The Limits of Predicting Individual-Level Longevity

Luca Badolato[a,1], Ari Decter-Frain[b,1], Nicholas J. Irons[c,1], Maria Miranda[d,1], Erin Walk[e,1], Elnura Zhalieva[f,1], Monica Alexander[g], Ugofilippo Basellini[d,2], and Emilio Zagheni[d]

November 1, 2023

[a]Department of Sociology, The Ohio State University, Columbus, OH 43210; [b]Jeb E. Brooks School of Public Policy, Cornell University; [c]Department of Statistics, University of Washington, Seattle, WA 98195; [d]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, 18057 Rostock, Germany; [e]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, MA 02139; [f]Department of Machine Learning, Mohamed Bin Zayed University of Artificial Intelligence, 5902 Abu Dhabi, United Arab Emirates; [g]Departments of Statistical Sciences and Sociology, University of Toronto, Toronto, Ontario, Canada

[1]L.B., A.D., N.J.I., M.M., E.W., and E.Z. contributed equally to this work and should all be considered first authors.

[2]Corresponding author: basellini@demogr.mpg.de

Abstract

Individual-level mortality prediction is a fundamental challenge with implications for life planning, social policies and public spending. We model and predict individual-level lifespan using 12 traditional and state-of-the-art models and over 150 predictors derived from the U.S. Health and Retirement Study. Machine learning and statistical models report comparable accuracy and relatively high discriminative performance, but fail to account for most lifespan heterogeneity at the individual level. We observe consistent inequalities in mortality predictability and risk discrimination, with lower accuracy for men, non-Hispanic Blacks, and low-educated individuals. Additionally, people in these groups show lower accuracy in their subjective predictions of their own lifespan. Finally, top features across groups are similar, with variables related to habits, health history, and finances being relevant predictors. We conclude by highlighting the limits of predicting mortality from representative surveys and the inequalities across social groups, providing baselines and guidance for future research and public policies.

Introduction

How long will we live? Answering this question means addressing a fundamental issue of human nature. Mortality is the ultimate life outcome, making disparities in the length of life the most extreme type of inequality (van Raalte et al., 2018). Demographers and actuaries have tackled this question since the 17th century, providing convincing evidence of mortality "laws" and regularities at the population level (see, e.g., Gompertz, 1825; Oeppen and Vaupel, 2002; Riley, 2001; Thatcher et al., 1998). Conversely, predicting mortality at the individual level remains challenging. Death is highly unpredictable (Einav et al., 2018), even in clinical and controlled settings where health records are gathered and medical knowledge can guide predictions (Henderson and Keiding, 2005).

Linked to this question, a key debate in aging research that has arisen in recent years concerns whether there is an inherent limit to human longevity. On one hand, some scholars believe that such a limit does not exist. The deceleration or plateau of age-specific mortality rates at the oldest ages is often put forward as a motivation for the lack of such a limit, along with recently observed mortality improvements for the elderly (Alvarez et al., 2021; Barbi et al., 2018; Horiuchi and Wilmoth, 1998; Rau et al., 2008; Wilmoth et al., 2000). On the other hand, other scholars have questioned the existence of a mortality plateau, arguing that they are a result of inaccurate data or age exaggeration, and that consequently human longevity has a fixed limit and will not increase indefinitely (Gavrilov and Gavrilova, 2011; Newman, 2018; Olshansky et al., 1990). Furthermore, there remain significant inequalities in mortality and lifespan variation across individuals based on educational attainment, income levels, and race and ethnicity (Goldman et al., 2017; van Raalte et al., 2018).

Recent advances in statistical methods and artificial intelligence hold great potential for improving mortality predictions and contributing to the longevity limit debate. Accurate lifespan predictions would enable individuals to make more informed choices regarding financing retirement and provision or receipt of support (van Raalte et al., 2018); moreover, they would allow for targeting high-risk individuals and for better organizing and managing health care spending, pensions, and other social policies (Einav et al., 2018). Such predictions could also shed novel insights on the debate regarding human senescence. There is thus a pressing need for investigating whether recent developments in statistical methods can improve longevity predictions beyond well-established approaches.

In parallel with the above advances, a body of work attempting to characterize the limits of prediction in complex social and health systems, often employing machine learning, has formed (Hofman et al., 2017). Such work emphasizes both existing biases as well as the need for better mechanisms to evaluate predictions in a standardized manner. The performance of machine learning models relative to classical statistical methods has been varied, with machine learning obtaining considerable gains in prediction accuracy in certain applications (Dong et al., 2019; Francesco et al., 2023; Jean et al., 2016) and at best marginal gains in other settings (Dressel and Farid, 2018; Joel et al., 2020). In this study, we find that machine learning methods trained on vast amounts of data exhibit minimal performance improvements over statistical models estimated from a handful of demographic predictors known to be closely tied to the risk of mortality. Furthermore, for all models, we also observe substantial disparities in predictive performance across race and ethnicity, gender, and socioeconomic status.

A frequently proposed explanation for the existence of bias in prediction models is the underrepresentation of minority groups in training data (Chen et al., 2021; Koenecke et al., 2020; Larrazabal et al., 2020; Li et al., 2022; Martin et al., 2019). Nevertheless, despite the fact that we train models on the HRS, a nationally representative aging survey in the US, as well as on oversampled synthetic datasets derived from the HRS that balance demographic groups, we find persistent inequalities in the accuracy of predictive models for disadvantaged populations as reflected in previous work Obermeyer et al. (2019). By focusing on the task of predicting individual outcomes, beyond estimating aggregate trends across demographic strata, our findings reveal that these groups, which are known to be subject to lifespan inequality, also face greater unexplained variation in mortality due to factors unrelated to the numerous behavioral, demographic, health, and social indicators currently measured by the HRS. Such differences can perpetuate or even increase existing health disparities as these models are deployed to make real-world decisions, such as pricing health insurance policies, implementing targeted social programs, and guiding decisions in clinical settings (Char et al., 2018). We observe these disparities despite employing state-of-the-art machine learning models capable of capturing complex interactions between covariates, thereby ruling out heterogeneous covariate effects across groups as a main source of disparities in predictive accuracy.

Taken together, our findings affirm that mortality is a complex process driven by poorly-characterized factors that vary across social strata. Uncovering mechanisms to better understand mortality at the individual level will require thoughtful data collection and well-designed research studies no less than application of the most advanced statistical and computational tools available. In line with recent calls for greater focus on the links between structural racism and population health in social and health research (Hummer, 2023), our results motivate further investigation into the drivers of mortality and lifespan inequality in marginalized groups. In addition to policy implications, the findings of such an endeavor could be used to inform the design of future aging surveys to ask questions that better capture the health of diverse communities.

In the remainder of this paper we break the overarching challenge of predicting and understanding mortality risk at the micro level down into four inter-related questions. First, we aim to systematically assess the limits to micro-level predictability of longevity by comparing both statistical and machine learning methods applied on one of the most authoritative longitudinal surveys of aging in the United States, the Health and Retirement Study (HRS). In employing a variety of models we explore whether increases in available data and advances in computational techniques substantially improve such predictions. Second, we measure how prediction accuracy and uncertainty in life outcomes vary across socioeconomic groups, reflecting underlying lifespan inequality. Third, we identify which key variables accurately predict mortality and analyze whether or not they differ across socioeconomic groups. Finally, we compare individuals' self-reported survival predictions with the outputs of statistical and machine learning survival models. These four complementary analyses provide a rich baseline to guide future research on mortality predictions and, more broadly, to assess the limits and biases of using machine learning models for demographic analyses, as highlighted in the conclusions of the paper.

Results

Lifespan predictability

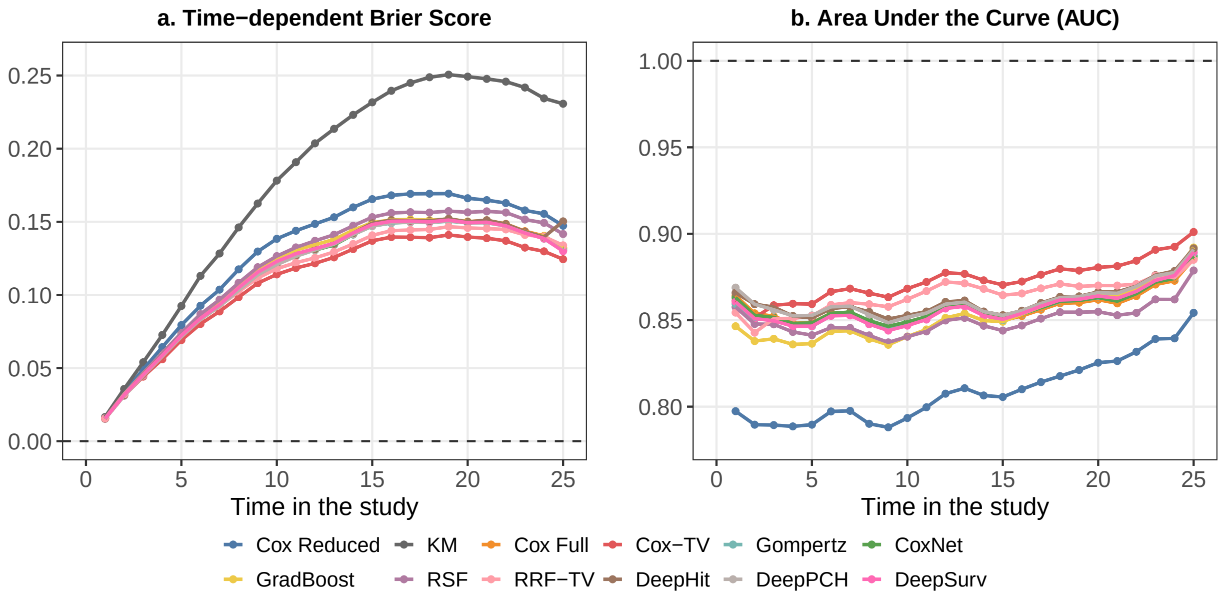

Figure 1: Evaluation metrics: Time-dependent Brier Score (panel a.) and Area Under the Curve (panel b.). The dashed horizontal lines indicate the optimal Brier Score (0.00) and the Area Under the Curve (1.00). The AUC for Kaplan-Meier, 0.5 by construction, has not been reported. Time in study is measured in years.

We assess the limits to individual-level predictability of lifespan by estimating 12 baseline, traditional, and machine learning models, as detailed in the Materials and methods section. We report the time-dependent Brier Score (BS) and the Area Under the Curve (AUC) for each model evaluated on the test set in Figure 1. Table S1 (Supporting Information) reports the integrated BS and mean AUC, summary measures of predictive performance over the entire time horizon of the study, for each model evaluated on the test set. The BS shows that more complex methods and predictors improve prediction accuracy compared to the baseline CoxReduced and Kaplan-Meier models, sometimes substantially. For example, the integrated BS difference between the CoxReduced and the time-varying Cox model (Cox-TV) is 0.021, a 16.0% improvement in prediction accuracy. Our results indicate that time-varying methods, which fully exploit the longitudinal structure of the HRS data and are generally overlooked in previous studies, consistently report better performance. Indeed, Cox-TV and the time-varying relative risk forest (RRF-TV) report, respectively, the lowest integrated BS among the traditional and machine learning models implemented. In particular, Cox-TV prediction accuracy is 6% higher than Cox Full, while RRF-TV prediction accuracy is 6.6% higher than the Random Survival Forest (RSF) in terms of integrated Brier score. Trends in discriminative accuracy across models, as measured by the Area Under the Curve (AUC), are similar to those of the Brier score. We note that all models except the Kaplan-Meier estimator report mean AUC above 0.8, considered excellent discriminative accuracy according to diagnostic standards (Mandrekar, 2010). More complex statistical models and predictors can significantly improve discrimination, which is crucial to target high-risk individuals. For example, Cox-TV reports a mean AUC of 0.874, 7.9% higher than CoxReduced. Again, the time-varying models, Cox-TV and RRF-TV, report the highest accuracy among traditional and machine learning models.

How predictable is lifespan? Previous results in clinical settings highlight the inherent challenges of predicting mortality (Henderson and Keiding, 2005). On the other hand, our AUC results suggest that statistical modeling can attain high discriminative accuracy on these data. However, we note that discrimination, which assesses a survival model's ability to rank subjects in terms of relative risk, is a lower benchmark than calibration, which assesses the accuracy of lifespan predictions. It can be difficult to interpret the Brier score in absolute terms, as opposed to the diagnostic thresholds established for the AUC (Mandrekar, 2010). Therefore, to further assess the performance of these models in lifespan prediction, we calculated predicted survival time in years as the area under the predicted survival curves by excluding censored individuals. We then compared our predictions with the observed survival times of subjects in the test set by calculating the Pearson correlation between these quantities for each model, which is shown in Figure S3 (Supoprting Information). Across models, the correlation coefficient varies between 0.4 and 0.5, implying that most of the heterogeneity in longevity at the individual level remains unexplained by the models and predictors used. We elaborate on these results in the Supporting Information and comment on their implications in the Discussion. Finally, for each model we calculated the mean absolute error (MAE) of predicted survival time on the test set, shown in Table S1 (Supporting Information). The average MAE among traditional and machine learning models is 2.39 years, which surpasses 21% of the average survival time of subjects in the test set whose deaths we observe, which is 11.3 years. Taken together, these results imply that the models' predictions of subject lifespans are not as accurate as their "excellent" mean AUC scores would suggest.

Notably, the time-varying models that make use of updated measurements in subsequent survey waves are the most performant, but also require and use additional information. Overall, however, we conclude from Figure 1 and Table S1 (Supporting Information) that traditional and machine learning models exhibit similar prediction accuracy on this dataset. Furthermore, we note that the improvement in predictive performance when going from the Kaplan-Meier estimator (which does not take into account covariates) to the reduced Cox model (based on 4 predictors) is much larger than the improvement when going from the reduced Cox model to the other models trained on the full set of over 150 predictors. These observations further highlight the difficulty of the prediction task at hand. While the time-invariant machine learning models report similar performance to the time-invariant traditional models in terms of integrated BS and mean AUC, we note that utilizing machine learning models comes with additional costs including greater difficulties in implementation, interpretability, and uncertainty quantification. Supposing these trends hold for other similar datasets, deciding which model is more suitable depends on the specific purpose of the study and data availability. If updated observations are unavailable, precluding the use of time-varying models, there is no clear superior choice. For public policy and social interventions aimed at targeting high-risk individuals, in which decisions are impacted by the statistical discriminative accuracy (or AUC) of a model, no method definitively outperforms the others. Similarly, for individual lifespan prediction, which is assessed by the Brier score, the differences between models may be irrelevant. On the other hand, for a life insurance company relying only on baseline covariates, switching to a model in which premia are assessed based on time-varying information may significantly improve revenue.

Inequalities in predictability among socioeconomic groups

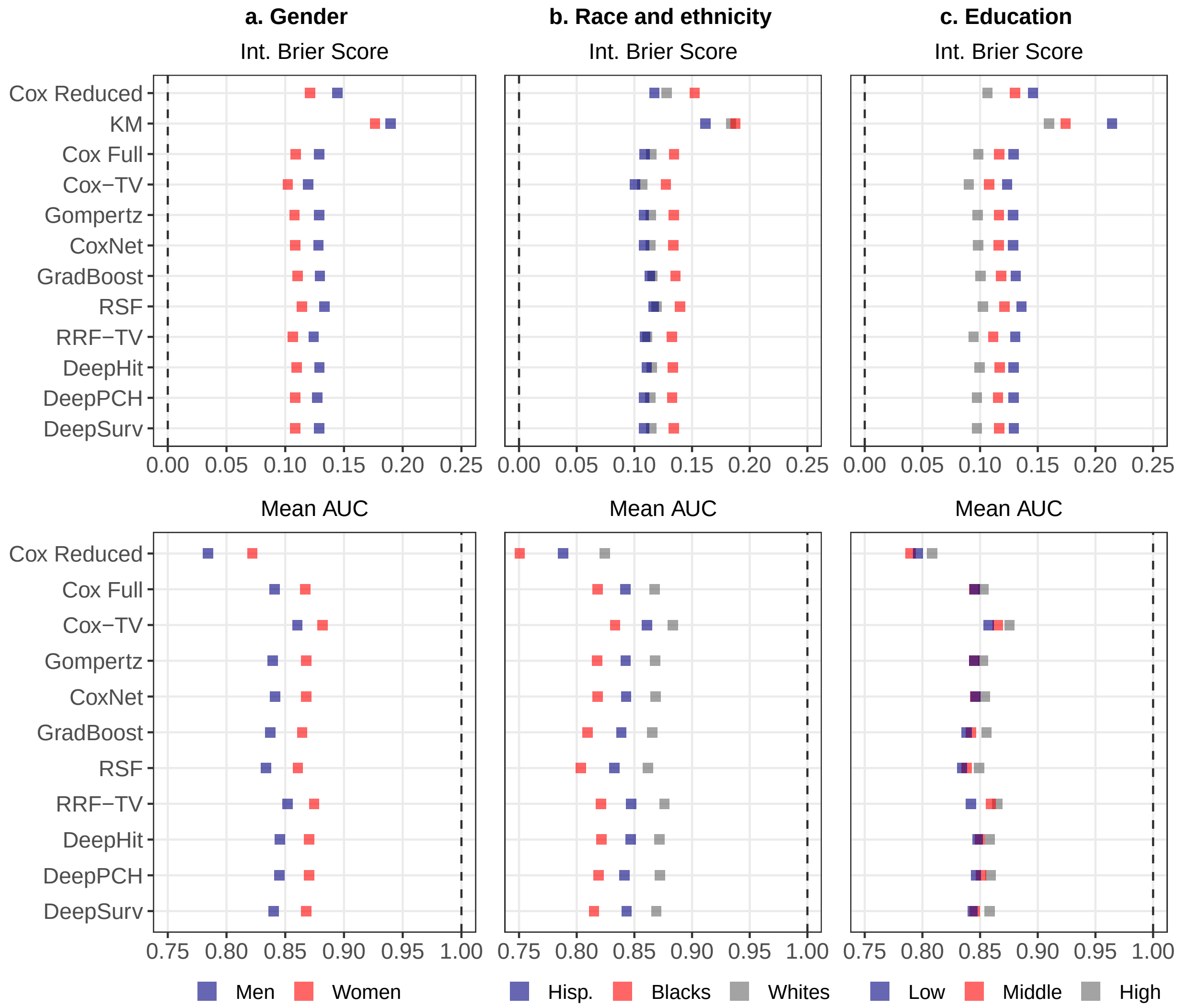

Figure 2: Integrated Brier Score and Mean Area Under the Curve by gender (panel a.), race and ethnicity (panel b.), and education (panel c.). The dashed vertical lines indicate the optimal Brier Score (0.00) and the Area Under the Curve (1.00). The Mean AUC for Kaplan-Meier, 0.5 by construction, has not been reported.

We investigated whether prediction accuracy varies between social groups, reflecting inequalities in mortality and morbidity underlined by previous research (Mackenbach et al., 2018; van Raalte et al., 2011, 2018). We estimated statistical models using a common training set and computed prediction accuracy in the test data among specific social groups to assess performance. We stratify our analyses by gender (men and women), race and ethnicity (non-Hispanic whites, non-Hispanic Blacks, and Hispanics), and education (low, middle, and high), as described in the data section. In Figure 2, we report integrated Brier score and mean AUC for each model evaluated on the test set stratified by each group.

Traditional and machine learning models show consistent inequalities in lifespan predictability across social groups. We find that the observed inequalities are robust to the choice of statistical model, and, in particular, that more novel machine learning models also exhibit inequalities in mortality prediction. Among men, Non-Hispanic Blacks, and low-educated respondents, all models report lower accuracy in predicting mortality, both in terms of calibration (integrated BS) and discrimination (mean AUC), as compared to their counterparts. For example, for whites we observe a mean AUC that is consistently 0.05 higher than for Blacks. In other words, statistical models better discriminate between high and low-risk white individuals, allowing better targeting for policy interventions, than between high and low-risk Black individuals.

As a robustness check, we ran the same analyses on augmented training sets oversampled to correct for imbalances in gender, race and ethnicity, and education for a subset of models. For instance, the observed inequalities in lifespan survival predictability could be driven by sample size differences among groups, as white respondents are, for example, over-represented in the training set compared to Black and Hispanic respondents. The results of the robustness check, which are qualitatively and quantitatively consistent with the those described here, are reported in the Supporting Information with more detailed information on the oversampling procedure.

Explaining why longevity is harder to predict for certain groups is key for future research on the root causes of disparities in mortality and lifespan predictability, with relevant consequences for life planning and social interventions. We identified the key variables to predict mortality across social groups as a preliminary approach to analyzing possible mechanisms driving the observed inequalities. As described in the data section and the Supporting Information, we computed variable importance using permutation importance with negative integrated brier score as the scoring metric.

We see minimal variation in the top features across groups, which are primarily related to habits, health history, and finances. Age, which is consistently the most predictive feature, improves the integrated brier score between 0.07 and 0.009 depending on the model used. Other top predictors, such as diabetes diagnosis and whether a participant has ever smoked, tend to have a significant but smaller impact, between 0.001 and 0.004. The order changes slightly across groups in different models. For example, items which may indicate healthcare access such as sum of medications and prescription for psychiatric medications are more predictive in White and higher educated populations. In the random forest model, diabetes diagnosis is more predictive of mortality for Blacks than Hispanics or Whites, similar to findings from Goldman et al. (2017). However, in general, the difference in feature importance across models is larger than across groups within a single model. Differences in top features also reflect some group-specific predictors, such as the age of the last menstrual period for women. Top variables for the parametric models are difficult to interpret, possibly due to collinearity issues since these models are not intended for such a large number of variables. We list the top 10 variables for each model in the Variable Importance section of the Supporting Information.

Inequalities in subjective predictability

Finally, we investigated how HRS participants' predictions about their own longevity compare to predictive models and to reality, and how these trends differ among social groups. HRS respondents are asked: "what do you think are the chances that you will live to be 75 or more?". This question is a measure of an individual's subjective survival probability at 75, which has been used, for example, to compute subjective cohort life tables (Perozek, 2008). Individuals form expectations holistically by considering their health background, environment, socioeconomic status, extended family experience, and genetics, among other information that may not be captured in surveys (Perozek, 2008), and such individual health assessments have been shown to be relatively predictive of mortality in some cases Goldman et al. (2016). This elicits the question: with access to such information, are individuals better able to predict their own survival? In the following analyses, we restrict the sample to respondents who were less than 75 at entry into the study.

In Figure 3, we report a calibration plot stratified by gender, race and ethnicity, and education, with the predicted survival probability on the horizontal axis and the observed proportion of respondents alive at 75 on the vertical axis. A perfect calibration is expected to follow the diagonal line - e.g., among respondents who estimate a 0.2 survival probability at 75, 20% are expected to be observed alive. We notice that subjective lines (highlighted in the graph) are generally flat. Individuals who report a low expected probability of being alive at 75 substantially underestimate survival, while those who report a high probability substantially overestimate survival. Conversely, model-based predictions better align with the diagonal line. We observe relevant differences among social groups.

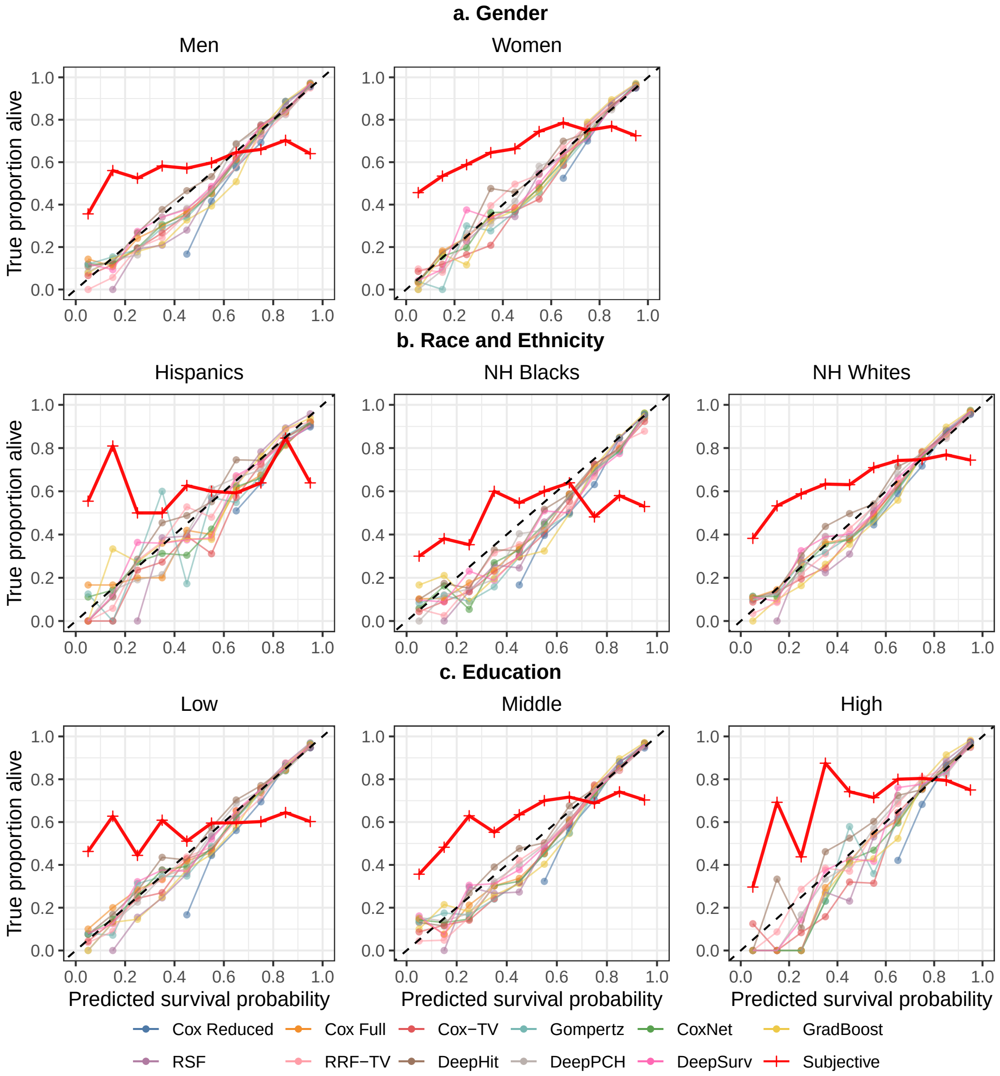

Figure 3: Calibration plot at age 75 of predicted survival probability and observed survival by gender (panel a.), race and ethnicity (panel b.), and education (panel c.). The dashed diagonal line represents a perfect calibration. Predicted survival probabilities have been binned in ten groups: 0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, 0.9-1.

For example, among respondents who reported a survival probability between 0.9 and 1, men, non-Hispanic Blacks, and low-educated individuals show lower observed survival. These groups face higher unexpected mortality relative to their counterparts.

In Figure S2 (Supporting Information), we also report the mean squared error (MSE) in predicted survival probability at age 75 for the estimated survival models and survey respondents ("Subjective") by social group. We underline two relevant results. First, individuals are significantly less accurate than statistical modeling - subjective survival predictions consistently score, on average, an additional 0.1 in terms of MSE. Second, the inequalities in lifespan predictability across social groups observed for the statistical models persist for the subjective survival probabilities reported by individuals in the survey. Remarkably, men, non-Hispanic Blacks, and low-educated respondents are less accurate in predicting their survival at 75 than women, non-Hispanic Whites, and highly educated respondents, respectively, consistent with the modeling results of the previous section.

Discussion

In this paper we sought to predict individual-level mortality from survey data, a demographic challenge with potential implications for individuals, policymakers, and researchers alike. We considered questions around prediction, model usage, and inequalities between groups and pushed the limits of predictability by including over 150 predictors in a broad range of classic statistical and machine-learning survival analysis models spanning Cox models, random forests, and deep neural networks. Our results generally suggest that models incorporating richer information demonstrate improved predictive performance and can attain relatively high discrimative accuracy, but achieving superior calibration of lifespan predictions remains a difficult task. In particular, we find that survival models accounting for time-varying covariates have the best performance. This is in line with previous research starting from the early work of Allison (1984). However, the additional prediction accuracy derived from time-varying covariates is not substantial, and we generally observe similar performance across traditional, penalized, and machine learning models. In addition, we find inequalities in prediction that cannot be explained by sampling procedures. Individual level prediction is consistently less accurate for certain groups, regardless of which model is used, motivating a need to delve more deeply into what may be causing such discrepancies. Finally, despite differences in the order of variable importance across models, most tended to draw predictive power from the same indicators like age, smoking status, and the presence of chronic diseases like diabetes, underscoring the success of past research in determining key mortality predictors.

Our results have several important implications. First, upon incorporating comprehensive life information, the discriminative performance of predictive models, which captures the ability to identify individuals at higher risk, can reach levels deemed 'excellent' by diagnostic standards (Mandrekar, 2010). Statistical modelling can thus be used to better inform individual-level life planning and organize more targeted social interventions and support programs. In particular, like other important social outcomes (e.g., as in Heller et al. (2022)), individual-level predictive survival models can help identify and target high-risk individuals. However, though time varying information is important, the emergence of a similar set of key factors across models provides a basis for prediction in lower information settings, and such factors may also be consistent across countries Goldman et al. (2016). Future work may consider at what point adding additional variables results in diminishing returns in terms of improved performance, which may aid countries with less comprehensive data collection efforts.

We have focused on one way to predict mortality by generating predicted survival curves based on fitted survival regression models. We briefly address two alternative approaches. One alternative approach would be to directly predict age at death, as is often desirable in healthcare settings (Henderson and Keiding, 2005). Breen and Seltzer attempt to predict lifespans in this way by focusing on a single birth cohort born in 1910 and identifying their sociodemographic characteristics and age at death from the US Census and Social Security records, respectively (Breen and Seltzer, 2022). We evaluated the accuracy of survival time predictions output by our models in a fashion analogous to theirs, which is available in the Supporting Information. However, our data are not well-suited to the evaluation of lifespan prediction due to the presence of right censoring. Respondents can drop out of the survey without having their deaths recorded, or they may outlive the length of the longitudinal period. The survival metrics we utilize above account for censoring and therefore provide a more valid assessment of the performance of the models considered in our context. Furthermore, we expect that results from our approach may be more relevant in practice, since the vast majority of data sources recording mortality contain censoring.

Another way to predict mortality would be to focus on forecasting the probability of death between waves of the survey. For instance, one could develop a discrete-time model using information from waves one through thirteen of the HRS to predict which respondents present at wave thirteen will die before wave fourteen. This 'discrete-time' approach has been used in previous work predicting social outcomes (Arpino et al., 2022; Heller et al., 2022; Salganik et al., 2020). This is an interesting approach that can be explored in future work.

Finally, we observed consistent inequalities in predictability: men, non-Hispanic Blacks, and low-educated respondents had less predictable mortality than their counterparts. These same inequalities reemerge in survey-respondents' own predictions about their survival, reflecting the persistent, potentially internalized nature of inequalities in lifespan (van Raalte et al., 2018). These inequalities present a non-trivial obstacle to the equitable use of predictive models, and must be accounted for when using models to target interventions. The persistence of differences in predictability even after including a rich set of predictors and complex modelling remains to be explained. One possibility may be that mortality is more driven by genetic, biological, or multi-generational factors for some groups compared to others (Christensen et al., 2006). Another is that certain groups may die more frequently from unpredictable causes (e.g., traffic accidents) than others. These possibilities remain to be explored in future work. In particular, a relevant dimension to consider is the role of genetics. We included a large set of behavioral, demographic, health, and social indicators, but genetic determinants of human longevity could further improve prediction accuracy and explain variation across individuals and groups. For instance, twin studies have consistently found that around 25% of human lifespan variation is driven by genetic differences, and that genetic influences on lifespan are minimal until 60 but increase after this age (Christensen et al., 2006). Understanding the differences which lead to such inequalities remains an important question in order to also assess equity in prediction of life expectancy.

Our findings emphasize the importance of interdisciplinary collaboration in addressing predictive disparities and establishing frameworks for evaluating models and predictions reliably, especially when applied to the social sciences. While the adoption of machine learning may offer marginal improvements in predictive accuracy, it often comes at the cost of transparency and interpretability and may not always supersede the benefits of traditional models. Moreover, it's crucial to assess models not only on a macro level but also with respect to specific subpopulations, considering factors such as race and gender as well as how they intersect. Diligent research in this field has the potential to benefit not only individuals in their long-term planning but also policymakers and other stakeholders dedicated to assisting marginalized communities and formulating end-of-life policies.

Materials and methods

Data

We use data from the US Health and Retirement Study (HRS, 2022), a representative sample of individuals over 50 years of age that has been run in two-year waves since 1992[1]. In particular, we use a longitudinally harmonized dataset ending in 2018 for a total of 14 waves[2] . Our data cleaning procedure is composed of four steps: variable selection based on raw categories (childhood, cognitive, demographic, habit, job, mental health, physical health, social, support, wealth, and welfare), variable selection based on repetition and missingness, special treatment for certain missing values, and removal of remaining variables with over 50% missingness. Our final dataset contains more than 150 predictors, spanning behavioral, biological, demographic, health, and social indicators for 39,248 respondents. An overview of the selected variables is available in the Supporting Information along with further discussion of our data cleaning procedure and the HRS and its limitations. Remaining missing values are imputed using random forest imputation (Stekhoven and Buhlmann, 2012), as done in previous work on mortality predictions using HRS data (Puterman et al., 2020), run separately on the train and test sets, which comprise a 60%-40% split of the full dataset. We stratify

[1] The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

[2] This analysis uses data or information from the Harmonized HRS dataset and Codebook, Version C as of Jan 2022 developed by the Gateway to Global Aging Data. The development of the Harmonized HRS was funded by the National Institute on Aging (R01 AG030153, RC2 AG036619, 1R03AG043052). For more information, please refer to www.g2aging.org.

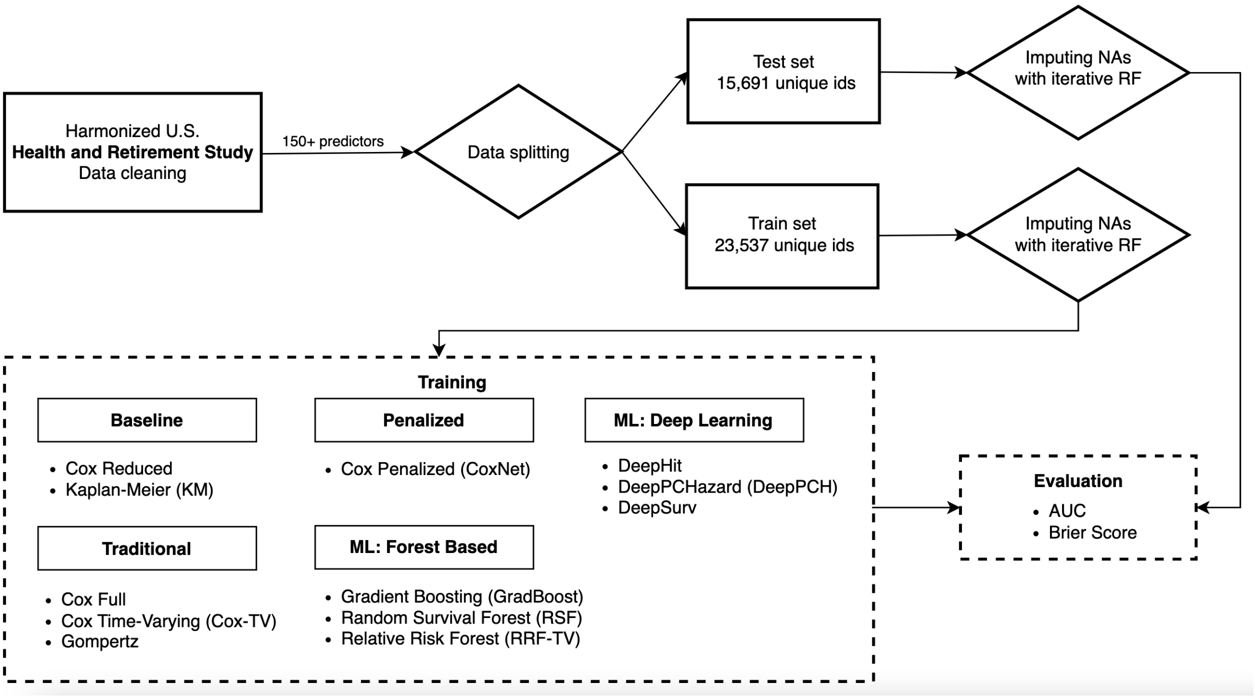

Figure 4: Data, methods, and evaluation flowchart diagram.

the analyses across gender, race and ethnicity, and education. Gender is categorized as either men or women, as reported in the HRS survey. Race and ethnicity is a categorical variable with three options: Non-Hispanic white, Non-Hispanic Black, and Hispanic. Education is a categorical variable which follows the simplified version of the International Standard Classification of Education scale (ISCED): low (less than high school), middle (upper secondary and vocational training), and high (tertiary education).

Models

We model each individual's survival curve as a function of time in the study from 0 at time of entry to 27 years, which is the maximum follow-up time in the HRS. We use time in study instead of age as the outcome in order to alleviate survivorship bias arising from left-truncation resulting from subjects entering the study at different ages. While left truncation is accounted for in most off-the-shelf Cox proportional hazard model software, many packages implementing machine learning survival models, including widely popular random survival forests, are not currently equipped to handle time-to-event data with left truncation. We categorize the methods we implement into five broad categories, which span a range of established and widely-used survival analysis models, in addition to more recently-developed machine learning models. We consider 12 models in total. Further remarks detailing the specification and implementation of the models presented here, including discussion of hyperparameter tuning, are included in the Supporting Information.

Baseline

The first class, which we refer to as the "Baseline" category, consists of two simple and popular survival models: the nonparametric Kaplan-Meier estimate (Kaplan and Meier, 1958) of the sample survival curve, which does not take into account covariates; and a Cox proportional hazards model (Cox, 1972) (hereafter referred to simply as a Cox model) fitted to a reduced set of covariates comprising the subject's age, gender, race, and education level, which we term "CoxReduced". These models are implemented using Python's scikit-survival package (Pölsterl, 2020) and R's survival package (Therneau, 2022), respectively. Baseline models serve as a benchmark on which the performance of increasingly complex models can be compared.

Traditional

The second class, which we refer to as the "Traditional" category, consists of classical parametric and semi-parametric survival models. The first of these is a Cox model fitted to the full set of predictors from each subject's first survey wave, termed "CoxFull". In practice, due to the large number of predictors, we apply a small amount of ridge (or ℓ_2) regularization to the partial likelihood to assuage singularities in the design matrix. This model is "time-invariant" in the sense that it uses only the covariate information provided at the time of entry to the study, i.e., the records from the first HRS wave in which an individual appears. Going forward, any models described should be assumed time-invariant in this sense unless otherwise stated. The remaining models are also fitted to the full set of predictors. Next, we fit a time-varying Cox model, denoted "Cox-TV", which makes full use of the longitudinal nature of the HRS, allowing covariates to vary across waves of the survey. Standard software implementations of the Cox model can directly handle time-varing covariates and left-truncated data using the Andersen-Gill counting process data formulation (Andersen and Gill, 1982). Finally, we fit a fully parametric Gompertz regression model (Gompertz, 1825), denoted "Gompertz", which is widely used to study mortality and assumes a log-linear baseline hazard. CoxFull, in constrast to CoxNet discussed in the penalized section, is implement within scikit-survival (Pölsterl, 2020). Cox-TV is implemented within the survival package (Therneau, 2022). The Gompertz model is implemented within Python's PySurvival package (Fotso et al., 2019).

Penalized

Despite its proximity to the classical Cox model, we differentiate the Cox model with an elastic net penalty (Park and Hastie, 2007; Simon et al., 2011), denoted "CoxNet", into its own category. Penalized likelihood methods, which induce shrinkage estimation and variable selection, are often included within the broad category of machine learning methods. CoxNet is implemented within scikit-survival (Pölsterl, 2020).

Machine Learning: Forest Based

The third class, termed "Machine Learning", consists of modern deep learning and decision-tree-based ensemble methods for survival analysis. The decision-tree-based ensemble methods that we fit include random survival forest (Ishwaran et al., 2008), denoted "RSF", and gradient-boosted trees with Cox proportional hazards loss (Friedman, 2002), denoted "GradBoost", both implemented within scikit-survival (Pölsterl, 2020). Tree ensemble models are flexible non-parametric methods that demonstrate superior performance in regression tasks with tabular data (Shwartz-Ziv and Armon, 2022). In particular, RSF does not make the strong proportional hazards assumption of the Cox model, which specifies that the hazard ratio between any two subjects remains constant over time. In the last few years, there has been work to extend popular machine learning survival models to handle data with left truncation and time-varying covariates (Fu and Simonoff, 2016; Moradian et al., 2022; Wongvibulsin et al., 2020; Yao et al., 2022). We fit the dynamic relative risk forest (RRF) of (Yao et al., 2022), an extension of the RRF (Ishwaran et al., 2004) to time-varying and left-truncated data, denoted here as "RRF-TV", which is implemented within the LTRCForests R package.

Machine Learning: Deep Learning

Finally, we introduce the three deep learning models implemented. DeepSurv is a nonlinear Cox model parameterizing the log-hazard via a deep neural network (Katzman et al., 2018). Although it allows for a nonlinear log-hazard function, DeepSurv still makes the proportional hazards assumption of the Cox model. The second deep learning model, DeepHit, makes no assumptions about the stochastic process of event times and instead estimates the distribution of survival times based on the covariates (Lee et al., 2018). While DeepHit can handle multiple competing risks, in our context we use it for a single risk case, mortality. Finally, DeepPCH models the continuous-time hazards by piece-wise constant functions of the covariates parametrized by neural networks (Kvamme and Borgan, 2019). The introduction of neural networks in these deep learning survival models allows for greater flexibility in modeling the survival curve, weakening the assumptions of the classical survival models and potentially capturing nonlinear interactions within the data. All of the deep learning models are implemented within the pycox package, which enables training survival models with PyTorch (Kvamme et al., 2019).

Model evaluation

Prediction metrics

After fitting each of the above models on the training dataset, we assess their predictive performance on the test set. Namely, for each individual in the test set, we generate a predicted survival curve from each model. We evaluate the accuracy of these survival curves with two widely-used survival prediction metrics – the time-dependent Brier score (BS) and area under the receiver operating characteristic curve (AUC), which assess the calibration and discrimination of a predictive model, respectively (Royston and Altman, 2013). We calculate the Brier score and AUC using Python's scikit-survival package (Pölsterl, 2020).

The Brier score is a strictly proper scoring rule (Gneiting and Raftery, 2007) that assesses model calibration by comparing the predicted survival curve at each time point to the subject's observed survival status using a squared error loss adjusted for the censoring distribution. The time-dependent Brier score is averaged over time to obtain an aggregate measure of predictive accuracy, the integrated Brier score. The time-dependent Brier score ranges from 0 to 1 with lower scores indicating better model calibration. A Brier score above 0.25 indicates predictive accuracy worse than a random "coin flip" prediction, which assigns a 0.5 probability of death to every individual at each time point. Essentially, the Brier score measures the average difference between the actual outcome and the outcome forecasted by our model.

The AUC assesses the discriminative performance of a survival model by comparing the risk scores assigned to pairs of subjects in relation to their observed survival status at each time point, while adjusting for the censoring distribution. A model discriminates appropriately if subjects living longer are assigned lower risk scores. The time-dependent AUC is averaged over time to yield an aggregate score, the mean AUC. The time-dependent AUC ranges from 0 to 1, with higher scores indicating a more discriminative model. An AUC below 0.5 indicates no discrimination, i.e., worse performance than a model that assigns the same risk score to each individual. Adopting the terminology of Mandrekar (2010), we consider an AUC within 0.7-0.8 as acceptable, 0.8-0.9 as excellent, and 0.9-1.0 as outstanding. The AUC is similar to the popular concordance index (or c-index), which we do not use due to its impropriety in our context (Blanche et al., 2018). If the Brier score measures how far the predictions are from the outcomes, the AUC denotes how common false positives, or in our case individuals incorrectly classified as dead, and false negatives, or individuals incorrectly classified as alive, are. This is an important metric because a model which predicted survival in every case could have a high accuracy when tested on a dataset where the majority of people survived, but would have a low AUC.

Variable importance

To interpret how each model classifies individuals in terms of survival we compute variable importance using permutation importance with negative integrated Brier score as the scoring metric. Permutation importance measures how model accuracy changes when a given variable is randomly shuffled (Breiman, 2001), thereby determining the predictors having the greatest impact on the model predictions. We run these comparisons for each model using all participants, as well as for each model separating participants by gender, race and ethnicity, and education level. We use the permutation importance method implemented within Python's scikit-learn package (Pedregosa et al., 2011) and change the scoring metric to the integrated Brier score with a custom scoring function.

Data Availability

The raw harmonized Health and Retirement Study (HRS) data used in this study are publicly available and can be downloaded from
https://hrsdata.isr.umich.edu/data-products/gateway-harmonized-hrs (HRS, 2022). All data processing, modeling, evaluation, and visualization were carried out in R (RStudio Team, 2019) and Python (Van Rossum and Drake Jr, 1995). All code is available and can be reproduced from Github: https://github.com/mpidr-mort-pred/Replication-Package.

Acknowledgments

We thank Jutta Gampe for providing valuable suggestions during research design, and Alyson van Raalte for useful comments. We also thank participants of the 2023 Giornate di Studio sulla Popolazione and the 2023 Annual Meeting of the Population Association of America for their feedback. This project was funded by the 2022 Summer Incubator Program of the Max Planck Institute for Demographic Research, and we thank all participants for their comments and suggestions.

References

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Number 46. Sage.

Alvarez, J.-A., Villavicencio, F., Strozza, C., and Camarda, C. G. (2021). Regularities in human mortality after age 105. *PloS one*, 16(7):e0253940.

Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100 – 1120.

Arpino, B., Le Moglie, M., and Mencarini, L. (2022). What Tears Couples Apart: A Machine Learning Analysis of Union Dissolution in Germany. *Demography*, 59(1):161–186.

Barbi, E., Lagona, F., Marsili, M., Vaupel, J. W., and Wachter, K. W. (2018). The plateau of human mortality: Demography of longevity pioneers. *Science*, 360(6396):1459–1461.

Blanche, P., Kattan, M. W., and Gerds, T. A. (2018). The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357.

Breen, C. and Seltzer, N. (2022). Using machine learning algorithms to predict longevity. *[Conference presentation]. ASA 2022 Conference, August 5-9 2022, Los Angeles, CA, United States.*

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care — addressing ethical challenges. *New England Journal of Medicine*, 378(11):981–983.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4(1):123–144.

Christensen, K., Johnson, T. E., and Vaupel, J. W. (2006). The quest for genetic determinants of human longevity: challenges and insights. *Nature Reviews Genetics*, 7(6):436–448.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dong, L., Ratti, C., and Zheng, S. (2019). Predicting neighborhoods' socioeconomic attributes using restaurant data. *PNAS*, 116(31):15447–15452.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Einav, L., Finkelstein, A., Mullainathan, S., and Obermeyer, Z. (2018). Predictive modeling of us health care spending in late life. *Science*, 360(6396):1462–1465.

Fotso, S. et al. (2019). PySurvival: Open source package for survival analysis modeling.

Francesco, D. D., Reiss, J. D., Roger, J., Tang, A. S., Chang, A. L., Becker, M., Phongpreecha, T., Espinosa, C., Morin, S., Berson, E., Thuraiappah, M., Le, B. L., Ravindra, N. G., Payrovnaziri, S. N., Mataraso, S., Kim, Y., Xue, L., Rosenstein, M. G., Oskotsky, T., Marić, I., Gaudilliere, B., Carvalho, B., Bateman, B. T., Angst, M. S., Prince, L. S., Blumenfeld, Y. J., Benitz, W. E., Fuerch, J. H., Shaw, G. M., Sylvester, K. G., Stevenson, D. K., Sirota, M., and Aghaeepour, N. (2023). Data-driven longitudinal characterization of neonatal health and morbidity. *Science Translational Medicine*, 15(683):eadc9854.

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378.

Fu, W. and Simonoff, J. S. (2016). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, 18(2):352–369.

Gavrilov, L. A. and Gavrilova, N. S. (2011). Mortality measurement at advanced ages: a study of the Social Security Administration Death Master File. *North American Actuarial Journal*, 15(3):432–447.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goldman, N., Glei, D., and Weinstein, M. (2016). What matters most for predicting survival? a multinational population-based cohort study. *PLoS ONE*, 11(7):e0159273.

Goldman, N., Glei, D. A., and Weinstein, M. (2017). The best predictors of survival: Do they vary by age, sex, and race? *Population and Development Review*, 43(3):541–560.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, 115:513–583.

Heller, S., Jakubowski, B., Jelveh, Z., and Kapustin, M. (2022). Machine Learning Can Predict Shooting Victimization Well Enough to Help Prevent It. Technical Report w30170, National Bureau of Economic Research, Cambridge, MA.

Henderson, R. and Keiding, N. (2005). Individual survival time prediction using statistical models. *Journal of Medical Ethics*, 31(12):703–706.

Hofman, J. M., Sharma, A., and Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324):486–488.

Horiuchi, S. and Wilmoth, J. R. (1998). Deceleration in the age pattern of mortality at older ages. *Demography*, 35:391–412.

HRS (2022). Health and Retirement Study RAND HRS Longitudinal File 2018 (V2) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI. , .

Hummer, R. A. (2023). Race and ethnicity, racism, and population health in the united states: The straightforward, the complex, innovations, and the future. *Demography*, 60(3):633–657.

Ishwaran, H., Blackstone, E. H., Pothier, C. E., and Lauer, M. S. (2004). Relative risk forests for exercise heart rate recovery as a predictor of mortality. *Journal of the American Statistical Association*, 99(467):591–600.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3).

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Joel, S., Eastwick, P. W., Allison, C. J., and Wolf, S. (2020). Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies. *PNAS*, 117(32):19061–19071.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1).

Koenecke, A., Nam, A., Lake, E., and Goel, S. (2020). Racial disparities in automated speech recognition. *PNAS*, 117(14):7684–7689.

Kvamme, H. and Borgan, Ø. (2019). Continuous and discrete-time survival prediction with neural networks.

Kvamme, H., Ørnulf Borgan, and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594.

Lee, C., Zame, W., Yoon, J., and van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., and Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, 8(11):eabj1812.

Mackenbach, J. P., Valverde, J. R., Artnik, B., Bopp, M., Brønnum-Hansen, H., Deboosere, P., Kalediene, R., Kovács, K., Leinsalu, M., Martikainen, P., et al. (2018). Trends in health inequalities in 27 european countries. *Proceedings of the National Academy of Sciences*, 115(25):6440–6445.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*, 51:584–591.

Moradian, H., Yao, W., Larocque, D., Simonoff, J. S., and Frydman, H. (2022). Dynamic estimation with random forests for discrete-time survival data. *Canadian Journal of Statistics*, 50(2):533–548.

Newman, S. J. (2018). Errors as a primary cause of late-life mortality deceleration and plateaus. *PLoS biology*, 16(12):e2006776.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Oeppen, J. and Vaupel, J. W. (2002). Broken limits to life expectancy. *Science*, 296(5570):1029–1031.

Olshansky, S. J., Carnes, B. A., and Cassel, C. (1990). In search of Methuselah: estimating the upper limits to human longevity. *Science*, 250(4981):634–640.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perozek, M. (2008). Using subjective expectations to forecast longevity: Do survey respondents know something we don't know? *Demography*, 45(1):95–113.

Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6.

Puterman, E., Weiss, J., Hives, B. A., Gemmill, A., Karasek, D., Mendes, W. B., and Rehkopf, D. H. (2020). Predicting mortality from 57 economic, behavioral, social, and psychological factors. *Proceedings of the National Academy of Sciences*, 117(28):16273–16282.

Rau, R., Soroko, E., Jasilionis, D., and Vaupel, J. W. (2008). Continued reductions in mortality at advanced ages. *Population and Development Review*, 34(4):747–768.

Riley, J. C. (2001). *Rising life expectancy: a global history*. Cambridge University Press.

Royston, P. and Altman, D. G. (2013). External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(33).

RStudio Team (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5).

Stekhoven, D. J. and Buhlmann, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Thatcher, A. R., Kannisto, V., and Vaupel, J. W. (1998). *The Force of Mortality at Ages 80 to 120*. Monographs on Population Aging, 5. Odense University Press, Odense.

Therneau, T. M. (2022). *A Package for Survival Analysis in R*. R package version 3.4-0.

van Raalte, A. A., Kunst, A. E., Deboosere, P., Leinsalu, M., Lundberg, O., Martikainen, P., Strand, B. H., Artnik, B., Wojtyniak, B., and Mackenbach, J. P. (2011). More variation in lifespan in lower educated groups: evidence from 10 european countries. *International journal of epidemiology*, 40(6):1703–1714.

van Raalte, A. A., Sasson, I., and Martikainen, P. (2018). The case for monitoring life-span inequality. *Science*, 362(6418):1002–1004.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wilmoth, J. R., Deegan, L. J., Lundstrom, H., and Horiuchi, S. (2000). Increase of maximum life-span in Sweden, 1861-1999. *Science*, 289(5488):2366–2368.

Wongvibulsin, S., Wu, K. C., and Zeger, S. L. (2020). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (rf-slam) data analysis. *BMC Med Res Methodol*, 20(1).

Yao, W., Frydman, H., Larocque, D., and Simonoff, J. S. (2022). Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research*, . PMID: 35895510.