Supplementary Methods

Supplementary Methods 1 Data collection and processing

The National Health and Nutrition Examination Survey (NHANES) from the National Center for Health Statistics (NCHS)[1] conducts interviews and physical examinations to assess the health and nutrition data for all ages in the United States. The interviews include demographic, socioeconomic, dietary, and health-related questions. The examinations include medical, dental, physiological measurements, and laboratory tests administered by highly trained medical personnel. Since 1999, data were collected and released at 2-year intervals. Each year NHANES examines a nationally representative sample of roughly 5,000 individuals across the Unites States. In this study, we include NHANES data sampled between 1999 and 2014. All-cause mortality is ascertained by a linked NHANES mortality file that provides follow-up mortality data from the date of survey participation through December 31, 2015.

Our study includes samples with known mortality status who participated in NHANES 1999-2014 ($n = 47,261$). We include all demographic, laboratory, examination, and questionnaire features that could be automatically matched across different NHANES cycles. We exclude variables that are missing for more than 50% of the participants and highly correlated features with correlations greater than 0.98; after filtering and one-hot encoding, 151 features remain (Supplementary appendix 2). We impute missing data using MissForest [5], a nonparametric random forest-based multiple imputation method for mixed-type data, with seven iterations. We predict all-cause mortality for two broad categories: (1) follow-up times of 1-year, 3-year, and 5-year and (2) age groups of $<$40, 40-65, 65-80, and $\geq$80 years old. For different follow-up times, we remove samples with unconfirmed mortality status. For different age groups, we predict 5-year mortality. The demographic characteristics and sample size of the data for different tasks are shown in Supplementary Table 1.

We use UK Biobank samples as an external validation dataset. Participants were enrolled in the UK Biobank from April, 2007, to July, 2010, from 21 assessment centres across England, Wales, and Scotland using standardised procedures. When participants agreed to take part in UK Biobank, they visited their closest assessment centre to provide baseline information, physical measures, and biological samples. We include the 51 features that are overlapping between NHANES and UK Biobank dataset (Supplementary appendix 2). We exclude samples with missing values. All-cause mortality included all deaths occurring before May, 2021. We include 384,762 samples aged 37-72 years with confirmed 5-year mortality status. Of these samples, 6,336 died after 5 years. The histograms of age, gender and body mass index of UK Biobank samples are shown in Supplementary Figure 2.

Supplementary Methods 2 Predictive modeling

To model mortality, we use gradient boosted trees (GBTs). GBTs are nonparametric methods composed of iteratively trained decision trees. The final ensemble of trees captures non-linearity and interactions between predictors. The dataset is randomly divided into training (80%) and testing (20%) sets. We use the implementation XGBoost [1][2] with a learning rate set to 0.002 , subsample ratio set to 0.5 and 10,000 trees of max depth 3. For comparison, we also train logistic regression models and deep neural networks. For logistic

[1] http://www.cdc.gov/nchs/nhanes.htm
[2] https://xgboost.readthedocs.io/en/latest/python/index.html

regression, we use L2 regularization. The L2 regularization weight was set to 100. For neural networks, we use a single layer with 1,000 nodes, and max iteration set to 1,000. The hyperparameters specified above are chosen by GridSearch and 5-fold cross validation. Other hyperparameter values are left at their default values. Models' performance is measured with the area under the receiver operator characteristic curve (AUROC). We bootstrap the test set to assess the statistical significance of the difference in AUC for pairs of models. Specifically, we resample with replacement from the test set 1,000 times and compare the models' performance on resampled test sets. We report a p-value which is the percentage of time that logistic regression or the neural network's performance is better than or equal to gradient boosted trees, divided by the number of resampled test sets. All models are built using the Scikit-learn package in Python 3.7.

Supplementary Methods 3 Model interpretation

To explain the GBT models, we utilize TreeExplainer [4], which provides a local explanation of the impact of input features on individual predictions. Specifically, TreeExplainer calculates exact SHAP [3] (SHapley Additive exPlanations) values for tree-based models. When explaining the mortality prediction models, we randomly select 10,000 background samples from the training set and 5,000 foreground samples from the test set.

Supplementary Methods 3.1 SHAP (SHapley Additive exPlanation) values

SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. The change of the model's prediction when the feature is masked is recorded across all possible subsets of features, yielding an average change in prediction resulting from the inclusion of a feature in the model:

$$\phi_i(f, x) = \sum_{R \in \mathcal{R}} \frac{1}{M!} \left[f_x(P_i^R \cup i) - f_x(P_i^R) \right], \tag{1}$$

where ϕ_i is the feature attribution (SHAP value) of feature i in model f for data point x, $\mathcal{R}$ is the set of all feature permutations, P_i^R is the set of all features before i in the ordering R, M is the number of input features, and f_x is an estimate of the conditional expectation of the model's prediction: $f_x(S) \approx \mathrm{E[f(x) \mid x_S]}$ where x_S is the set of observed features.

SHAP values which guarantee a set of desirable theoretical properties, including additivity and consistency. Additivity states that when approximating the original model f for a specific input x, the SHAP values sum up to the output $f(x)$:

$$f(x) = \phi_0(f) + \sum_{i=1}^{M} \phi_i(f, x), \tag{2}$$

The sum of feature attributions (SHAP values) matches the original model output f(x), where $\phi_0(f) = \mathrm{E[f(z)]} = \mathrm{f_x(\emptyset)}$. Consistency states that if a model changes so that some feature's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease. Therefore, SHAP values are consistent and accurate calculations of each feature's contribution to the model's prediction.

Supplementary Methods 3.2 SHAP interaction values and main effects

The SHAP interaction effects is based on the Shapley interaction index from game theory. While standard feature attribution results in a vector of values, one for each feature, attributions based on the Shapley interaction index result in a matrix of feature attributions. The main effects are on the diagonal and the interaction effects on the off-diagonal. The **SHAP interaction values** are defined as:

$$\Phi_{i,j}(f,x) = \sum_{S \subseteq \mathcal{M} \setminus \{i,j\}} \frac{|S|!(M-|S|-2)!}{2(M-1)!} \nabla_{i,j}(f,x,S), \tag{3}$$

when $i \neq j$, and

$$\nabla_{ij}(f,x,S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S). \tag{4}$$

where $\mathcal{M}$ is the set of all M input features. In Equation 3 the SHAP interaction value between feature i and feature j is split equally between each feature so $\Phi_{i,f}(f,x) = \Phi_{j,i}(f,x)$ and the total interaction effect is $\Phi_{i,f}(f,x) + \Phi_{j,i}(f,x)$.

The **main effects** for a prediction can then be defined as the difference between the SHAP values and the off-diagonal SHAP interaction values for a feature:

$$\Phi_{i,i}(f,x) = \phi_i(f,x) - \sum_{j \neq i} \Phi_{i,j}(f,x). \tag{5}$$

Supplementary Methods 3.3 Partial dependence plots and additional perspective to reference interval

We use partial dependence plots to show the change in mortality risk for all values of a laboratory feature. Partial dependence plots show the marginal effect one feature has on the predicted outcome of a machine learning model. The relative mortality risk is defined as the average value of the model predicted probability when we fix a specific feature to a given value divided by the average value of the model predicted probability. The relative risk percentage is the maximum relative risk for the values within the reference interval divided by the maximum relative risk for all values of a laboratory feature. High relative risk percentage indicates that the values within the reference interval have a relatively high mortality risk. The partial dependence plots of selected laboratory feature values on 1-, 3-, and 10-year mortality risk are shown in Supplementary Figure 7.

Supplementary Methods 4 Model interpretation plots

In this section we describe a number of plotting types for model explanation visualization.

SHAP value, SHAP main effect value and SHAP interaction value plots In SHAP value/SHAP main effect value/SHAP interaction value plots, every point corresponds to a single sample where the x-axis is the value of the feature and the y-axis is the SHAP value/SHAP main effect value/SHAP interaction value. The coloring of the points often denotes the value of a separate feature.

Summary plot Summary plots show the feature attributions (SHAP values) for many samples and multiple features in order of global feature importance (the mean absolute SHAP values). Summary plots stack

multiple subplots for each feature. For the feature plots, every point corresponds to a single sample where the x-axis is the feature attribution value and the y-axis is vertical dispersion representing the frequency of samples with a particular feature attribution value. Finally, the color of each point represents the normalized feature value, with red representing a high value and blue representing a low one. Intermediary feature values are interpolations between red and blue.

Individualized explanation plot Individualized explanation plots show the feature attributions (SHAP values) for an individual in terms of how they drive the model's prediction for the individual away from the average model prediction across the baseline distribution. The width of the bars indicate the SHAP value with red indicating a positive affect and blue indicating a negative one. The features corresponding to the largest bars are below with their actual values for the individual.

Supplementary Methods 5 Supervised distance

Supplementary Methods 5.1 Supervised distance and hierarchical clustering

Supervised distance can accurately measure feature redundancy based on a specific prediction task. As Supplementary Figure 4 shows, to calculate the supervised distance between feature i and feature j, we first train a uni-variate GBT model to predict the label (e.g. 5-year mortality in our study) using feature i. Then, we can obtain the $Prediction_i$ which is the output of the fitted uni-variate GBT. Next, we fit another uni-variate GBT to predict $Prediction_i$ using feature j. We define the output of the new GBT as $Prediction_i^j$. All hyperparameter values of the uni-variate GBTs are set to their default values. Following the same above steps, we can obtain $Prediction_j^i$. The supervised distance between feature i and feature j (supervised distance(i,j)) is defined as:

$$supervised\,R^2(i,j) = max(0, 1 - mean(\frac{(Prediction_i - Prediction_i^j)^2}{var(Prediction_i)})) \tag{6}$$

$$supervised\,distance(i,j) = max(1 - supervised\,R^2(i,j), 1 - supervised\,R^2(j,i)) \tag{7}$$

where $var(x)$ is the variance of the vector x, $mean(x)$ is the average of the vector x. Supervised distance is scaled roughly between 0 and 1, where 0 distance means the features perfectly redundant and 1 means they are completely independent.

To explore the redundant feature groups, we hierarchically cluster all features according to the supervised distance. Specifically, we use complete linkage hierarchical clustering which merges in each step the two clusters whose merger has the smallest diameter. The hierarchical clustering tree is shown in Supplementary Figure 5.

Supplementary Methods 5.2 Redundant feature groups experiments training details

Reducing redundancy model To identify the most representative feature in a redundant feature group, we train GBTs using one feature in the redundancy group and all features outside the group for 5-year mortality prediction. Then we compare the feature importance ranking of the redundant features by calculating the mean absolute SHAP values using TreeExplainer. The hyperparameters of the GBTs are chosen

by GridSearch and 5-fold cross validation. The max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. Other hyperparameter values are left at their default values.

Single feature model We further analyze the predictive power of the redundant features by fitting 5-year mortality prediction GBTs using one feature in the redundant feature group. Specifically, we use one feature in the redundant feature group and two important confounders, age and gender, to train a GBTs for 5-year mortality prediction. All hyperparameter values are set to their default values. We compare the AUCs of the models. We bootstrap the test set for 1,000 times and compare the models' performance on resampled test sets. The averages of the AUCs are reported.

Supplementary Methods 5.3 Supervised distance-based feature selection

,—We propose a supervised distance-based feature selection method to select predictive and less-redundant feature sets. The workflow of our feature selection method is shown in Supplementary Figure 4. The dataset is randomly divided into training (80%) and testing (20%) sets. Firstly, we fit a GBT for 5-year mortality prediction on all features using the training set and rank the features by mean absolute SHAP values from TreeExplainer. The hyperparameters of the GBTs are chosen by GridSearch and 5-fold cross validation. The max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. The max number of trees is set to 1000. We use 20% of the training samples as validation set for early stopping. The number of early stopping rounds is set to 100. Since age and gender are important confounders, we would like to keep them in the selected feature set. Therefore, we cluster features except age and gender into a specific number of groups using supervised distances-based hierarchical clustering and select the most important feature in each cluster. Then, we add age and gender to the selected feature set and re-fit the model. Next, we rerun the clustering using the new feature set except age and gender. This process is repeated until all remaining features cluster to a single group. In every iteration, we remove 5 features. The models are evaluated on the testing set with bootstrapping for 1,000 times. We report the average of the AUCs and the minimum supervised distance within the selected feature sets. The selected features in each iteration are listed in Supplementary Appendix 1.

Supplementary Methods 6 5-year mortality risk scores

Supplementary Methods 6.1 Mortality risk scores training details

IMPACT mortality risk scores are defined to be the prediction of the 5-year mortality prediction models. For comparison, we train linear[3] and gradient boosted tree-based Cox proportional hazard models[4]. We do a temporal validation of the risk scores by assessing their performances in the samples collected in 2009-2014 ($N = 7,034$). Specifically, the samples collected in 1999-2008 ($N = 28,820$) are randomly divided into training (80%) and testing (20%) sets. To compare with Intermountain gender-specific risk scores, we evaluate the models on different gender groups. The models are trained on the whole training set and evaluate on different gender groups in the testing set. Furthermore, considering the different feature collection cost for the general public and medical professionals, we build the risk scores starting from different feature sets. For the general public, the models are trained on all demographics, questionnaire features and examination

[3]https://scikit-survival.readthedocs.io/en/latest/api/generated/sksurv.linear_model.CoxPHSurvivalAnalysis.html
[4]https://scikit-survival.readthedocs.io/en/latest/api/generated/sksurv.ensemble.GradientBoostingSurvivalAnalysis.html

features that are accessible at home for general public, For medical professionals, the models are trained on all demographics and laboratory features. All trained models are evaluated on different gender groups of the samples collected in 2009-2014 for temporal validation.

The hyperparameters are chosen by GridSearch and 5-fold cross validation. For XGBoost 5-year mortality prediction models, the max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. The max number of trees is set to 1000. We use 20% of the training samples as validation set for early stopping. The number of early stopping rounds is set to 100. For linear Cox proportional hazard models, the regularization parameter α is selected from $\{0.01, 0.1, 1, 10, 100\}$. For tree-based Cox proportional hazard models, the max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. Other hyperparameter values are left at their default values.

We explain the mortality prediction model in terms of its probability predictions. Specifically, we rescale the SHAP values (in the log-odds space) to be in the probability space directly. The rescaled SHAP values now sum to the probability output of the model.

To compare with the popular mortality risk scores and biological ages, we repeat the same process for 1-year and 10-year mortality prediction. For 1-year mortality prediction, we do a temporal validation in the samples collected in 2013-2014 ($N = 6,082$). For 10-year mortality prediction, we do a temporal validation in the samples collected in 2005-2014 ($N = 4,945$).

Supplementary Methods 6.2 Recursive feature elimination

Recursive feature elimination works by searching for a subset of features by starting with all features in the training dataset and successively removing features until the desired number of features remains. Firstly, we train a model on the full dataset with all features. Then we rank features by importance (mean absolute SHAP values) and remove the least important features. Another model is trained on the resulting feature set, and the process iterates until only the desired number of features are left. Starting from 151 features, we remove 6 features at the first iteration. Then, We remove 5 features in each iteration until only one feature is left. We bootstrap the test set and assess the predictive performance. Specifically, we resample with replacement from the test set 1,000 times and report the average and the 95% confidence interval of the AUCs. The selected features in each iteration are listed in Supplementary appendix 2.

Supplementary Methods 6.3 Intermountain mortality risk scores and exhaustive feature selection

Intermountain mortality risk scores [2] are built using complete blood count and basic metabolic profile. Specifically, 13 laboratory features are used to predict 30 days, 1-year and 5-year mortality. Logistic regression was used to model the risk prediction equations with adjustment for age and sex. Dummy variables modeled each category, with the referent defined as the lowest risk group (except for age categories: 18-29, 30-39, 40-49 [referent], 50-59, 60-69, 70-79, and $\geq$80 years). A scalar score value was derived for each variable category by multiplying its β-coefficient by 3 and rounding to the nearest integer (referent value = zero). Each individual's risk score became the sum of the score values based on his or her individual data.

We implement exhaustive search to select features of Intermountain risk scores. The number of features ranges from 1 to 14 (including age). Given the number of features, we search all possible feature combinations. The risk score becomes the sum of the score values of the selected features. The 5-year mortality risk scores are evaluated on the training set. We select the feature combination that achieve the highest AUC on the

training set. Then, the risk scores of the selected feature combinations are evaluated on the testing set with bootstrapping for 1,000 times.

References

[1] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[2] Benjamin D Horne et al. "Exceptional mortality prediction by risk scores from common laboratory tests". In: *The American journal of medicine* 122.6 (2009), pp. 550–558.

[3] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems*. 2017, pp. 4765–4774.

[4] Scott M Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 2522–5839.

[5] Daniel J Stekhoven and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1 (2012), pp. 112–118.