Mortality and Machine Learning: A Glance at Death, Holistically and Precisely

Jiani Yan, University of Oxford

Introduction

Research on mortality typically explores factors from either a biological or non-biological perspective. Academics have identified an abundance of theoretical risk factors which 'contribute' to and are statistically associated with death. However, we recognise several areas for improvement in this body of work:

Predictive: There is little evidence concerning the predictability of death. The use of predictive frameworks is not a common research design in social sciences, presenting an area for further exploration.

Holistic: Existing studies generally draw evidence from one or two disciplines. A holistic view of risk factors, particularly from a non-biological perspective, is rarely explored.

Precise: As methodologies advance, we present several concerns regarding prediction precision in health-related outcomes.

Precise: Seed Variability

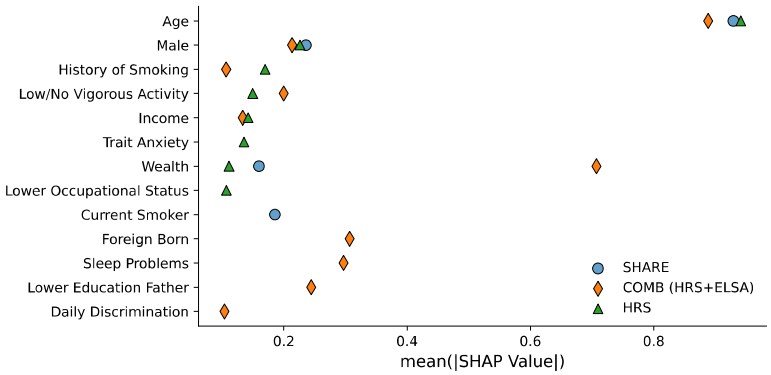
Evaluation of Model Performance in Death Prediction over 10 seeds

	HRS			SHARE			ELSA		
Metrics	SL	LightGBM	LR	SL	LightGBM	LR	SL	LightGBM	LR
IMV	0.200	0.204	0.211	0.069	0.072	-0.024	0.016	0.015	0.018
ROC AUC	0.822	0.825	0.831	0.795	0.800	0.448	0.908	0.909	0.894
PR AUC	0.689	0.687	0.707	0.497	0.507	0.161	0.262	0.246	0.259
EFRON R^2	0.288	0.290	0.307	0.196	0.204	-0.050	0.096	0.080	0.133
FFC R^2	0.562	0.564	0.574	0.446	0.452	0.277	0.224	0.210	0.255
IP	0.293	0.293	0.293	0.196	0.196	0.196	0.057	0.057	0.057

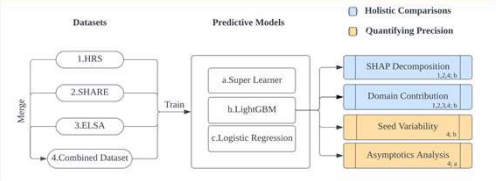
	HRS + SHARE			HRS + ELSA			SHARE + ELSA		
Metrics	SL	LightGBM	LR	SL	LightGBM	LR	SL	LightGBM	LR
IMV	0.106	0.108	0.108	0.120	0.122	0.116	0.065	0.067	0.062
ROC AUC	0.808	0.810	0.810	0.860	0.862	0.852	0.833	0.837	0.825
PR AUC	0.580	0.585	0.585	0.624	0.625	0.629	0.481	0.497	0.485
EFRON R^2	0.238	0.241	0.242	0.304	0.307	0.306	0.217	0.228	0.219
FFC R^2	0.511	0.513	0.513	0.528	0.530	0.530	0.410	0.419	0.412
IP	0.236	0.236	0.236	0.211	0.211	0.211	0.157	0.157	0.157

Holistic: SHAP Decomposition

Top Important Risk Factors of Three Datasets



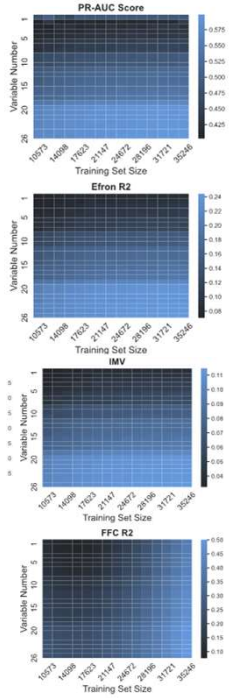
Methods and Data



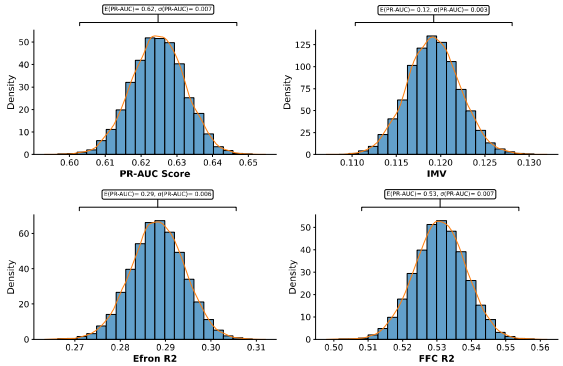
1. **U.S. Health and Retirement Study**
   Death Prevalence: 30.0%
2. **Survey of Health, Ageing and Retirement in Europe**
   Death Prevalence: 19.5%
3. **English Longitudinal Study of Ageing**
   Death Prevalence: 5.40%

Domain of Risk Factors:
Demography, Socioeconomics, Psychology, Adulthood Adversity, Childhood Adversity, Social Connections, Health Behaviours

Precise: Asymptotic



Precise: Seed Variability



Predictive Performance of 10000 Seeds in Train-Test splitting

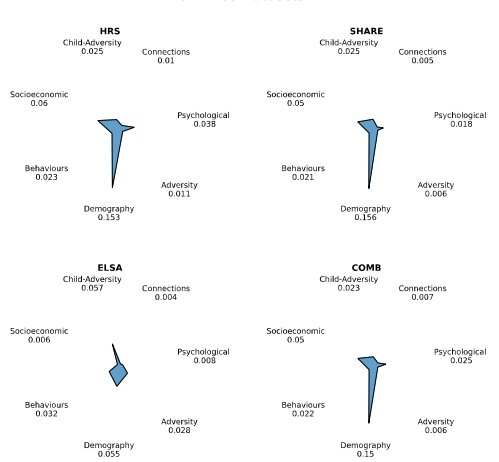
Outcome Distribution Characteristics of 10000 seeds

Type	Metric/Predictor	Mean	Min	Max	Standard Deviation
Evaluation Metrics	PR-AUC	0.625	0.597	0.651	0.007
	IMV	0.119	0.109	0.131	0.003
	Efron R^2	0.288	0.266	0.309	0.006
	FFC R^2	0.531	0.501	0.558	0.007
SHAP	Current Smoker	0.141	0.090	0.187	0.012
	Age	0.965	0.926	1.004	0.011
	Low/No Moderate Activity	0.115	0.071	0.15	0.010
	Male	0.229	0.196	0.272	0.010

Note: $FFC\ R^2 = 1 - \frac{\sum(y_i - \pi_i)^2}{\sum(y_i - \bar{y}_{train})^2}$;
$IMV = \frac{\omega_1 - \omega_0}{\omega_0}$, ω_0 is in-sample prevalence, ω_1 is the probability that defines the entropy of the model

Holistic: Domain Contribution

Domain-level Prediction Contribution of Three Datasets



OXFORD POPULATION HEALTH DSU — UNIVERSITY OF OXFORD — LEVERHULME CENTRE FOR DEMOGRAPHIC SCIENCE

Visit booth 312 and scan the code to find out more