictive performance is inherently shaped by the amount of relevant information presented in the data. For instance, in the HRS dataset, the psychological domain includes the largest number of risk factors, which partially explains its substantial contribution to domain-level predictive performance. However, the number of predictors alone does not fully determine domain importance. For example, in SHARE, the demographic and psychological domains contain three and six factors, respectively, yet the demographic domain's contribution is more than twenty times greater than that of the psychological domain.

In ELSA, the absence of the childhood-adversity domain may result in an incomplete assessment of risk factor importance, particularly in capturing the marginal impact associated with the psychological and social connection domain. As demonstrated by Kessler et al. [20], childhood adversities have a significant impact on adult mental disorders across 21 countries. Additionally, the distinct pattern of domain importance observed in ELSA may, in part, be attributed to the short prediction window, which restricts the dataset's capacity to capture variable importance beyond the demographic domain.

Cross Sectional Dissimilarities: Risk Factor Importance

Beyond assessing risk factor importance from a model performance perspective, we calculate SHAP values for a representative model from each dataset to decompose predicted death probabilities. Our objective is to provide insights into variable importance within robust predictive frameworks, aligning with the interpretative nature of SHAP values. Notably, SHAP values can accurately reflect factor importance only when the predictive algorithm is reliable. To mitigate the risk of over-interpreting unstable results, we present only those risk factors with stable SHAP values (i.e. their mean $|\text{SHAP}| \leq 0.1$) that rank among the top tier in risk

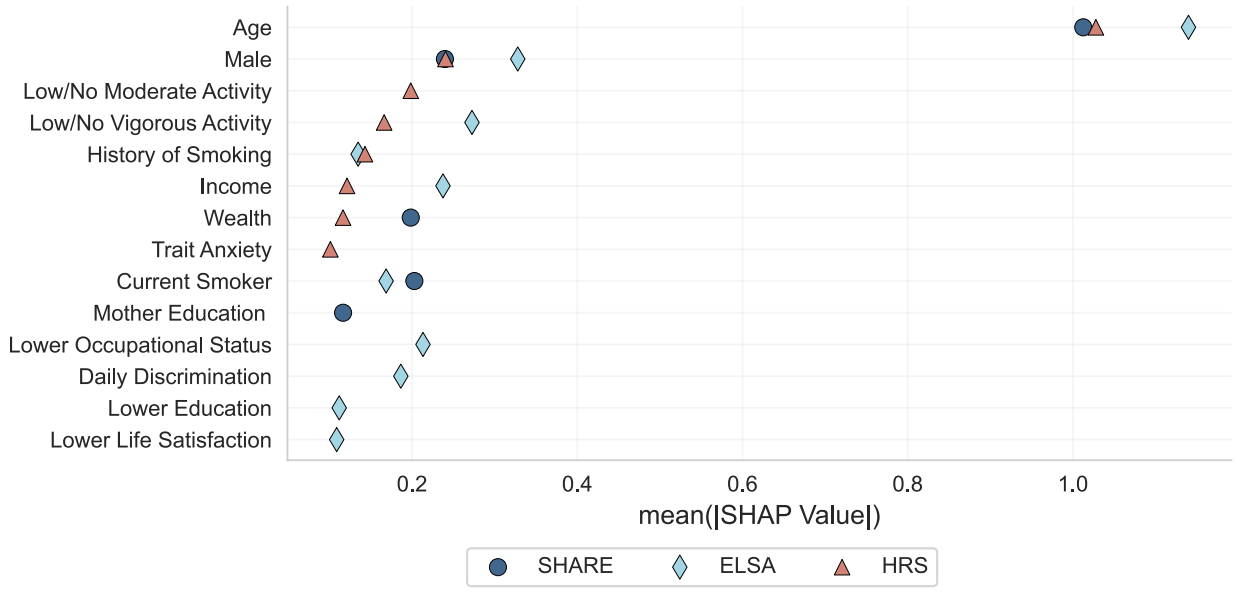

Figure 4: Risk Factor Importance of Three Datasets. Risk factors with mean |SHAP| > 0.1 from the predictions of HRS, SHARE, and ELSA are displayed in the figure. Mean |SHAP| values were calculated across the entire dataset (train + test) to comprehensively reflect the overall feature contribution, capturing both model fitting and generalisation behaviours. Risk factor ranks are calibrated based on their order in HRS. The ranks and factors vary across different predictions. Common risk factors across the three predictions are age and gender.

comparisons.

Figure 4 highlights the key differences among the datasets, which primarily stem from three aspects: the specific risk factors identified, their relative rankings, and their levels of importance. In HRS, the most influential risk factors (ranked from highest to lowest) are age, gender, low/no moderate activity, low/no vigorous activity, history of smoking, income, wealth, and trait anxiety. In SHARE, the most critical predictors include age, gender, current smoking status, wealth, and maternal education. In ELSA, the top-ranked risk factors encompass age, gender, low/no vigorous activity, income, lower occupational status, daily discrimination, current smoking, history of smoking, lower education and lower life satisfaction. Notably, ELSA exhibits the broadest range of significant risk factors, spanning nearly all domains except the psychological domain.

Among these top-ranked risk factors, only age and gender consistently emerge across all datasets. Age demonstrates a clear gradient effect in all cases. Gender, despite exhibiting different levels of importance across datasets, maintains the second-most important rank among

all datasets and a relatively stable trend across both age groups and datasets. Income and wealth collectively emerge as the most influential socioeconomic risk factors across all datasets. In ELSA, income ranks as the forth most significant predictor, whereas its impact—along with wealth—is less pronounced in SHARE and HRS.

Some risk factors appear to be context-dependent. Beyond age and gender, six additional risk factors are present across all datasets in Figure 4: history of smoking, wealth, low/no vigorous activity, current smoking, lower education, and lower occupational status. However, the importance of these factors varies by context. In the U.S. context, current smoking status, lower education, and lower occupational status do not rank among the most important factors (mean $|\text{SHAP}| > 0.1$). Conversely, in the European datasets (SHARE), history of smoking, low/no vigorous activity, and lower occupational status are not among the most influential predictors.

Furthermore, even among common risk factors, their relative importance varies. For example, income is a stronger predictor in ELSA than in HRS, which may be attributed to ELSA's exclusive reliance on income rather than both income and wealth. Similarly, low/no vigorous activity plays a more significant role in ELSA compared to HRS, despite their similar mean values of approximately 0.7. In terms of smoking status, SHARE stands out by identifying only current smoking as a significant risk factor, whereas history of smoking is more relevant in the other datasets. This distinction is likely due to the substantially higher prevalence of current smokers in SHARE (mean value: 0.48) compared to HRS (0.12) and ELSA (0.18).

Given our focus on predicting mortality outcomes in the elderly population, we further stratify the mean absolute SHAP values ($|\text{SHAP}|$) by age groups within each dataset, as shown in Figure 5. This stratification helps identify age-related gradients in feature importance and enables more nuanced cross-sectional comparisons across datasets.

Among the common predictors, age exhibits a consistent 'U-' or 'V-' shaped pattern in all three datasets. Notably, the lowest $|\text{SHAP}|$ values occur within the 60–75 age range, which corresponds approximately to the mean age at death in each dataset. This indicates that age contributes less to mortality prediction near the average death age, and becomes

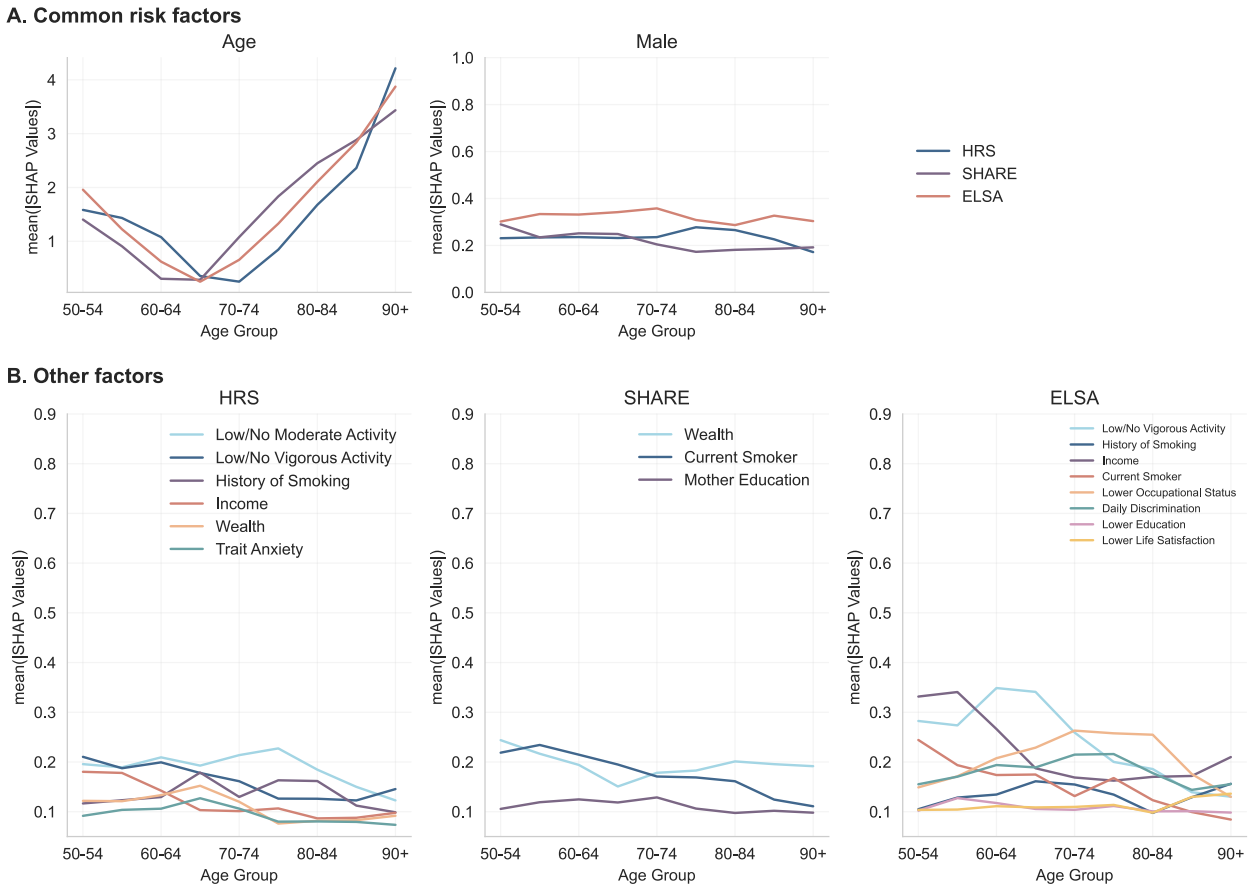

Figure 5: Age-specific Risk Factor Importance. Subfigure A shows the age-based importance of the three common variables, represented by the mean |SHAP| values. Subfigure B shows the importance trend of the top variables having overall mean |SHAP| values > 0.1 based on different datasets.

more influential for individuals at both younger and older extremes. Details about the SHAP interpretation of age can be found in SI: SHAP explanation of age.

Although age remains the most influential risk factor, its predictive importance in identifying mortality risk diminishes around this mean death age. Instead, age becomes more critical in predicting mortality among the oldest-old (those over 80 years), where the mean |SHAP| values exceed 1.5 in all datasets. Gender, by comparison, shows a relatively stable contribution to mortality prediction across age groups. The consistent prominence of age and gender—both demographic variables—highlights the dominant role of the demographic domain in predicting mortality risk in older populations across diverse settings.

Panel B further illustrates age-specific trends in variable importance across different risk factors and datasets. In the U.S. dataset, while most risk factors exhibit moderate fluctu-

ations across age groups, income, low/no vigorous activity, and trait anxiety demonstrate a mild downward trend with increasing age. In SHARE, current smoking status declines in importance as a predictor of mortality after age 60. The ELSA dataset displays the highest degree of variation, which may be partly attributed to the SHAP decomposition algorithm, where variable importance is computed based on predicted values. Income follows a 'V-shaped' pattern, with higher importance among both younger and the oldest age groups. In contrast, lower occupational status exhibits a peak around ages 70–74, following an inverted 'U-shape' across different age groups.

Despite their overall importance, health behaviour risk factors—such as smoking history, current smoking status, and low or no engagement in vigorous physical activity—exhibit a declining trend in predictive importance with advancing age across all datasets. This trend may be attributed to the general decline in individuals' ability to engage in physical activity as they age, which reduces the variability and predictive power of this factor in older age groups.

Discussion

In this study, we first address the question of how well mortality can be predicted using social determinants by selecting the best-performing model among multiple machine learning algorithms. Utilising information from multiple domains, we achieve high predictive performance across all datasets: in HRS, SHARE, and ELSA, the highest ROC-AUC scores are 0.820, 0.815, and 0.832, respectively, while PR-AUC scores reach 0.698, 0.586, and 0.533, respectively. Notably, the predictive performance of ELSA improves significantly when integrated with other datasets, particularly in combination with HRS (PR-AUC = 0.650).

At the domain level, we develop a leave-one-out algorithm to assess the marginal contribution of each domain, averaging its predictive performance impact across all possible combinations. Despite variations in the number of risk factors across domains, consistent patterns emerge. In all datasets, demographic and socioeconomic factors are the most influential predictors. In HRS and SHARE, the child-adversity domain ranks third in importance. Health

behaviours domain holds this position in ELSA. It also suggests that while key domains influencing mortality are relatively stable across countries, their relative importance may vary. These findings align with previous research indicating that while individual domains may have limited predictive power in isolation, their combined effects are critical for capturing mortality risk [13].

Dissimilarities are largely discovered in the decomposition of prediction using SHAP values. This decomposition enables us to identify the most influential individual risk factors for mortality prediction within each dataset, reflecting the unique characteristics and contextual influences of different populations. Age and gender consistently emerge as critical predictors across all datasets, underscoring the foundational role of demographic factors in mortality modelling among ageing populations. This finding aligns with existing literature. For example, Aida et al. [21] shows a large-scale change of survival days in both ELSA and the Japan Gerontological Evaluation Study (JAGES) with increasing age. Lantz et al. [22] demonstrates significant survival differences between females and males among U.S. adults.

However, beyond these commonalities, different datasets prioritise different risk factors. In HRS, the top-tier risk factors (mean $|\text{SHAP}| > 0.1$) highlight the significance of socioeconomic status, health behaviours, and mental health within the U.S. context. In SHARE, socioeconomic conditions, health behaviours, and childhood adversity emerge as key predictors, aligning with domain-level analyses. ELSA exhibits the broadest set of high-importance risk factors, encompassing multiple domains, which may be attributable to its inferior model performance and incomplete coverage of risk factors at the domain level.

Notable regional variations exist. For instance, current smoking status is highly influential in SHARE but not in HRS. Conversely, factors such as history of smoking and low/no vigorous activity play a more prominent role in HRS but not in the European datasets. Moreover, while ELSA shares many high-importance risk factors with both HRS and SHARE, it uniquely emphasises lower occupational status and lower education as key predictors. These differences highlight the necessity of context-specific approaches to mortality risk assessment, reinforcing the importance of integrating both generalised and localised perspectives in demographic and

health research.

Figure 5 also shows a general downward trend in health behaviour factors, suggesting a diminishing relative contribution of these factors in predicting mortality among the oldest-old, though they remain significant (mean $|\text{SHAP}| > 0.1$). While extensive research has established the link between health behaviour and mortality, particularly all-cause mortality (see Duncan et al. [23] for an example), few studies have systematically decomposed their importance across different age groups. This decline may be attributed to the overriding influence of age in late-life mortality risk or to physical limitations among the oldest-old, which may render certain health behaviours, such as vigorous physical activity, less relevant in distinguishing mortality risk within this group.

Despite the strengths of this study, several limitations must be acknowledged. First, discrepancies in data collection methodologies across datasets result in imperfect alignment of risk factors. This mismatch reduces the number of directly comparable variables when integrating datasets, thereby constraining cross-contextual comparisons (see Figure S2). For instance, only 25 out of 61 risk factors in HRS have direct counterparts in SHARE.

Second, our analysis is limited to mortality prediction within high-income countries—the U.S., Europe, and the U.K. While prioritising variable consistency enables more robust modelling, the exclusion of datasets from low- and middle-income countries restricts the generalisability of our findings. Incorporating data from comparable large-scale studies, such as the China Health and Retirement Longitudinal Study and the Longitudinal Aging Study in India, would enable a more comprehensive examination of global health disparities.

Finally, the exclusion of biological and genetic variables due to data limitations constrains the comprehensiveness of our models. Integrating biomarkers and genetic data could enhance predictive accuracy and provide deeper insights into the complex interplay between biological and social determinants of health. As such data become increasingly available, future studies should seek to incorporate these dimensions to refine mortality prediction models further.

References

[1] Joseph Carroll. "Death in Literature". In: *Evolutionary Perspectives on Death*. Ed. by Todd K. Shackelford and Virgil Zeigler-Hill. Cham: Springer International Publishing, 2019, pp. 137–159. ISBN: 978-3-030-25466-7. DOI: 10.1007/978-3-030-25466-7_7.

[2] Lisa F. Berkman and Ichiro Kawachi. "A Historical Framework for Social Epidemiology: Social Determinants of Population Health". In: *Social Epidemiology*. Ed. by Lisa F. Berkman, Ichiro Kawachi, and M. Maria Glymour. Oxford University Press, July 1, 2014, p. 0. ISBN: 978-0-19-537790-3. DOI: 10.1093/med/9780195377903.003.0001.

[3] Bruce G. Link and Jo Phelan. "Social Conditions As Fundamental Causes of Disease". In: *Journal of Health and Social Behavior* (1995). Publisher: [American Sociological Association, Sage Publications, Inc.], pp. 80–94. ISSN: 0022-1465. DOI: 10.2307/2626958.

[4] Dan Lewer et al. "Premature mortality attributable to socioeconomic inequality in England between 2003 and 2018: an observational study". In: *The Lancet Public Health* 5.1 (Jan. 1, 2020). Publisher: Elsevier, e33–e41. ISSN: 2468-2667. DOI: 10.1016/S2468-2667(19)30219-1.

[5] V. Lorant et al. "Socioeconomic inequalities in depression: a meta-analysis". eng. In: *American Journal of Epidemiology* 157.2 (Jan. 2003), pp. 98–112. ISSN: 0002-9262. DOI: 10.1093/aje/kwf182.

[6] Mai Stafford and M.G. Marmot. "Neighbourhood deprivation and health: does it affect us all equally?" In: *International Journal of Epidemiology* 32.3 (June 1, 2003), pp. 357–366. ISSN: 0300-5771. DOI: 10.1093/ije/dyg084.

[7] Julianne Holt-Lunstad, Timothy B. Smith, and J. Bradley Layton. "Social Relationships and Mortality Risk: A Meta-analytic Review". In: *PLOS Medicine* 7.7 (July 27, 2010). Publisher: Public Library of Science, e1000316. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1000316.

[8] Carla M. Perissinotto, Irena Stijacic Cenzer, and Kenneth E. Covinsky. "Loneliness in Older Persons: A Predictor of Functional Decline and Death". In: *Archives of Internal Medicine* 172.14 (July 23, 2012), pp. 1078–1084. ISSN: 0003-9926. DOI: 10.1001/archinternmed.2012.1993.

[9] Lucinda Rachel Grummitt et al. "Association of Childhood Adversity With Morbidity and Mortality in US Adults: A Systematic Review". In: *JAMA Pediatrics* 175.12 (Dec. 1, 2021), pp. 1269–1278. ISSN: 2168-6203. DOI: 10.1001/jamapediatrics.2021.2320.

[10] G. L. Engel. "The need for a new medical model: a challenge for biomedicine". In: *Science (New York, N.Y.)* 196.4286 (Apr. 8, 1977), pp. 129–136. ISSN: 0036-8075. DOI: 10.1126/science.847460.

[11] Kimberlé Crenshaw. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". In: *U. Chi. Legal F.* 1989 (Jan. 1, 1989), p. 139.

[12] Eli Puterman et al. "Predicting mortality from 57 economic, behavioral, social, and psychological factors". In: *Proceedings of the National Academy of Sciences* 117.28 (July 14, 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 16273–16282. DOI: 10.1073/pnas.1918455117.

[13] Casey F. Breen and Nathan Seltzer. *Demographic Perspectives on Predicting Individual-level Mortality*. Apr. 8, 2023. DOI: 10.31235/osf.io/znsqg.

[14] Andrius Vabalas et al. "Deep learning-based prediction of one-year mortality in Finland is an accurate but unfair aging marker". In: *Nature Aging* 4.7 (July 2024). Publisher: Nature Publishing Group, pp. 1014–1027. ISSN: 2662-8465. DOI: 10.1038/s43587-024-00657-5.

[15] Zoe Sanipreeya Rice and Pranee Liamputtong. "Cultural Determinants of Health, Cross-Cultural Research and Global Public Health". In: *Handbook of Social Sciences and Global Public Health*. Ed. by Pranee Liamputtong. Cham: Springer International Publishing, 2023, pp. 1–14. ISBN: 978-3-030-96778-9. DOI: 10.1007/978-3-030-96778-9_44-1.

[16] Viju Raghupathi and Wullianallur Raghupathi. "The influence of education on health: an empirical assessment of OECD countries for the period 1995–2015". In: *Archives of Public Health* 78.1 (Dec. 2020). Number: 1 Publisher: BioMed Central, pp. 1–18. ISSN: 2049-3258. DOI: 10.1186/s13690-020-00402-5.

[17] Jiani Yan and Charles Rahal. "On the unknowable limits to prediction". en. In: *Nature Computational Science* 5.3 (Mar. 2025). Publisher: Nature Publishing Group, pp. 188–190. ISSN: 2662-8457. DOI: 10.1038/s43588-025-00776-y.

[18] Scott Lundberg, Gabriel Erion, and SuIn Lee. "Consistent Individualized Feature Attribution for Tree Ensembles". In: *arXiv:1802.03888 [cs, stat]* (Mar. 6, 2019). arXiv: 1802.03888.

[19] Yoshua Bengio. *Practical recommendations for gradient-based training of deep architectures*. arXiv.org. June 24, 2012. URL: https://arxiv.org/abs/1206.5533v2 (visited on 08/30/2024).

[20] Ronald C. Kessler et al. "Childhood adversities and adult psychopathology in the WHO World Mental Health Surveys". In: *The British Journal of Psychiatry: The Journal of Mental Science* 197.5 (Nov. 2010), pp. 378–385. ISSN: 1472-1465. DOI: 10.1192/bjp.bp.110.080499.

[21] Jun Aida et al. "Social and Behavioural Determinants of the Difference in Survival among Older Adults in Japan and England". In: *Gerontology* 64.3 (Jan. 18, 2018), p. 266. DOI: 10.1159/000485797.

[22] Paula M. Lantz et al. "Socioeconomic and Behavioral Risk Factors for Mortality in a National 19-Year Prospective Study of U.S. Adults". In: *Social science & medicine (1982)* 70.10 (May 2010), pp. 1558–1566. ISSN: 0277-9536. DOI: 10.1016/j.socscimed.2010.02.003.

[23] Mitch J. Duncan et al. "The associations between physical activity, sedentary behaviour, and sleep with mortality and incident cardiovascular disease, cancer, diabetes and mental health in adults: a systematic review and meta-analysis of prospective cohort studies".

In: *Journal of Activity, Sedentary and Sleep Behaviors* 2.1 (Sept. 4, 2023), p. 19. ISSN: 2731-4391. DOI: 10.1186/s44167-023-00026-4.

Revisiting the social determinants of health with explainable AI: a cross-country perspective
Supplementary Information

Jiani Yan

Contents

1 Methodology

1.1 Super Learner

Developed by Laan, Polley, and Hubbard [1], a Super Learner (SL) is an ensemble algorithm that combines multiple different machine learning models together to produce a prediction that performs better than a single model. Ensemble models have long been discussed and explored; normally scientists choose to ensemble models either by averaging or applying some pre-defined weights, usually generated from the model performance. However, Laan, Polley, and Hubbard [1] has devised the SL, which aims to use a learning algorithm to cross-validate and find the optimal weights to combine a collection of predictive models.

To facilitate understanding, we divide the whole process into two parts: the cross-validation stage and the ensemble learning stage. We start illustrating by defining a set of m base predictive algorithms as $g_1, g_2, g_3, ..., g_m$, each representing a single candidate learner applied in the SL. In our case, as we aim to predict a binary outcome, we choose a set of classifiers including a stochastic gradient descent classifier, K-neighbors Classifier, Logistic Regression, Decision Tree classifier, Support Vector Machines, Gaussian Naive Bayesian, Adaptive Boosting Classifier, Bagging Classifier, Random Forest Classifier, Extra Trees Classifier, Light Gradient Boosting Machine Classifier, eXtreme Gradient Boosting Machine Classifier as the basic learners.

In the cross-validation stage, the ultimate goal is to construct and retrieve the performance of each learner in the training set. Specifically, the training set will be split into v mutual exclusive and nearly equal size sub-sets where candidate learners will be cross-validates on each of the v splits. In other words, further by splitting the original training set into v splits, selecting one of them v_i as a sub-validation set and the rest as sub-training sets ($v-1$ splits), every learner will be trained on the sub-training set and validated on the sub-validation set, generating a corresponding prediction of v_i. Repeating this procedure v times (validation set v_i, $i = 1, ..., v$) to generate a complete list of v predictions. Therefore, after v times of repeating validation on m single learners, we will generate a new space of possible outcomes Z, which has the same number of rows as the training set and m columns.

Then we move to the ensemble learning stage where the main goal is to train a user-defined learning algorithm $\tilde{\Psi}$ that minimises the squared error loss function. Specifically, $\tilde{\Psi}$ is designed to estimate the regression $E(Y|Z)$, where Y is the real label of the training set. The $\tilde{\Psi}$ can also be defined as the minimum cross-validated risk predictor as it minimises the cross-validated risk generated in the last step. Lastly, the fitted $\tilde{\Psi}$ will be applied to the single learners that are trained on the whole training set. Based on the data P_n and value X, the SL is given by

$$\hat{\Psi}\left(P_n\right)\left(X\right) \equiv \hat{\Psi}^*\left(P_n\right)\left(\hat{\Psi}_j\left(P_n\right)\left(X\right), j = 1, \ldots, m\right) \tag{1}$$

where $\hat{\Psi}$ is the mapping function ensembles the m learners from their risks obtained from the cross-validate step and $\hat{\Psi}^*\left(P_n\right)$ is the actual obtained predictor from the mapping function. In other words, in a SL, the Y for a value X is obtained by evaluating the predictor $\hat{\Psi}^*\left(P_n\right)$ at the J predicted values at X of the m candidate learners. In our case, we select LightGBM as the ensemble mapping learner.

1.2 LightGBM

As LightGBM is selected as a benchmark method, we briefly review its mechanism. Developed by Ke et al. [2], LightGBM is a cutting-edge variant of gradient boosting machine (GBM). In essence, it is a powerful ensemble learning method that aggregates multiple weak classifiers together to form a single learner in the manner of reducing errors along the gradient direction. It can be described with equations as follows: suppose we have T regressions trees $f_t(X)$ and the ultimate predicted value is generated through an integrated regression tree of the T regressions, i.e. $F_t(x_i) = \sum_{t=1}^{T} f_t(X)$. The integration is conducted through minimising a specific loss function $\sum_i l\left(\hat{y}_i, y_i\right)$ where the y_i is the true value of death occurrence and $\hat{y}_i$ is the predicted value. We choose the log loss here as our outcome variable is binary. During integration at the t^{th} step, the LightGBM method aims to minimise the value of following equation:

$$\min \Gamma^{(t)} = \min \sum_{i=1}^{n} l\left(y_i, F_{t-1}\left(x_i\right) + f_t\left(x_i\right)\right) \tag{2}$$

The minimisation task is undertaken by Newton's method as follows:

$$\Gamma_t \cong \sum_{i=1}^{n} \left(g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) \right) \tag{3}$$

where again, $g_i = \partial_{\hat{y}_i} L\left(\hat{y}_i, y_i\right)$ is the first-order gradient function of the loss function and $h_i = \partial^2_{\hat{y}_i} L\left(\hat{y}_i, y_i\right)$ is the second-order. Compared to other GBMs, Light GBM is novel in two unique techniques to tackle big data: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). With GOSS, it can reduce the sample size possessing small gradients as it is defined in information gains that only larger gradients contribute to higher information gain. EFB is designed to reduce the sparsity of the feature space through bundling features that are exclusive and rarely having non-zero values at the same time. Combining the two techniques together, the LightGBM has become one of the most efficient and scaleable boosting decision tree methods.

1.3 Shapley Values

Shapley Values are designed to unravel the input variable importance within machine learning algorithms. For a single feature i, the classical Shapley Value is obtained through the following formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right] \tag{4}$$

where S is a non-zero subset of the whole input features. M and N are the total number and the whole set of input features respectively (see Table 1 and Table S1 for the information for all datasets). $f_x(S)$ is the prediction function for feature in set S such as LightGBM. Under such a definition, the Shapley Value is a decomposition method that calculates the weighted difference between all possible subsets of the input features with and without the specific input feature i. Following the definition, Lundberg, Erion, and Lee [3] proposed the following equation to calculate Shapley Values:

$$f_x(S) = f(h_x(z')) = E[f(x)|x_s] \tag{5}$$

where $h_x(z')$ is a mapping function that maps between the original function and the left-out features (z', in a binary manner by setting $z'_i = 1$ or 0). $E[f(x)|x_s]$ is the expected function value conditioned on the subset S. The resulting Shapley Value will be called SHAP hereafter. In our research, the SHAP value for a risk factor i stands for the log odds being predicted as dead, which is the marginal contribution of i on the death prediction. A positive SHAP indicates a higher possibility of being predicted as dead and a negative value indicates a higher possibility of being predicted as being alive. As SHAP can be both positive and negative, to interpret SHAP, we adopt the mean absolute SHAP. Note that, unlike conventional regressions where a feature will only have one uniform 'β' presenting its contribution, SHAP is calculated at the individual level and therefore a distribution of i will be obtained. Appreciating the essence of predicting frailty in this research and to avoid interpretability at the single-factor level, the individual-level SHAP values will not be presented.

2 Evaluation Metrics

In this work, we will evaluate the out-of-sample predictive performance with two broad categories of evaluation metrics: conventional social science metrics and machine learning evaluation metrics. The Pseudo R^2 is calculated to provide conventional insights into the model goodness of fit, providing the following functions:

$$\text{Pseudo } \boldsymbol{R}^2 = 1 - \frac{\sum (y_i - \pi_i)^2}{\sum (y_i - \bar{y}_{test})^2} \tag{6}$$

where y_i stands for the real death situation of individual i (0 or 1) and π_i is the corresponding predicted death probability. $\bar{y_{test}}$ and $\bar{y_{train}}$ are the death prevalence of the test set and train set.

We will then use the classic PR-AUC score and the innovative Inter-Model Vigorish (IMV) score as machine learning evaluation metrics. The PR-AUC score is the Area Under the Curve of Precision and Recall, which emphasises the harmonic mean of precision and recall. It takes precision ($\frac{\text{TP}}{\text{TP+FP}}$) as the y-axis value, measuring the number of true positive observations against the total number of observations predicted as positive. In the x-axis, the recall rate ($\frac{\text{TP}}{\text{TP+FN}}$) measures the proportion of true positive in the real positive observations. One of the benefits of PR-AUC is that it focuses on the positive labels (death in our study) and therefore avoids a high score if the sample is unevenly distributed. As most of our samples have a very small portion of death (True) labels, PR-AUC is a better choice than the more prevalent ROC-AUC score, as ROC-AUC can easily gain a higher score if we randomly assign a False label to all samples. Using the PR-AUC grants us a more comprehensive understanding of our models. A higher PR-AUC is preferred as a better performance of the prediction, whereas in-sample prevalence (IP, True label portion of the training set) is viewed as a benchmark to interpret the PR-AUC score, any value higher than the IP is a sign of a 'good score'.

The IMV is a portable metric designed by Domingue et al. [4, 5] to understand the goodness of fit in the case of a binary outcome. We use the score as it's a comparable score between models and datasets, where our analysis involves evaluating the performance of four different datasets. To compute the IMV score, we start by establishing two binary predictive systems,

one of which is the baseline and the other one is the enhanced system. The enhanced system contains 'additional' information that is blind to the baseline system.

Recall that the entropy is defined as $-1(w \log w + (1-w)\log(1-w))$, to compute IMV, we aim to calculate the probability ω_0 and ω_1 that identifies the entropy of the baseline and enhanced system respectively that fulfils the following function:

$$\omega \log(\omega) + (1-\omega)\log(1-\omega) = \log(A) \tag{7}$$

where A is the geometric mean of the likelihood of the selected system, defined as:

$$A = \left(\prod_{i=1}^{n} L_i\right)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} p_i^{y_i}\,(1-p_i)^{1-y_i}\right)^{\frac{1}{n}} \tag{8}$$

here p_i is the predicted death probability of individual i and y_i is the true death label of the corresponding individual. The ω_0 and ω_1 are calculated by minimising the difference between the entropy definition and $\log(A)$:

$$\min |\omega \times \log(\omega)| + (1-\omega) \times \log(1-\omega) - \log(A) \tag{9}$$

and the IMV score is simply the relative ratio of the two probabilities:

$$\text{IMV} = \frac{w_1 - w_0}{w_0} \tag{10}$$

It reflects the extent the enhanced system (usually the fitted models) outperforms the benchmark system (usually a simple model). We select the in-sample prevalence as w_0, corresponding to the PR-AUC score described above.

3 SHAP Explanation of Age

To further interpret the 'V-' or 'U-' pattern of age, we note that age is the only feature whose mean SHAP values (not just absolute values) cross the zero line. As shown in Figure S5, the original SHAP values for age exhibit a steadily increasing trend: younger participants have negative SHAP values, while older participants have positive ones. Since SHAP values represent the marginal contribution of a feature to the model output relative to the mean prediction, a negative SHAP value for age implies a lower predicted probability of death compared to the dataset average, and vice versa. Therefore, the observed 'U-' or 'V-' shaped pattern in |SHAP| values can be understood as reflecting the non-linear influence of age. Age becomes more important for individuals who deviate substantially from the mean age at death—either because they are significantly younger (indicating low risk) or significantly older (indicating high risk).

4 Model Performance of Combined Datasets

From the comparison of data performance, we can observe that combining a single dataset with another dataset that has better predictive performance improve the overall predictive capability. For example, any combination of ELSA with HRS and SHARE significantly enhances

its performance. Specifically, the PR-AUC increases from 0.533 for ELSA alone to 0.650 for HRS+ELSA and 0.624 for HRS+SHARE+ELSA. For ELSA, the primary contributors to this performance improvement are the in-sample prevalence (increasing from 0.169 to 0.259 and 0.239) and the sample size (increasing from 8,389 to 21,599 and 39,417). These two effects even outweigh the reduction in the number of risk factors in SHARE+ELSA, where the model still shows improvement despite a reduction of 13 risk factors. Using the combined dataset, we estimate the importance of exploring how variations in the scale of sample size and the number of risk factors affect model performance. A similar conclusion could be drawn from SHARE where combining a dataset with better performance could achieve higher scores: both SHARE+HRS and SHARE+HRS+ELSA have higher performance compared to the model trained on SHARE alone, where HRS+SHARE is better than the HRS+SHARE+ELSA model.

Conversely, combining a dataset with worse predictive performance slightly deteriorates the combined scores. For instance, the PR-AUC score of SHARE+ELSA is lower than that of SHARE alone. In this case, both the number of risk factors (decreasing from 25 to 13) and the in-sample prevalence (declining from 0.214 to 0.200) decrease, while only the sample size increases (from 17,818 to 26,207). Another example is the HRS dataset, which exhibits a slight overall negative impact due to the reduction in risk factors and in-sample prevalence, despite an increase in sample size.

5 Tables

Table S1: Death Window and Prevalence of Combined Datasets

Dataset	HRS+ELSA	HRS+SHARE	SHARE + ELSA	All
Age range	50-100	50-100	50-99	50-100
Female portion	0.577	0.565	0.548	0.563
Risk factor number	25	25	13	13
Sample size	21599	31028	26207	39417
Death prevalence	0.259	0.258	0.200	0.239
prediction window	2005 - 2019	2006 - 2021	2005 - 2021	2005 - 2021

Note that there are 13 death cases of ELSA from 2004, which happened after data collection.

Table S2: Risk Factors by Dataset and Domain

Dataset	Domain	Count	Variables
HRS	Demography	5	Male, Black, Age, Foreign Born, Hispanic
HRS	Child-Adversity	7	Father Education , Father was Unemployed in Childhood, Relocated Homes in Childhood, Childhood Psychosocial Adversities, Family Received Financial Help in Childhood, Mother Education , Father Occupational Status
HRS	Socioeconomic	13	History of Food Insecurity, History of Unemployment, Lower Neighborhood Safety, Lower Neighborhood Cohesion, History of Food Stamps, History of Medicaid, History of Renting, Wealth, Lower Education, Lower Occupational Status, Income, Neighborhood Disorder, Recent Financial Difficulties
HRS	Behaviours	6	History of Smoking, Alcohol Abuse, Sleep Problems, Low/No Vigorous Activity, Low/No Moderate Activity, Current Smoker
HRS	Adversity	3	Major Discrimination, Daily Discrimination, Adulthood Psychosocial Adversity
HRS	Connections	7	History of Divorce, Lower Positive Interactions with Family, Negative Interactions with Family, Lower Positive Interactions with Children, Negative Interactions with Children, Negative Interactions with Friends, Never Married
HRS	Psychological	20	Lower Purpose in Life, Lower Sense of Mastery, Pessimism, Lower Conscientiousness, Lower Neuroticism, Negative Affectivity, Perceptions of Obstacles, Lower Extroversion, Lower Optimism, Lower Openness to Experiences, Loneliness, Hopelessness, Cynical Hostility, Anger In, Anger Out, Lower Religiosity, Lower Life Satisfaction, Lower Agreeableness , Trait Anxiety, Lower Positive Affectivity
SHARE	Demography	3	Male, Age, Foreign Born
SHARE	Child-Adversity	3	Father Education , Mother Education , Father Occupational Status
SHARE	Socioeconomic	5	History of Unemployment, History of Renting, Wealth, Lower Education, Lower Occupational Status
SHARE	Behaviours	5	History of Smoking, Sleep Problems, Low/No Vigorous Activity, Low/No Moderate Activity, Current Smoker
SHARE	Adversity	1	Adulthood Psychosocial Adversity
SHARE	Connections	2	History of Divorce, Never Married
SHARE	Psychological	6	Pessimism, Negative Affectivity, Perceptions of Obstacles, Lower Optimism, Hopelessness, Lower Positive Affectivity
ELSA	Demography	3	Male, Age, Foreign Born
ELSA	Socioeconomic	8	History of Unemployment, Lower Neighborhood Cohesion, History of Renting, Wealth, Lower Education, Lower Occupational Status, Income, Neighborhood Disorder
ELSA	Behaviours	4	History of Smoking, Alcohol Abuse, Low/No Vigorous Activity, Current Smoker
ELSA	Adversity	1	Daily Discrimination
ELSA	Connections	7	History of Divorce, Lower Positive Interactions with Family, Negative Interactions with Family, Lower Positive Interactions with Children, Negative Interactions with Children, Negative Interactions with Friends, Never Married
ELSA	Psychological	2	Loneliness, Lower Life Satisfaction

Note: Sleep Problem, Mother education and father education are removed from ELSA due to their disproportional cross-distributions with deaths, which largely biased the death prediction.

Table S3: Risk Factors of Combined Datasets

Dataset	Domain	Variable Count	Variables
HRS+SHARE+ELSA	Demography	3	Age, Male, Foreign Born
HRS+SHARE+ELSA	Child-Adversity	0	
HRS+SHARE+ELSA	Socioeconomic	5	Lower Education, Wealth, History of Renting, History of Unemployment, Lower Occupational Status
HRS+SHARE+ELSA	Behaviours	3	Current Smoker, History of Smoking, Low/No Vigorous Activity
HRS+SHARE+ELSA	Adversity	0	
HRS+SHARE+ELSA	Connections	2	History of Divorce, Never Married
HRS+SHARE+ELSA	Psychological	0	
HRS+SHARE	Demography	3	Age, Male, Foreign Born
HRS+SHARE	Child-Adversity	3	Father Education , Mother Education , Father Occupational Status
HRS+SHARE	Socioeconomic	5	Lower Education, Wealth, History of Renting, History of Unemployment, Lower Occupational Status
HRS+SHARE	Behaviours	5	Current Smoker, History of Smoking, Low/No Moderate Activity, Sleep Problems, Low/No Vigorous Activity
HRS+SHARE	Adversity	1	Adulthood Psychosocial Adversity
HRS+SHARE	Connections	2	History of Divorce, Never Married
HRS+SHARE	Psychological	6	Hopelessness, Negative Affectivity, Lower Optimism, Perceptions of Obstacles, Pessimism, Lower Positive Affectivity
HRS+ELSA	Demography	3	Age, Male, Foreign Born
HRS+ELSA	Child-Adversity	0	
HRS+ELSA	Socioeconomic	8	Lower Education, Income, Lower Neighborhood Cohesion, Neighborhood Disorder, Wealth, History of Renting, History of Unemployment, Lower Occupational Status
HRS+ELSA	Behaviours	4	Alcohol Abuse, Current Smoker, History of Smoking, Low/No Vigorous Activity
HRS+ELSA	Adversity	1	Daily Discrimination
HRS+ELSA	Connections	7	Negative Interactions with Children, Negative Interactions with Family, Negative Interactions with Friends, Lower Positive Interactions with Children, Lower Positive Interactions with Family, History of Divorce, Never Married
HRS+ELSA	Psychological	2	Lower Life Satisfaction, Loneliness
SHARE + ELSA	Demography	3	Age, Male, Foreign Born
SHARE + ELSA	Child-Adversity	0	
SHARE + ELSA	Socioeconomic	5	Lower Education, Wealth, History of Renting, History of Unemployment, Lower Occupational Status
SHARE + ELSA	Behaviours	3	Current Smoker, History of Smoking, Low/No Vigorous Activity
SHARE + ELSA	Adversity	0	
SHARE + ELSA	Connections	2	History of Divorce, Never Married
SHARE + ELSA	Psychological	0	

6 Figures

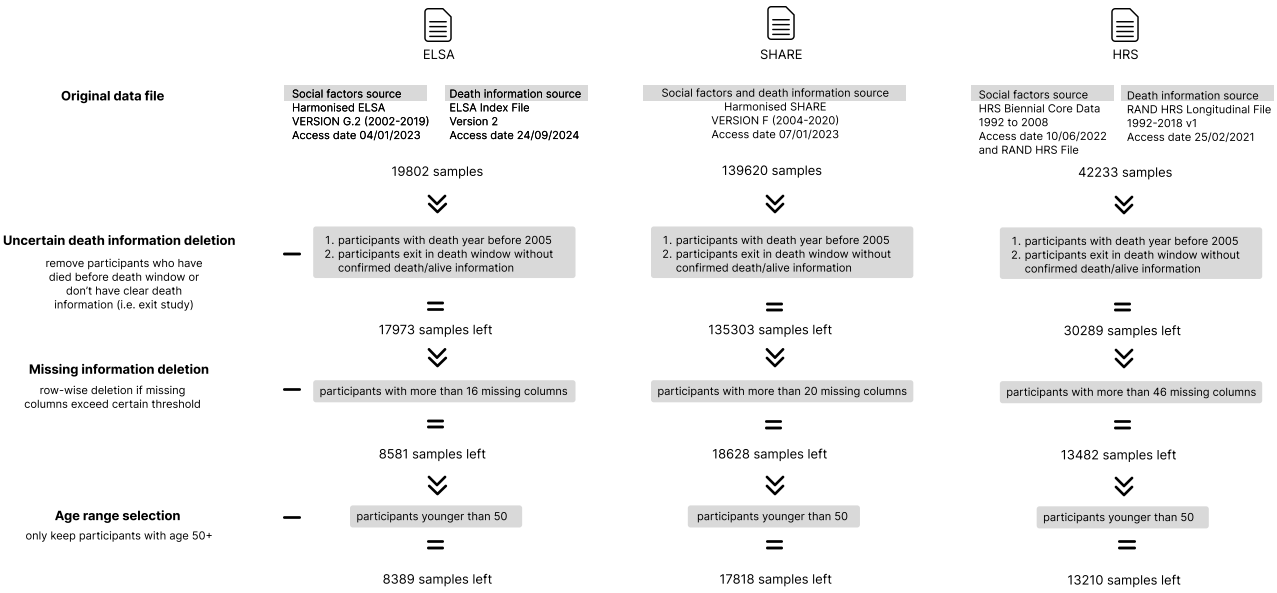

Figure S1: Data pruning process illustration for all datasets

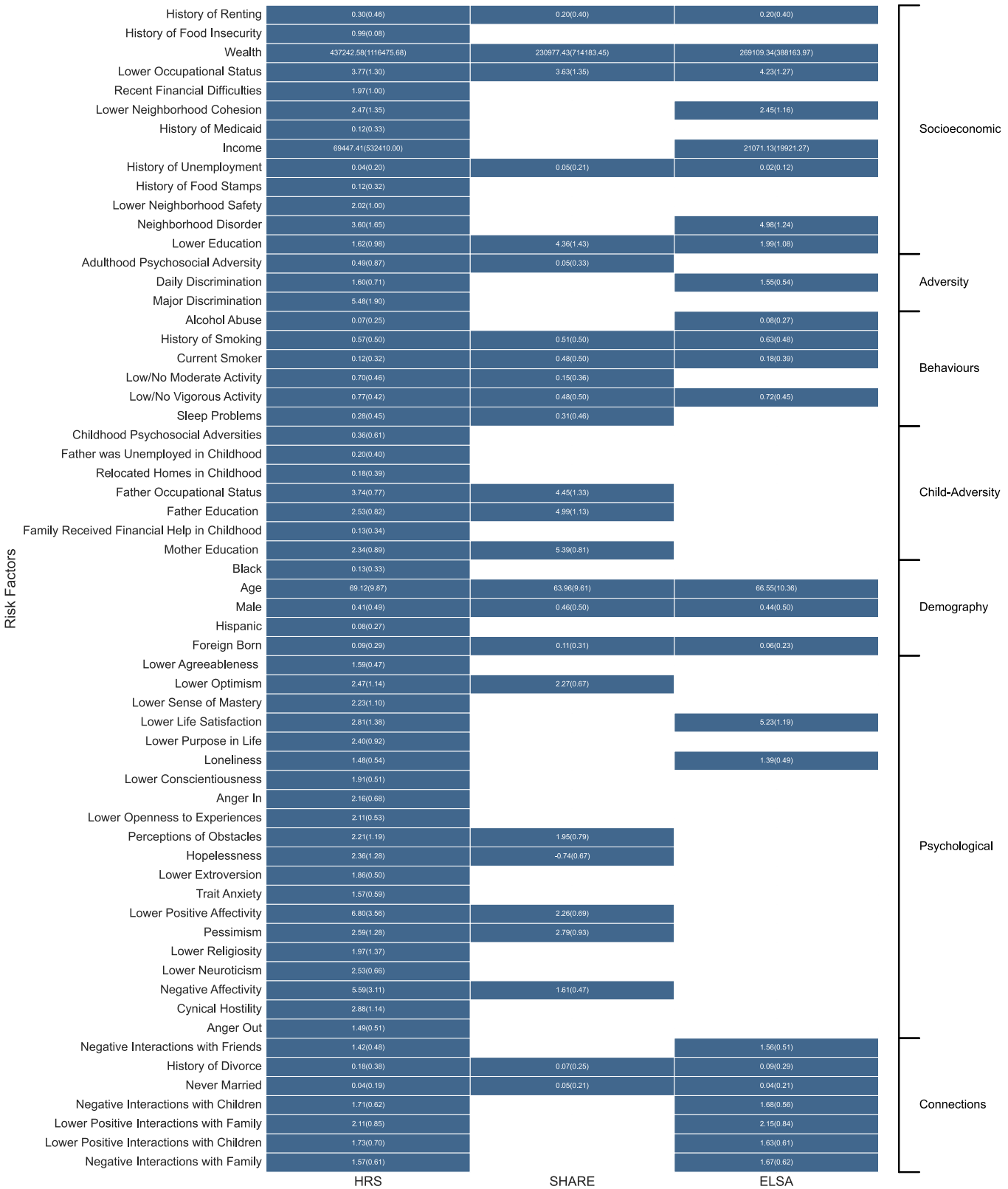

Figure S2: Risk factor inclusion and exclusion, mean value and standard deviation by dataset

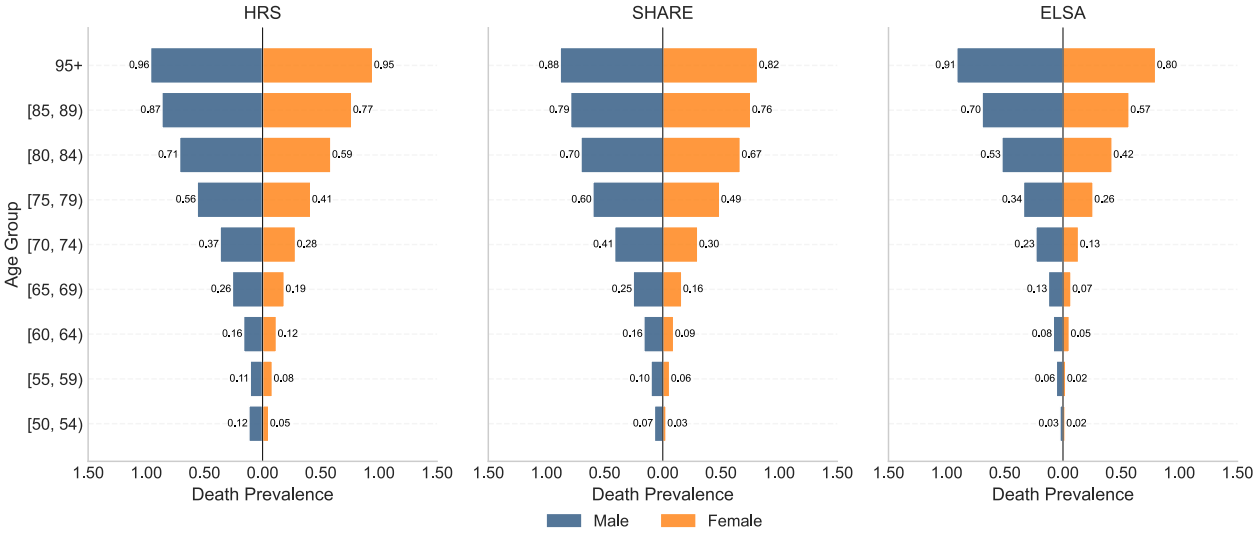

Figure S3: Death prevalence by age and gender for all datasets. For all datasets, we observe an increasing trend of death prevalence with increasing age groups, and for both genders. Compared to males, females of the same age have lower death prevalence in each dataset.

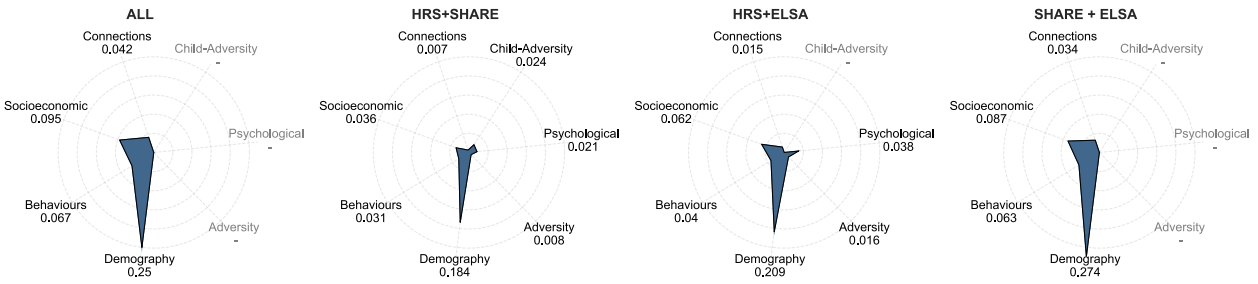

Figure S4: Domain contribution for all combined datasets. The observed pattern persists: demographic domain ranks the highest and then socioeconomic domain.

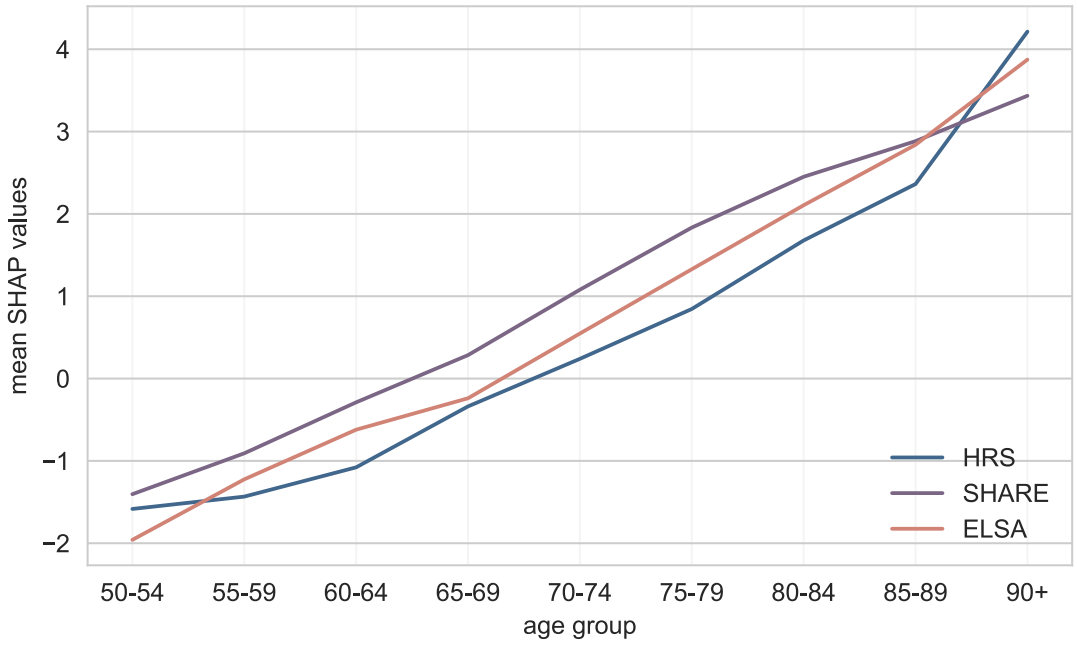

Figure S5: Mean SHAP Values of Age Across Different Age Groups. There is a clear and continuous upward trend of mean SHAP values for age across the three datasets, from younger to older age groups. There is a distinct pattern in mean SHAP values across the three datasets, transitioning from negative to positive as age increases. Negative mean SHAP values indicate that younger ages reduce the predicted risk of death relative to the average age of the sample, while positive values show that older ages increase predicted risk. The magnitude of SHAP values on either side of zero reflects how strongly age influences predictions, with larger absolute values indicating stronger relative influence, whether protective (negative SHAP) or risk-enhancing (positive SHAP).

References

[1] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (Sept. 16, 2007). Publisher: De Gruyter. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1309.

[2] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[3] Scott Lundberg, Gabriel Erion, and SuIn Lee. "Consistent Individualized Feature Attribution for Tree Ensembles". In: *arXiv:1802.03888 [cs, stat]* (Mar. 6, 2019). arXiv: 1802.03888.

[4] Benjamin W. Domingue et al. "The InterModel Vigorish (IMV) as a flexible and portable approach for quantifying predictive accuracy with binary outcomes". en. In: *PLOS ONE* 20.3 (Mar. 2025). Publisher: Public Library of Science, e0316491. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0316491.

[5] Benjamin W. Domingue et al. "The InterModel Vigorish as a Lens for Understanding (and Quantifying) the Value of Item Response Models for Dichotomously Coded Items". eng. In: *Psychometrika* 89.3 (Sept. 2024), pp. 1034–1054. ISSN: 1860-0980. DOI: 10.1007/s11336-024-09977-2.