Development of a digitally-obtainable 10-year all-cause mortality risk score based on data from 497,712 UK Biobank participants

Michele Colombo

Huma Therapeutics Ltd

Nikola Dolezalova

Huma Therapeutics Ltd

Aleksa Despotovic

Huma Therapeutics Ltd

Angus B. Reed

Huma Therapeutics Ltd

Davide Morelli

Huma Therapeutics Ltd

Mert Aral

Huma Therapeutics Ltd

David Plans (✉ david.plans@huma.com)

Huma Therapeutics Ltd

Research Article

Keywords: Mortality, prediction model, machine learning, survival analysis, digital health

Posted Date: February 21st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1368928/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

All-cause mortality (ACM) scores are a useful tool for identifying individuals with decreased life expectancy. An interpretable score consisting of smartphone-obtainable variables could allow for long-term management of individual health and support the next generation of healthcare monitoring and preventative practices. The aim of this study was to develop a 10-year ACM risk score using the UK Biobank dataset, using only digitally-obtainable variables.

Methods

The models were developed using the full UK Biobank cohort comprising nearly 500,000 individuals. We extracted 399 features from the dataset and, through a data-driven feature selection process with subsequent clinical review, identified 34 features for the final model. As part of the study, we compared two survival analysis approaches: Cox proportional hazards model and DeepSurv, a deep learning-based survival analysis algorithm.

Results

Before feature selection, Cox performed similarly to DeepSurv, achieving a c-index of 0.771 (95% CI 0.770–0.772) and 0.774 (95% CI 0.772–0.775) on the test dataset, respectively. Using the selected 34 features, the c-index of Cox decreased slightly to 0.770 (95% CI 0.769–0.770) and DeepSurv to 0.758 (95% CI 0.755–0.762). The models show excellent calibration at 10 years.

Discussion

This study describes a smartphone-compatible ACM score, with excellent predictive ability. The score described here improves on a previous smartphone-compatible score, C-Score, by incorporating non-modifiable factors in addition to variables which can be actively modified to reduce risk.

Conclusion

We have developed a smartphone-compatible ACM score that is comprehensive, easily interpretable and actionable, and as such, could provide a powerful tool for preventative healthcare.

Background

The rapid increase in life expectancy and decrease in birth rates in many countries around the world in recent decades has brought about a change in demographic landscape^{1,2}. Populations are ageing, conferring increased healthcare expenditure due to the higher number of morbidities in the elderly and the average higher cost per morbidity in this demographic³. Being able to identify individuals who have decreased life expectancy has important implications for policy and clinical practice, as well as for the individuals themselves, particularly if they are supported in identifying any pathways to reduce this risk, such as by changing certain lifestyle factors. Prognostic

models of death from any cause ('all-cause mortality', ACM) over a specified time period have been a helpful tool for evaluation of overall health status.

The National Institute for Health Care and Excellence (NICE) reviewed 41 existing tools for mortality predictions in 2016. It recommended that, owing to ubiquitous, shared limitations, further research should be undertaken to develop reliable tools for use in clinical practice ⁴. Many of these predictive models were developed using cohorts of older individuals (> 65 years) with a prediction horizon between one and five years ⁵⁻⁸. The UK Biobank (UKB) ⁹, a cohort study of ~ 500,000 UK participants aged 38–73, provides a unique opportunity to study risk factors for a broader age range over a longer time period.

Implementing an ACM score in a smartphone application would maximise access to tools that could support individuals' long-term health management. Such a score should be easily interpretable, actionable, and visibly dynamic to incentivise sustained lifestyle changes. Indeed, modifiable risk factors such as tobacco use, activity, and diet have been shown to be strongly associated with mortality ¹⁰⁻¹² and subsequently used in other risk models ¹³. Our previous effort to build a risk score within a smartphone application, named C-Score ¹⁴, incorporated heart rate, sleep duration, waist-to-height ratio, number of cigarettes per day, alcohol intake, reaction time, and self-rated health for predictions of 10-year ACM. This score deliberately included only modifiable predictors, resulting in a concordance index (c-index) of 0.66.

Here, we aim to build from this proof of concept and expand potential predictors to medical history, family history, sociodemographic and environmental factors, physical activity, mental health, and diet; many of which are known predictors of mortality ^{7,15,16}. All variables available for most UKB participants will be used in the initial set, following the exclusion of those that are not easily acquired by smartphone (via user input or passive recording) or are country-specific. Contrary to previous studies, we aim to use an entirely data-driven approach to select the most significant predictors from this initial set of variables, with a clinical review of the final predictor selection. Our modelling approach comprises traditional Cox proportional hazards modelling alongside a machine learning approach to survival analysis, the Cox proportional hazards deep neural network (DeepSurv) ¹⁷.

Objective

This study aims to develop a data-driven prognostic model for 10-year ACM using the UK Biobank dataset that can be implemented in a smartphone setting to support user engagement with their health.

Methods

Study Population

Data comes from the UKB ⁹, approved under UKB application number 55668. UKB participants were recruited for a prospective cohort study from the general population between 2006 and 2010. Data up to the 30th September 2020 update were used, which we further consider as the end of the follow-up period.

Input Features

We selected 77 fields based on literature review and clinical plausibility, ensuring that the information could be collected on a smartphone and applied to different geographies. This initial set included basic demographics (age,

sex, education level), anthropometrics (body measurements, weight, BMI), biometrics (heart rate), alcohol and smoking habits, sleep habits, self-rated health, medical and family history, physical activity habits, dietary habits, UV exposure and protection, and environmental variables (air pollution, proximity to roads).

Preprocessing

ACM outcome was defined as death from any cause during the follow-up period as per UKB field 40000. Additional insights were obtained analysing the underlying causes of death, field 40001. The length of follow-up was defined as the period between assessment date and either date of death or the end date of the study.

Main data transformations were: mean-imputation of missing values; merging groups of highly specific fields into a summary field (e.g. average weekly alcohol consumption was derived from a sum of consumption of different drink types); merging sex-specific fields (e.g. male-only and female-only fields for various medications); or deriving ratios of original features (e.g. waist-to-height ratio). Lastly, all categorical information was one-hot encoded, followed by excluding categories occurring less than 0.1%. Processing steps are summarised in Supplementary Table 1.

Experimental Settings

The dataset was split into training (75%) and test (25%) sets; the latter was used only for the final model's validation.

Two survival analysis approaches were tested, the Cox Proportional Hazard (CPH) model¹⁸ and its deep learning variant, DeepSurv¹⁷, which exploits artificial neural networks to model the relationship between prognostic factors and survival time. In the first instance, we used CPH to minimise the number of features without significant performance degradation. Both CPH and DeepSurv were then trained and evaluated using the resulting set of features.

CPH Model and Feature Selection

As CPH models are semi-parametric, the model's selection phase practically reduces to feature selection only.

Using the *lifelines* package¹⁹ an initial model was obtained by adjusting for age only. A baseline model with all the features was then trained and a stepwise variable selection process employed to remove features which do not have significant impact on performance. A set of six features (nine following one-hot encoding of self-rated health) was manually fixed within the model to extend the previously developed C-Score¹⁴.

We trained a univariate model for each feature during forward selection, keeping only those with p -value < 0.10 . A model was trained with all the remaining candidate variables during backward selection and its performance assessed using 5-fold cross-validation. Models excluding features in decreasing p -value order were then tested and if performance did not significantly degrade, the feature was eliminated. The process was continued until all variables were tested for removal. Features were initially tested in chunks of decreasing size in order to accelerate the process.

The final step of feature selection involved manual review in which features were eliminated where they were deemed clinically insignificant and where there was minimal performance contribution among the initially fixed features.

DeepSurv

DeepSurv models¹⁷, in contrast to the CPH model, require extensive hyperparameter optimisation. The focus, at first, was finding the best hyper-parameterisation for the replete baseline model to assess whether the problem involved non-linear components that the CPH model would not capture. A separate set of optimal hyperparameters was defined for the final reduced model using the same procedure. Since results suggested no significant improvement could be achieved by using DeepSurv on the baseline input space, no further experiments for features selection using DeepSurv were performed.

Models were trained employing an extension of the deep learning library *PyTorch*^{20,21}. Hyperparameter space was explored through a Tree-Structured Parzen Estimator (TPE)²², as provided by the *Optuna* library²³. Each model was tested employing three-fold cross-validation. Feed-forward neural networks with up to three hidden layers were tested, details of methods and search space are provided in Supplementary Table 2.

Statistical Analysis

Statistical analysis of baseline characteristics and train and test datasets were performed using Python *tableone* library²⁴. The discrimination metric for all models was the concordance index (c-index), while the Integrated Calibration Index (ICI), implemented in the lifelines library¹⁹, was used to evaluate calibration at the 10-year timepoint. Confidence intervals (CIs) were obtained using percentile bootstrap resampling with 50 resampling rounds.

Results

Population characteristics

The entire UKB cohort was used in this study. After excluding participants with missing data, the dataset contained 497,712 participants. There were 29,615 (5.96%) participants who died during follow-up (Fig. 1a). There were no statistical differences between train and test datasets among the features included in the final model (Supplementary Table 3).

The analysis of mortality causes in the studied cohort is summarised in Supplementary Table 4 and revealed that 53.3% of the deaths resulted from cancers (most commonly lung, breast, and pancreas cancers) and 20.3% from diseases of the cardiovascular system (particularly chronic ischaemic heart disease, myocardial infarction, and stroke). The remainder of the top-5 are diseases of the respiratory (7.3%), nervous (4.9%), and digestive system (3.8%). All other causes each contributed < 3% of the total deaths.

The demographic analysis of the cohort is presented in Supplementary Table 5, both in the overall sample and separated by outcome. Among the participants, 54.4% were women, with a median age of 58 at recruitment, and predominantly white (> 94%). The median follow-up time was 11.6 years (IQR 10.87–12.33).

Feature selection and CPH model

Model performance is reported in Table 1. The CPH model comprising only age obtained 0.690 c-index on the training dataset and 0.694 on the test dataset. The model trained with all 399 input features led to a c-index of 0.779 (95% CI 0.778–0.779) on the training dataset and 0.771 (95% CI 0.770–0.772) in the test dataset.

Table 1

CPH models results reported for different sets of input features. Shown are concordance indices obtained during training and internal validation on the test dataset, along with 95% bootstrap confidence intervals.

Features Set	Number of input features	Train C-Index	Test C-Index
Age	1	0.6900 [0.6900–0.6900]	0.6940 [0.6940–0.6940]
All Features	399	0.7786 [0.7783–0.7790]	0.7712 [0.7702–0.7719]
Step-Wise Selected	37	0.7724 [0.7721–0.7727]	0.7697 [0.7694–0.7700]
Manually Reviewed	34	0.7722 [0.7719–0.7725]	0.7695 [0.7691–0.7697]

Supplementary Table 1 outlines the features selected according to the stepwise variable selection procedure. Numbers of input features in the individual steps of the feature selection process are also summarised in Fig. 1b. 80 features were removed from the candidate set without any measurable degradation of performance following forward selection. Following backward elimination, 37 features were selected. These features were further subjected to manual review, excluding initially ‘fixed’ features with negligible impact (sleep duration and cigarettes-per-day) or those with problematic clinical explanation (experienced headaches in the past month being a protective feature), resulting in 34 features. The performance after manual review remained equivalent: 0.772 on the training dataset and 0.770 on the test dataset. The contribution of individual features to the overall performance is shown in Supplementary Fig. 1, while the plot of coefficients for individual features is presented in Fig. 2 (detailed results in Supplementary Table 6).

While the baseline model slightly overestimated the predicted risk (ICI 0.10%), the final model showed excellent calibration (ICI 0.03%, Supplementary Fig. 2). The mean observed 10-year risk in the cohort was 4.79% (95% CI 4.75–4.82), while the 10-year risk predicted by the final model was 4.82% (95% CI 4.78–4.85).

DeepSurv

Optimal hyperparameters for baseline (399 features) and final (34 features) model were selected using three-fold cross-validation. Performance comparable to the CPH model was obtained in fewer than 50 iterations of the TPE algorithm (Supplementary Fig. 3). Subsequently, only negligible performance improvement was achieved. Hence, we limited the number of trials to 200 to avoid potential overfitting.

The resulting hyperparameters for the baseline model led to a c-index of 0.774 (95% CI 0.772–0.775; trial 85) in the test dataset. For the final model with 34 features, the best performance was 0.758 (95% CI 0.755–0.762; trial 181). There was minimal difference between performance on the training and test datasets for both models, indicating no overfitting (Table 2).

Table 2

Best baseline and reduced models selected within the first 200 trials. Median concordance indices with 95% bootstrap confidence intervals shown.

	Trial #	Hyper-Parameters	Train C-Index	Test C-Index
Baseline model (399 features)	85/200	Activation: SELU	0.7795	0.7736
		Batch Normalization: Yes	[0.7785–0.7806]	[0.7724–0.7746]
		Dropout: 0.1311		
		Weight Decay: 0.0402		
		Learning Rate: 0.010119		
		Optimizer: SGD		
		Momentum: 0.3650		
		Hidden Layers Shape: 128x128		
Final model (34 features)	181/200	Activation: ReLU	0.7620	0.7584
		Batch Normalization: Yes	[0.7571–0.7652]	[0.7546–0.7618]
		Dropout: 0.1585		
		Weight Decay: 12.8300		
		Learning Rate: 0.000550		
		Optimizer: Adam		
		Hidden Layers Shape: 256		

Discussion

By virtue of the UKB's comprehensive and diverse data, coupled with a long follow-up period, we were able to create a 10-year ACM CPH model with excellent predictive capability. Age and age-related conditions such as Parkinson's disease, which is known to contribute to ACM ²⁵, were predictably identified as both having high importance to the model alongside high hazard ratios (HR). Additionally, a number of pre-existing conditions, including cardiovascular (stroke and myocardial infarction), respiratory (COPD, emphysema, and bronchitis), diabetes, cancer, and psychiatric and neurological disorders, significantly contribute to ACM in our model. All retained pre-existing conditions are known to affect life expectancy ^{26–30}. The majority of these conditions are non-communicable diseases, which are largely preventable through appropriate modifications in lifestyle and behavioural aspects of health ³¹, as well as early medical intervention.

Besides pre-existing conditions, the features with the highest HR in our model – alcohol dependency, slow usual walking pace, active smoking, higher waist-to-height ratio, and increased resting heart rate – have all previously been shown to contribute to ACM ³². These features point to the fundamental aspects of one's health and their relationship with ACM, specifically physical activity, nutrition, alcohol intake, and smoking status ³². Interestingly, never or rarely using UV protection was another lifestyle factor that is significantly associated with increased risk for ACM in our

model. The relationship between UV exposure and development of skin cancers has been established in the literature³³, but the exact long-term effects of sunscreen protection are yet to be fully understood³⁴.

Contrastingly, the bulk of protective factors are common knowledge — brisk walking pace, positive self-reported health, and a never-smoker status or history of smoking cessation. Again, this points to the preventable aspect of disease occurrence, and emphasising again the well known benefit of smoking cessation even after years of smoking³⁵. Lastly, regular glucosamine use was identified as protective in our model. Often used for treatment of joint pain, glucosamine's beneficial effect on ACM has been established in literature by reducing one's risk of developing several age-related diseases³⁶.

In addition to the CPH model, we tested the deep learning approach to survival analysis, DeepSurv. This model achieved comparable performance for the baseline model with all 399 features but slightly underperformed CPH for the final model. The lack of significant improvement when implementing deep learning is not uncommon with ACM, as was shown in¹⁷, seemingly as minimal contribution of non-linear associations between factors; thus DeepSurv's ability to take advantage of non-linear relationships has not been exploited in this setting. Additionally, there is limited interpretability of the individual feature contributions in black-box models such as DeepSurv, making them less suitable for clinical translation.

Our model significantly improves on the previously published smartphone-compatible algorithm, C-Score, achieving a c-index of 0.77 vs. 0.66, respectively¹⁴. Among other studies using UKB, Ganna and Ingelsson (2015) built a CPH model for the prediction of 5-year ACM, achieving a c-index of 0.80 for men and 0.79 for women⁸. Separately, Weng *et al.* employed both a traditional statistical approach (c-index 0.75) and machine learning (0.78–0.79) to train models for prediction of 10-year premature ACM³⁷. Unlike these studies, we employed survival analysis in both traditional statistical and machine learning modeling which allowed us to account for length of survival rather than binary outcome at a single time point. Compared to our results, these models contain notable differences in the final features, likely due to different methodological approaches to feature selection. Our selection process allowed us to create a geographically-agnostic model (e.g. absence of UK-specific 'Townsend deprivation Index'), which requires at the minimum only an internet connection to complete, while still maintaining good predictive capability.

The value in such a model is two-fold: first, if used on an individual level, accessible ACM models can form the backbone of behaviour-change programmes by presenting the user with interactable, dynamic health forecasts based on their lifestyle choices; second, if used on a regional or population level, such models could be used to inform local funding initiatives targeted to the most prevalent risk factors within their sub-population.

The primary limitation of this study concerns the UKB dataset. First, the majority of the UKB population is of White ethnicity (94%), which can lead to poor replicability when implemented across other ethnic groups. Second, the cohort's age range is restricted to 37–73 years, which may impart a similar impact on generalisability. Third, the UKB population is considered to be healthier and wealthier than the general population³⁸. These limitations mean external validation is needed to solidify its applicability both in the UK and across other populations.

Conclusion

We have developed a 10-year ACM model with very good predictive capability that can be readily accessible through smartphones by the general population. A focus on factors that are modifiable either by an individual or at a population level further supports the needed shift towards preventative healthcare and promotes longevity. Future studies on more diverse samples should be carried out to enable its widespread use.

Declarations

Ethical approval: The project falls under the ethical approval given for the UK Biobank by North West - Haydock Research Ethics Committee (16/NW/0274).

Consent for publication: Not Applicable

Data availability statement: This research has been conducted using the UK Biobank Resource under application number 55668. The data can be accessed via a data access application directly to UKB.

Competing interest: MC, ND, AD, ABR, DM, MA, and DP are employees of Huma Therapeutics Ltd.

Funding: This research was funded by Huma Therapeutics Ltd.

Author contributions: MC, ND, ABR, DM, conceived and designed the study, carried out statistical analysis, interpreted the results and drafted and reviewed manuscripts. AD, MA, and DP, interpreted the results, reviewed manuscript for intellectual content, and aided in drafting final versions. All authors confirm they had access to the data, a role in writing the manuscript, and approved the final version for publication.

Acknowledgements

The authors would like to acknowledge Adam Cunningham for his contribution during the preparation of this manuscript.

References

1. Roser M, Ortiz-Ospina E, Ritchie H. Life Expectancy. *Our World in Data*. Published online May 23, 2013. Accessed February 16, 2021. <https://ourworldindata.org/life-expectancy>
2. Roser M. Fertility Rate. *Our World in Data*. Published online February 19, 2014. Accessed February 16, 2021. https://ourworldindata.org/fertility-rate?source=content_type%3Areact%7Cfirst_level_url%3Aarticle%7Csection%3Amain_content%7Cbutton%3Abody_link
3. Howdon D, Rice N. Health care expenditures, age, proximity to death and morbidity: Implications for an ageing population. *J Health Econ*. 2018;57:60-74.
4. National Institute for Clinical Excellence. Multimorbidity: clinical assessment and management, NICE guidelines NG56. *NICE, ed London*. 2016;443.
5. Hippisley-Cox J, Coupland C. Development and validation of QMortality risk prediction algorithm to estimate short term risk of death and assess frailty: cohort study. *BMJ*. 2017;358:j4208.
6. van Walraven C. The hospital-patient one-year mortality risk score accurately predicted long-term death risk in hospitalized patients. *J Clin Epidemiol*. 2014;67(9):1025-1034.
7. Austin PC, van Walraven C, Wodchis WP, Newman A, Anderson GM. Using the Johns Hopkins Aggregated Diagnosis Groups (ADGs) to predict mortality in a general adult population cohort in Ontario, Canada. *Med Care*. 2011;49(10):932-939.
8. Ganna A, Ingelsson E. 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. *Lancet*. 2015;386(9993):533-540.
9. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.

10. Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *The Lancet*. 2020;395(10226):795-808. doi:10.1016/s0140-6736(19)32008-2
11. Li K, Hüsing A, Kaaks R. Lifestyle risk factors and residual life expectancy at age 40: a German cohort study. *BMC Med*. 2014;12:59.
12. Wijndaele K, Sharp SJ, Wareham NJ, Brage S. Mortality risk reductions from substituting screen-time by discretionary activities. *Med Sci Sports Exerc*. 2017;49(6):1111.
13. Baer HJ, Glynn RJ, Hu FB, et al. Risk Factors for Mortality in the Nurses' Health Study: A Competing Risks Analysis. *American Journal of Epidemiology*. 2011;173(3):319-329. doi:10.1093/aje/kwq368
14. Clift AK, Le Lannou E, Tighe CP, et al. Development and validation of risk scores for all-cause mortality for the purposes of a smartphone-based "general health score" application: a prospective cohort study using the UK Biobank. doi:10.1101/2020.11.23.20229161
15. Walter S, Mackenbach J, Vokó Z, et al. Genetic, physiological, and lifestyle predictors of mortality in the general population. *Am J Public Health*. 2012;102(4):e3-e10.
16. Hakulinen C, Pulkki-Råback L, Virtanen M, Jokela M, Kivimäki M, Elovainio M. Social isolation and loneliness as risk factors for myocardial infarction, stroke and mortality: UK Biobank cohort study of 479 054 men and women. *Heart*. 2018;104(18):1536-1542.
17. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24.
18. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*. 1972;34(2):187-202.
19. Davidson-Pilon C, Kalderstam J, Jacobson N, et al. *CamDavidsonPilon/lifelines: v0.25.9*; 2021. doi:10.5281/zenodo.4505728
20. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res*. 2019;20(129):1-30.
21. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019:8024-8035.
22. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: *25th Annual Conference on Neural Information Processing Systems (NIPS 2011)*. Vol 24. Neural Information Processing Systems Foundation; 2011. <https://hal.inria.fr/hal-00642998/>
23. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Association for Computing Machinery; 2019:2623-2631.
24. Pollard TJ, Johnson AEW, Raffa JD, Mark RG. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open*. 2018;1(1):26-31.
25. Park J-H, Kim D-H, Park Y-G, et al. Association of Parkinson disease with risk of cardiovascular disease and all-cause mortality: A nationwide, population-based cohort study. *Circulation*. 2020;141(14):1205-1207.
26. Mehta NK, Abrams LR, Myrskylä M. US life expectancy stalls due to cardiovascular disease, not drug deaths. *Proceedings of the National Academy of Sciences*. 2020;117(13):6998-7000. doi:10.1073/pnas.1920391117
27. Shavelle RM, Paculdo DR, Kush SJ, Mannino DM, Strauss DJ. Life expectancy and years of life lost in chronic obstructive pulmonary disease: findings from the NHANES III Follow-up Study. *Int J Chron Obstruct Pulmon Dis*. 2009;4:137-148.

28. Raghavan S, Vassy JL, Ho Y-L, et al. Diabetes mellitus-related all-cause and cardiovascular mortality in a national cohort of adults. *J Am Heart Assoc.* 2019;8(4):e011295.
29. Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. *World Psychiatry.* 2014;13(2):153-160. doi:10.1002/wps.20128
30. Mbizvo GK, Bennett K, Simpson CR, Duncan SE, Chin RFM. Epilepsy-related and other causes of mortality in people with epilepsy: A systematic review of systematic reviews. *Epilepsy Research.* 2019;157:106192. doi:10.1016/j.eplepsyres.2019.106192
31. Marmot M, Bell R. Social determinants and non-communicable diseases: time for integrated action. *BMJ.* 2019;364:l251.
32. Li Y, Pan A, Wang DD, et al. Impact of healthy lifestyle factors on life expectancies in the US population. *Circulation.* 2018;138(4):345-355.
33. Moan J, Grigalavicius M, Baturaite Z, Dahlback A, Juzeniene A. The relationship between UV exposure and incidence of skin cancer. *Photodermatol Photoimmunol Photomed.* 2015;31(1):26-35.
34. Lindstrom AR, von Schuckmann LA, Hughes MCB, Williams GM, Green AC, van der Pols JC. Regular sunscreen use and risk of mortality: Long-term follow-up of a skin cancer prevention trial. *Am J Prev Med.* 2019;56(5):742-746.
35. Müezziner A, Mons U, Gellert C, et al. Smoking and all-cause mortality in older adults: Results from the CHANCES consortium. *Am J Prev Med.* 2015;49(5):e53-e63.
36. Li Z-H, Zhong W-F, Huang Q-M, Zhang X-R, Mao C. Response to: "Correspondence to 'Associations of regular glucosamine use with all-cause and cause-specific mortality: a large prospective cohort study' by Li et al" by Yueh et al. *Annals of the Rheumatic Diseases.* Published online 2020:annrheumdis - 2020. doi:10.1136/annrheumdis-2020-218659
37. Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLOS ONE.* 2019;14(3):e0214365. doi:10.1371/journal.pone.0214365
38. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol.* 2017;186(9):1026-1034.

Figures

Figure 1

Flow diagram of participants and input variables in the study. (a) Participant numbers used in the study, including breakdown of the recorded death outcomes in the train and test datasets. **(b)** Size of the input space before and after processing and after feature selection.

Figure 2

Plot of Cox Proportional Hazards model coefficients. Points show $\log(HR) \pm 95\%$ CI. HR = hazard ratio, CI = confidence interval.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryinformationd1.docx](#)