nature communications

Article

https://doi.org/10.1038/s41467-025-61891-y

# Comprehensive interaction modeling with machine learning improves prediction of disease risk in the UK Biobank

Received: 30 July 2024

Accepted: 3 July 2025

Published online: 18 July 2025

Check for updates

Heli Julkunen ✉ & Juho Rousu

Understanding how risk factors interact to jointly influence disease risk can provide insights into disease development and improve risk prediction. Here we introduce survivalFM, a machine learning extension to the widely used Cox proportional hazards model that enables scalable estimation of all potential pairwise interaction effects on time-to-event outcomes. The method approximates interaction effects using a low-rank factorization, allowing it to overcome the computational and statistical limitations typically associated with high-dimensional interaction modeling. Applied to the UK Biobank dataset across nine disease examples and diverse clinical and omics risk factors, survivalFM improves prediction performance in terms of discrimination, explained variation, and reclassification in 30.6%, 41.7%, and 94.4% of the scenarios tested, respectively. In a clinical cardiovascular risk prediction scenario using the established QRISK3 model, the method adds predictive value by identifying interactions beyond the age interaction effects currently included. These results demonstrate that comprehensive modeling of interactions can facilitate advanced insights into disease development and improve risk predictions.

Risk prediction models are needed in modern preventive medicine to identify individuals at high risk of disease before clinical symptoms manifest. The ability to predict disease risk is particularly important in managing complex diseases, such as cardiovascular disease, chronic kidney disease, and diabetes, where early intervention can substantially alter patient outcomes. However, accurately predicting disease risk is challenging due to the inherent complexity of most human diseases, which arise from the interplay of genetic, environmental, and lifestyle factors. Traditional methods in survival analysis, such as the widely used Cox proportional hazards regression[1], assume linear effects of predictor variables on time-to-event outcomes. This assumption may lead to oversimplified models that overlook the complex interplay among predictors, potentially missing important biological insights and limiting risk prediction accuracy.

The accuracy of time-to-event prediction models can be improved by incorporating interaction terms, a well-established concept in epidemiology to assess the joint effects of predictors on outcomes[2,3]. For instance, interaction terms have been shown to be relevant in cardiovascular disease (CVD) risk prediction, where the effects of other risk factors can vary depending on age[4–7]. However, incorporating these terms in multivariable prediction models typically requires prior hypotheses about which interactions to include. Since the number of potential interaction terms increases quadratically with the number of predictor variables in consideration, inclusion of all potential interactions quickly becomes impractical without targeted hypotheses to guide the selection. Therefore, prior multivariable prediction models have typically been constrained to a restricted set of interaction terms known to alter outcome associations, such as those involving age. This limits the discovery of new, potentially relevant interactions. Another commonly employed strategy is to perform statistical testing of individual interaction terms, but this can miss interactions that only become relevant for prediction in the presence of other variables. This challenge becomes particularly pronounced

Department of Computer Science, Aalto University, Espoo, Finland. ✉e-mail: heli.julkunen@aalto.fi

with modern biomedical datasets, which can contain hundreds of potential predictors. While machine learning survival analysis extensions like random survival forests[8] and deep survival models[9,10] can capture complex non-linearities and interactions in the underlying data, they often compromise interpretability, which is crucial when the goal is to inform clinical decision-making or to obtain insights into the risk factors underlying disease development.

To enhance possibilities to understand and model the joint effects of risk factors on time-to-event disease outcomes, here we present survivalFM, a methodological extension to the Cox proportional hazards model that incorporates estimation of all potential pairwise interaction effects among predictor variables. The method is based on an efficient strategy of learning the interaction effects using a low-rank factorized approximation, a concept taken from factorization machines (FMs)[11] and here adapted to survival analysis. survivalFM combines the factorization of the interaction effects with an efficient quasi-Newton optimization algorithm, thereby overcoming the computational and statistical challenges of fitting comprehensive interaction effects in time-to-event prediction models involving many variables. The resulting model is fully interpretable, providing access to the estimates of both individual effects and the approximated interactions. We demonstrate the performance of survivalFM by developing and validating prediction models across various data modalities and disease outcomes using data from the UK Biobank. We further highlight an application in a clinical cardiovascular risk prediction scenario and show that survivalFM can learn predictive interaction effects which improve identification of high-risk individuals. While we highlight applications in disease risk prediction, the method is generally applicable to modeling any type of time-to-event outcomes.

## Results

### Overview of survivalFM

Figure 1 presents an overview of survivalFM. We developed survivalFM to estimate all potential pairwise interaction effects among input variables for right-censored survival data, such as time to disease

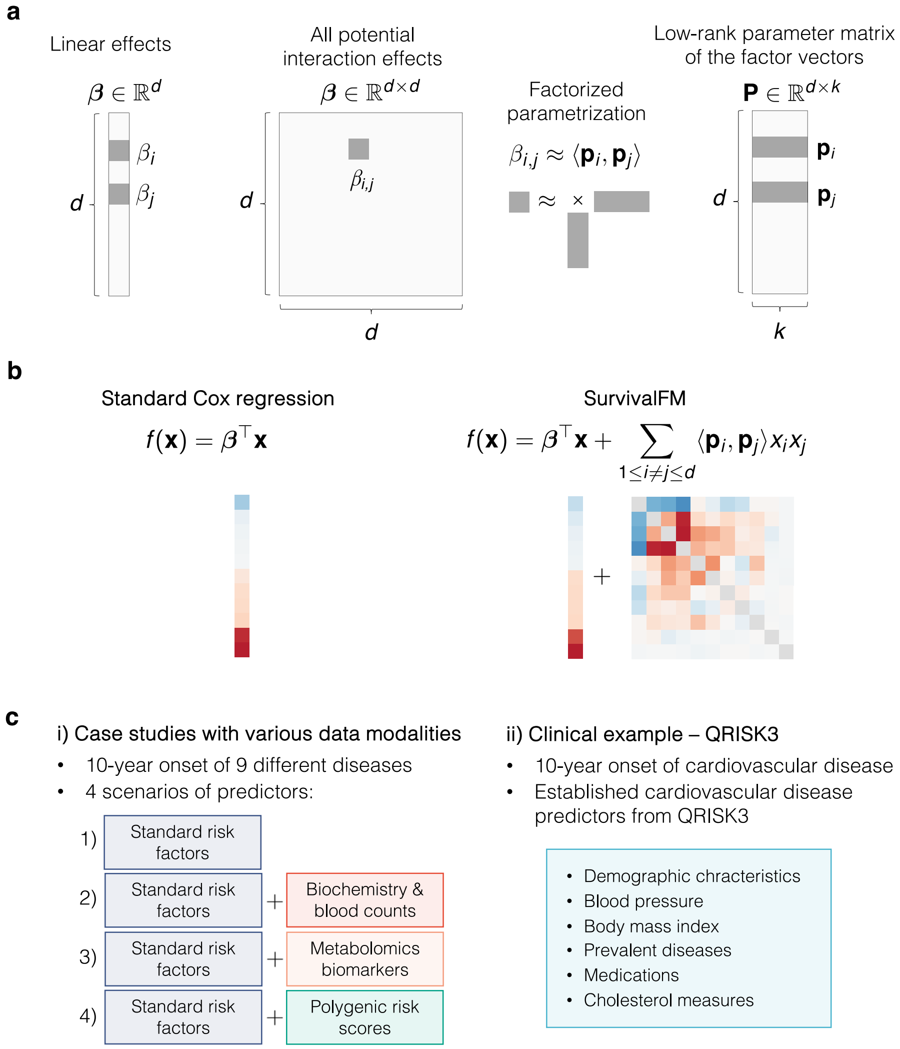
a
Linear effects
$\boldsymbol{\beta} \in \mathbb{R}^d$
$\beta_i$
$\beta_j$
$d$
All potential interaction effects
$\boldsymbol{\beta} \in \mathbb{R}^{d \times d}$
$\beta_{i,j}$
$d$
$d$
Factorized parametrization
$\beta_{i,j} \approx \langle \mathbf{p}_i, \mathbf{p}_j \rangle$
≈ ×
Low-rank parameter matrix of the factor vectors
$\mathbf{P} \in \mathbb{R}^{d \times k}$
$\mathbf{p}_i$
$\mathbf{p}_j$
$d$
$k$
b
Standard Cox regression
$f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$
SurvivalFM
$f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + \sum_{1 \le i \ne j \le d} \langle \mathbf{p}_i, \mathbf{p}_j \rangle x_i x_j$
+
c
i) Case studies with various data modalities
• 10-year onset of 9 different diseases
• 4 scenarios of predictors:
1) Standard risk factors
2) Standard risk factors + Biochemistry & blood counts
3) Standard risk factors + Metabolomics biomarkers
4) Standard risk factors + Polygenic risk scores
ii) Clinical example – QRISK3
• 10-year onset of cardiovascular disease
• Established cardiovascular disease predictors from QRISK3
• Demographic chracteristics
• Blood pressure
• Body mass index
• Prevalent diseases
• Medications
• Cholesterol measures

**Fig. 1 | Method overview and evaluation examples. a** A machine learning survival analysis method, survivalFM, is developed to estimate linear and all pairwise interaction effects between predictor variables using factorized parametrization of the interaction terms $\beta_{i,j} \approx \langle \mathbf{p}_i, \mathbf{p}_j \rangle$. $d$ denotes the number of predictor variables and $k$ is a hyperparameter defining the rank of the factorization of the interaction terms. Typically, the rank of the factorization is substantially lower than the number of predictor variables ($k \ll d$), which allows interaction effects to be computed efficiently even with high-dimensional input data. **b** The added value of incorporating comprehensive interaction terms using survivalFM is assessed by comparing its performance to the standard linear Cox proportional hazards regression. $f(\mathbf{x})$ parameterizes the partial hazard function in a proportional hazards model $h(t|\mathbf{x}) = h_0(t)\exp(f(\mathbf{x}))$. **c** The performance is evaluated in various disease prediction examples: i) case studies with four different predictor sets, each applied to nine disease examples; ii) a clinical example using predictors from the QRISK3 cardiovascular disease (CVD) risk evaluation tool.

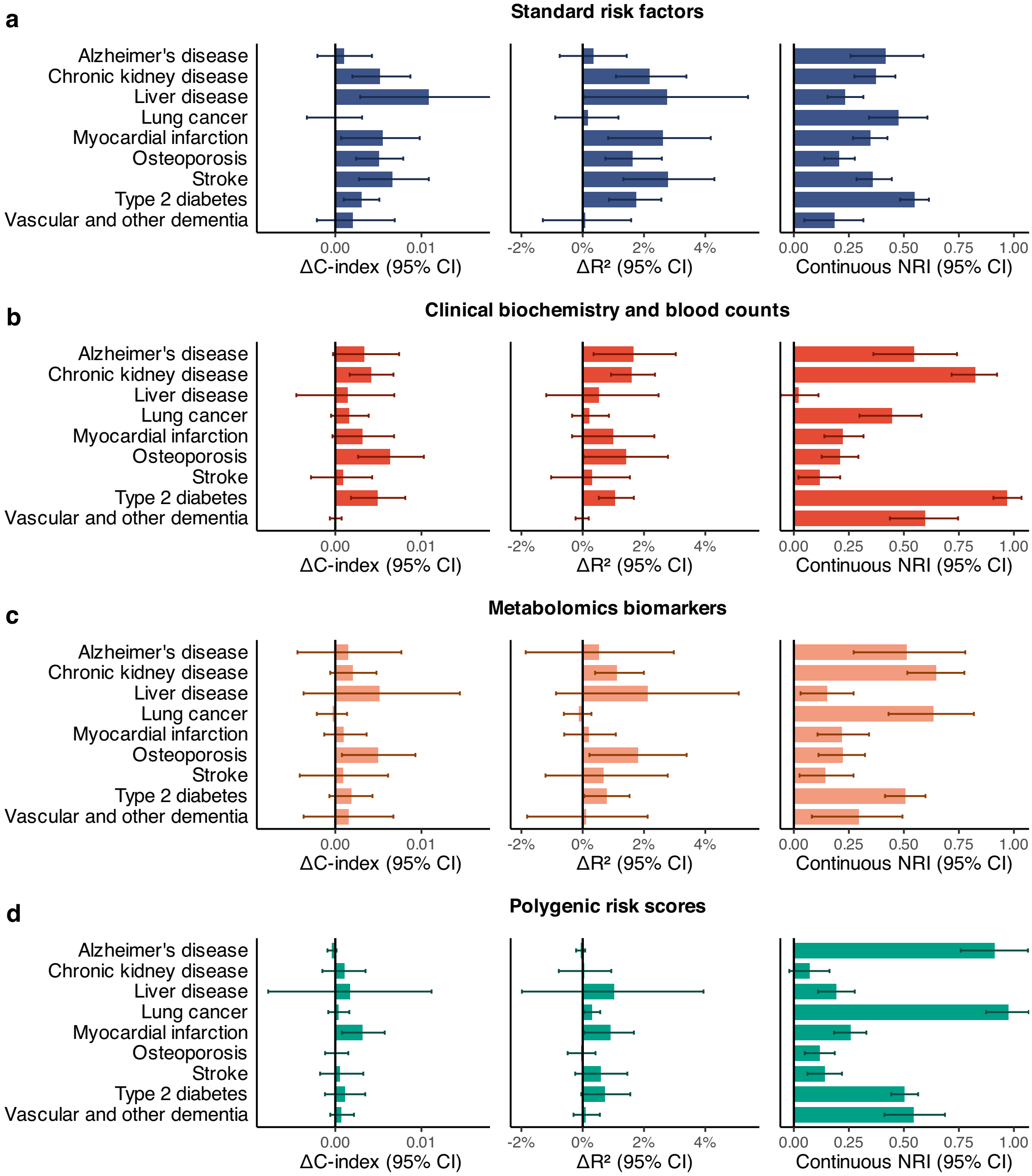

**Fig. 2 | Comprehensive interaction modeling by survivalFM improves risk prediction performance across various diseases and data modalities.** Comparison of the predictive performance of survivalFM to standard linear Cox proportional hazards regression in terms of improvements in concordance index (ΔC-index), Royston's $R^2$ ($\Delta R^2$) and continuous net reclassification (NRI). Results are shown for nine disease examples (y-axis) across four data modalities, computed for individuals in the Scotland test set: **a** standard risk factors (blue; included in all models), **b** clinical biochemistry and blood counts (red), **c** metabolomics biomarkers (orange) and **d** polygenic risk scores (green). Bars represent the observed value of the model performance metric; horizontal error bars denote 95% confidence intervals (CIs), estimated with bootstrapping over 1000 resamples. Sample sizes and event counts for each disease example are provided in Supplementary Data 3. Source data are provided as a Source Data file.

onset. It is based on the widely utilized proportional hazards model[1] which associates time-to-event outcomes with a set of predictor variables through a hazard function of the form:

$$h(t|\mathbf{x}) = h_0(t)\exp(f(\mathbf{x})) \quad (1)$$

where $h(t|\mathbf{x})$ represents the hazard for an individual at time point $t$, with the baseline hazard function $h_0(t)$ describing the time-varying hazard and the partial hazard $\exp(f(\mathbf{x}))$ quantifying the impact of the predictor variables $\mathbf{x}$ on the baseline hazard. In the standard formulation of the Cox proportional hazards model, the partial hazard $\exp(f(\mathbf{x}))$ is assumed to be parametrized by a linear combination of the predictor variables $f(\mathbf{x}) = \boldsymbol{\beta}^\top\mathbf{x}$, with $\boldsymbol{\beta}$ giving the weights for the individual variables.

In many applications, understanding how variables may interact to jointly impact the hazard rate can provide additional value beyond their independent linear effects. However, directly fitting all potential pairwise interaction effects in a multivariable prediction model quickly becomes challenging due to the quadratic increase in the number of interaction terms as a function of the number of input variables. Hence, we propose survivalFM, an extension which adds an approximation of all pairwise interaction effects using a factorized parametrization approach (Fig. 1a-b):

$$f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + \sum_{1 \le i \ne j \le d} \widetilde{\beta}_{i,j} x_i x_j = \boldsymbol{\beta}^\top \mathbf{x} + \sum_{1 \le i \ne j \le d} \langle \mathbf{p}_i, \mathbf{p}_j \rangle x_i x_j \tag{2}$$

where $\langle \cdot , \cdot \rangle$ denotes the inner product and $d$ denotes the number of predictor variables. The first part contains the linear effects of all predictor variables in the same way as in the standard formulation of the Cox proportional hazards model. The second part contains all pairwise interaction effects between the predictor variables $x_i$ and $x_j$. However, instead of directly estimating the interaction effects $\beta_{i,j}$, the effects are approximated through a factorized parameterization using an inner product between two low-rank latent vectors $\widetilde{\beta}_{i,j} = \langle \mathbf{p}_i, \mathbf{p}_j \rangle$. The parameter vectors $\mathbf{p}_i \in \mathbb{R}^k$ and $\mathbf{p}_j \in \mathbb{R}^k$ are the row vectors of a low-rank parameter matrix $\mathbf{P} \in \mathbb{R}^{d \times k}$ (Fig. 1a). Hence, this results in much fewer parameters to estimate, as the rank used in the factorization is typically substantially lower than the total number of predictor variables ($k \ll d$). With this approach, we avoid the statistical and computational problems that would be encountered with direct estimation of all interactions terms in the presence of many predictor variables, while still maintaining interpretability. The idea of using factorized parametrization strategy originates from factorization machines (FMs)[11], originally proposed for regression and classification tasks in the context of recommender systems. For more details of the model and the fitting procedure, see Methods.

## Study population, disease outcomes and data modalities

To evaluate whether survivalFM could improve risk prediction models and provide new insights into the joint effects of risk factors on disease onset, we performed analyses using data from the UK Biobank. This cohort comprises a total of ~500,000 participants from the UK, enrolled in 21 recruitment centers across the country. Upon agreeing to participate, individuals visited the nearest assessment center to provide baseline data, physical measurements, and biological samples.

Of the entire UK Biobank cohort, 93% of participants were recruited in assessment centers located in England and Wales, and 7% in Scotland. Previous studies have revealed differences in health-related characteristics between these regions[12,13]. Given these regional differences and following approaches used in other prediction studies in UK Biobank[12,14,15], we trained our models using data collected from participants enrolled in England and Wales and tested them using data from participants enrolled in Scotland.

The UK Biobank is renowned for its comprehensive phenotyping and molecular profiling, including routine blood biomarkers and advanced ʼomics measurements such as genomics and metabolomics. Baseline characteristics of the study population and a summary of the datasets studied here are summarized in Supplementary Data 1. As disease outcomes, we considered the 10-year incidence of nine example diseases, selected to comprise common diseases and diseases which can benefit from intervention if identified early (Supplementary Data 2 and 3). Disease endpoints were defined based on hospital episode statistics, death registries, self-reported outcomes, and, where available, primary care data (Methods). For lung cancer, we additionally incorporated data from the cancer registry. Prevalent cases were defined as individuals with a recorded first occurrence of a given outcome before the baseline assessment visit and were excluded from analyses for each respective disease endpoint.

To assess the performance across different data modalities, we considered four different prediction scenarios that incorporate an array of predictors ranging from traditional clinical predictors to more advanced omics-based data sources (Fig. 1c, Methods). In the first scenario, we started from a set of standard cardiovascular risk factors included in the ASCVD risk estimator plus[16], widely recognized in various primary prevention scores. Since these factors have been shown to be predictive beyond cardiovascular diseases[17–19], we included them as standard risk factors across all analyzed disease examples. We then added sets of more complex data layers to these standard risk factors (Fig. 1c). In the second scenario, we added a comprehensive set of hematologic and clinical biochemistry measures to the standard risk factors; in the third scenario, we incorporated a wide range of metabolomic biomarkers, recently shown promise as an assay to inform on multidisease risk[17,20]; and finally, we included a set of polygenic risk scores for both disease and quantitative traits[21], which have gained interest for their potential to enhance risk prediction models by providing complementary information to traditional risk factors[22–24].

## survivalFM improves risk prediction across diverse diseases and data modalities

The practical utility of any risk prediction model is determined by its ability to stratify risk and identify high-risk individuals. We evaluated the ability of survivalFM to predict future disease risk and benefit from the comprehensive interaction terms by comparing its performance to standard linear Cox proportional hazards regression (Fig. 1b), employing L2 (Ridge) regularization in both methods to control model complexity and prevent overfitting (Methods). Regularization parameters were optimized via 10-fold cross-validation within the training set (England and Wales participants), selecting the values that maximized the concordance index (C-index) across validation folds. The final obtained models were then tested on the participants enrolled in Scotland.

By modeling the comprehensive interactions present in the underlying data, survivalFM improved the discriminatorion performance across a majority of the studied examples, as measured by concordance index (C-index; Fig. 2). Specifically, statistically significant improvements were noted in 11 of the 36 evaluated scenarios (30.6%), with a mean improvement in C-index (ΔC-index) of 0.0054. Minor improvements were noted in another 23 of the 36 of scenarios (63.9%), with a mean ΔC-index of 0.0014. Importantly, none of the studied examples demonstrated a statistically significant decrease in performance with survivalFM, highlighting its robustness. Absolute values for the C-indices are detailed in Supplementary Data 4, demonstrating good discriminative performance across all models with C-indices ranging from 0.68 to 0.92.

We further evaluated performance of the models using Royston's $R^2$[25] (Fig. 2), which extends the concept of explained variation to survival outcomes, providing a measure of overall model fit. In terms of $R^2$, statistically significant increases in the proportion of explained variation were observed in 15 of the 36 examples (41.67%), with a mean $R^2$ improvement ($\Delta R^2$) of 1.62 percentage points. Minor improvements were observed in 17 of the 36 examples (47.2%), with a mean improvement of 0.53 percentage points. None of the studied examples demonstrated a statistically significant reduction in the explained variation with survivalFM. Absolute $R^2$ values are detailed in Supplementary Data 4, with the proportion of explained variation across the examples ranging from 24.9% to 95.0%.

Given that even modest improvements in discrimination performance at the population level can substantially affect individual risk predictions, we also evaluated model performance using continuous net reclassification improvement (NRI), which has been shown to provide complementary information on risk model performance[26,27]. The continuous NRI quantifies the extent to which the model appropriately increases the predicted probabilities for subjects who experience events and decreases them for those who do not. This metric is particularly useful in the absence of established clinical thresholds for high-risk groups, as it quantifies the improvement in risk prediction without relying on predefined risk cutoffs and thus facilitates comparisons across different diseases.

In terms of continuous NRI, survivalFM yielded significantly improved resclassification in 34 of 36 (94.4%) of the studied examples,

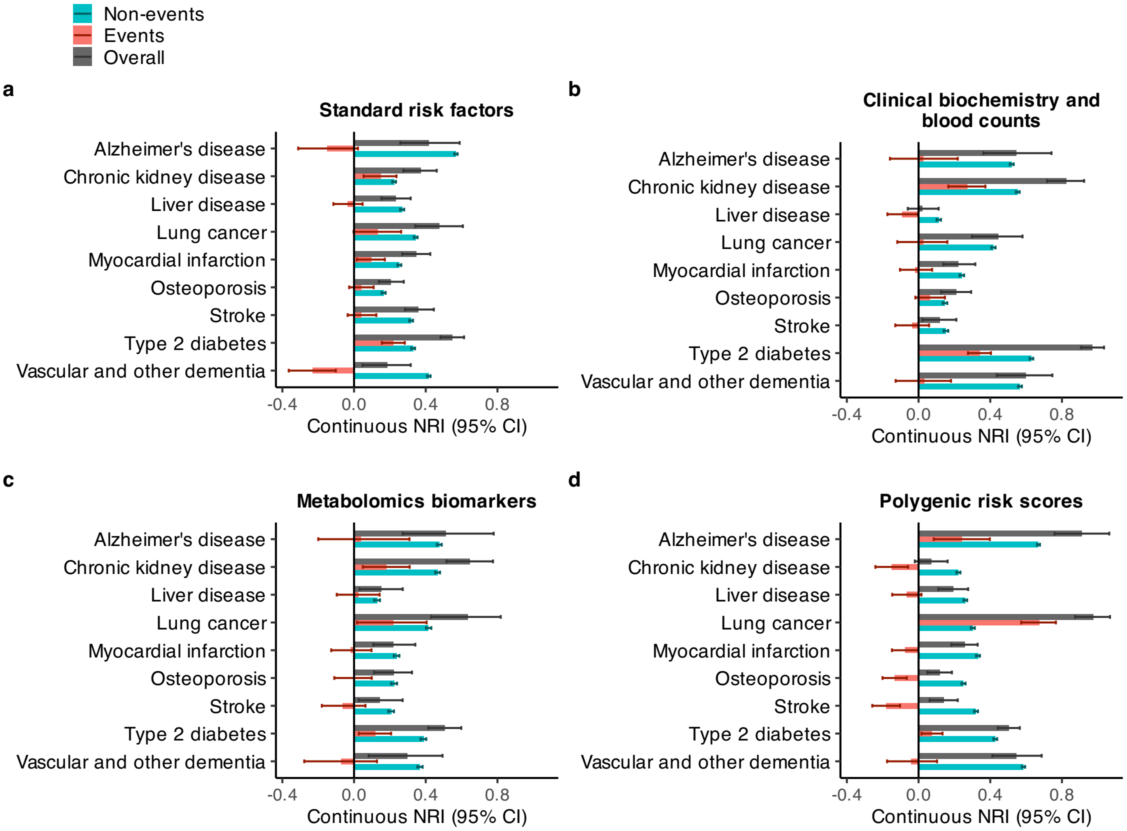

**Fig. 3 | Event- and non-event-specific reclassification improvements.** Continuous net reclassification improvements (NRI) (in dark gray), separated by events (red) and non-events (blue). Results are shown for nine disease examples (y-axis) across four data modalities: **a** standard risk factors (included in all models), **b** clinical biochemistry and blood counts, **c** metabolomics biomarkers, and **d** polygenic risk scores, computed for individuals in the Scotland test set. Bars represent the observed value of the model performance metric; horizontal error bars denote 95% confidence intervals (CIs), estimated with bootstrapping over 1000 resamples. Sample sizes and event counts are provided in Supplementary Data 3. Source data are provided as a Source Data file.

with a mean continuous NRI of 0.41. Minor improvements were noted in the remaining 2 of the 36 examples, with a mean continuous NRI of 0.046. Therefore, despite the relatively modest improvement magnitudes in the C-indices, the continuous NRI indicated notable positive changes in individual risk predictions. For instance, type 2 diabetes modeled using clinical biochemistry and blood counts data demonstrated the highest continuous net reclassification improvement of 0.97 (95% CI 0.91–1.03), corresponding to 34% (95% CI 28–40%) of events and 63% (95% CI 62–64%) of non-events having improved risk estimates (Fig. 3).

To further elucidate the components of the net reclassification improvements, we examined the continuous NRI separately for individuals who experienced events and those who did not (Fig. 3). While the overall continuous NRI was positive across all studied examples, the relative contributions of event- and non-event-specific reclassification varied across diseases and data modalities. Among individuals who experienced events, statistically significant improvements in $NRI_{events}$ were observed in 11 out of 36 examples (30.6%), with a mean improvement of 0.235. Minor improvements in $NRI_{events}$ were noted in 9 of the 36 examples (25.0%), with a mean improvement of 0.047. Conversely, statistically significant decreases in $NRI_{events}$ were observed in 6 out of 36 examples (16.7%), with a mean reduction of 0.143. However, even in these cases, the overall NRI remained positive due to substantial improvements in $NRI_{non\text{-}events}$. Among individuals who did not experience events, the $NRI_{non\text{-}events}$ demonstrated statistically significant improvements across all studied examples. Furthermore, in scenarios where $NRI_{events}$ was negative, the decrease was smaller in magnitude compared to the corresponding gains in $NRI_{non\text{-}events}$, resulting in a positive overall NRI.

To evaluate whether the presence of genetically related individuals between the training and test sets influenced predictive performance, we performed an additional sensitivity analysis. Specifically, we excluded from the Scotland test set all individuals who were genetically related to any member of the England and Wales training set. We defined relatedness as third-degree or closer relatives, using a kinship coefficient threshold of ≥0.0442[28]. Under this more stringent exclusion criterion, the predictive performance of survivalFM remained stable, with only minor numerical fluctuations and no meaningful reduction in the performance metrics (Supplementary Figs. 5, 6). These findings indicate that observed predictive performance is not influenced by the inclusion of related individuals in the test set.

Overall, the models demonstrated good calibration in the Scotland test set, with the exception of chronic kidney disease (Supplementary Figs. 1–4). In this case, both standard Cox regression and survivalFM models trained on the England and Wales training set systematically overestimated disease risk in the Scotland test set across all input types, likely due to differences in chronic kidney disease incidence rates between Scotland and England and Wales in the UK Biobank (Supplementary Data 3).

In addition to our primary analyses, we conducted a supplementary analysis combining all the input data types (standard risk factors, clinical biochemistry and blood counts, metabolomics biomarkers, and polygenic risk scores). This analysis was restricted to individuals in the UK Biobank who had measurements available for all these data types; baseline characteristics, sample sizes, and event counts for this combined analysis are detailed in Supplementary Data 5, 6. In this combined model, the performance gains from modeling comprehensive interactions remained largely consistent but were slightly attenuated (Supplementary Fig. 7), likely due to the increased number of predictor variables, some of which may already capture associations that overlap with the interaction effects. Specifically, in terms of discrimination, minor but statistically non-significant improvements in C-index were observed in 7 out of 9 examples (77.8%), with a mean ΔC-index of 0.0016. For Royston's $R^2$, a statistically significant increase was noted in one example (chronic kidney disease, $\Delta R^2 = 1.16$ percentage points), while minor improvements were present in 4 out of 9 examples (44.4%), with a mean $\Delta R^2$ of 0.23 percentage points. Despite the modest improvements in discrimination and explained variation, continuous NRI demonstrated more pronounced effects: statistically significant improvements were observed in 6 out of 9 examples (66.7%), with a mean NRI of 0.652, while minor improvements were noted in 2 out of 9 examples (22.2%), with a mean NRI = 0.0472.

These findings suggest that interaction terms carry additional predictive information across various disease and data modalities and survivalFM can model this residual contribution. While the extent of improvement varied depending on the specific disease and dataset under study, improvements were consistently observed across multiple disease areas and data types.

### Disease-specific interaction profiles

A key advantage of survivalFM is that despite introducing a more complex layer of non-linearity through the interaction terms, it still maintains interpretability and transparency of how the model predictions are made. To compare the interactions identified by survivalFM with those detected using a conventional approach of explicitly enumerating them in a standard Cox regression model, we performed additional analyses by fitting Cox models with all possible pairwise interactions explicitly included (Methods). Due to computational constraints, this analysis was limited to the standard risk factor dataset with the fewest input predictors.

Despite differences in model parameterization and optimization, both approaches for modeling interactions yielded similar improvements in predictive performance (Supplementary Fig. 8). Furthermore, a comparison of the estimated coefficients (for both main and interaction effects) between survivalFM and Cox regression showed strong concordance (Supplementary Fig. 9). Model predictions were nearly identical, with correlation coefficients exceeding 0.99 in all cases (Supplementary Fig. 10). These findings indicate that, despite methodological differences, both approaches produce highly comparable results. Practically, this suggests that the interaction coefficients in survivalFM ($\beta_{ij} = \langle \mathbf{p}_i, \mathbf{p}_j \rangle$) can be interpreted similarly to those in a standard Cox model ($\beta_{ij}$). Thus, survivalFM maintains the interpretability of interaction effects while offering a more compact representation that mitigates the computational burden associated with estimating numerous interactions simultaneously.

Analysis of the estimated interaction effects from survivalFM across the studied disease outcomes and different input datasets revealed that in many cases there was a diverse interaction landscape contributing to these predictions, demonstrating that the observed performance gains are likely to stem from the cumulative benefit of many small interaction effects rather than a few prominent ones (examples shown in Supplementary Figs. 11–13). Here, we will highlight a few examples with some of the most notable performance gains.

Including interaction terms significantly improved predictive performance across various diseases, with liver diseases modeled using standard risk factors or metabolomic biomarkers being among those showing the highest gains in C-index and $R^2$ (Fig. 2). In the liver disease model based on standard risk factors, prominent interactions emerged among different cholesterol measures, cholesterol-lowering medication, age, body mass index, and sex (Supplementary Fig. 11). These results suggest that the joint effects of these risk factors further explain the risk of chronic liver disease outcomes beyond their additive linear effects. Additionally, smoking status was highly weighted both individually and in the interactions, aligning with the earlier research suggesting that smoking may exacerbate the influence of the other risk factors in the development of chronic liver diseases[29]. In the liver disease model based on metabolomics (Supplementary Fig. 12), various measures of lipid subclasses were highly weighted as individual predictors, while interactions especially among various amino acids were prominent, aligning with previous research on altered amino acid metabolism in chronic liver disease[30,31]. Acetate exhibited a notably strong interaction profile, consistent with its established role in alcohol metabolism and lipid accumulation in the liver[32,33].

A contrasting example was type 2 diabetes modeled using clinical biochemistry and blood counts data, which obtained the highest observed continuous NRI. Unlike the other examples, analysis of the model coefficients revealed that the model weights were predominantly concentrated around glycated hemoglobin (HbA1c) and its interactions across the other variables (Supplementary Fig. 13). The highest interaction weight was attributed to the interaction between HbA1c and glucose, which was negatively weighted despite their positive individual effects. This likely reflects the fact that the simultaneous elevation of both HbA1c and glucose does not increase risk additively but rather relates to them being correlated measures of blood glucose regulation and overall glycemic control. Additionally, the model highlighted positively weighted interactions of HbA1c with age, white ethnicity, and urate levels, indicating these factors together might amplify the risk. In contrast, interactions between HbA1c and reticulocyte count and body mass index were negatively weighted.

### survivalFM benefits from large training data sizes

To understand the impact of training data size on model performance and the ability of survivalFM to leverage interaction terms, we conducted analyses with models trained on varying-sized subsets of the training data. Throughout these analyses, the Scotland test set remained fixed, allowing us to analyze how changes only in the number of training individuals influence model performance. Figure 4 shows the discrimination performance of survivalFM as a function of the number of training individuals for the input dataset involving standard risk factors (results for the other predictor sets are shown in Supplementary Figs. 14–16). For the standard risk factor input dataset (Fig. 4), which contains the fewest predictors and therefore permits feasible inclusion of all pairwise interaction terms, we also include a Cox model with interaction terms as a reference. This provides additional context by serving as a benchmark for performance when interactions are explicitly modeled using a standard approach. These results demonstrate a clear dependency on large sample sizes to uncover predictive interaction terms, with survivalFM generally requiring at least 50,000 individuals in training to outperform standard Cox regression. The discriminatory performance of survivalFM shows a positive trend and increasing gap to standard Cox regression with increasing sample sizes, although the gains often begin to plateau at the upper end of the sample size range.

### survivalFM improves prediction performance in a clinical cardiovascular risk prediction scenario

To explore whether comprehensive interaction modeling via survivalFM could also refine well-established clinical risk prediction models, we conducted analyses in a clinical CVD risk prediction setting using predictors from the QRISK3 model[5]. QRISK models are Cox proportional hazard models used for predicting the patient's 10-year risk of CVD, recommended by the healthcare guidelines in the UK. The latest version, QRISK3 from 2017[5], incorporates a variety of risk factors and comorbidities, along with a set of their interaction terms with age.

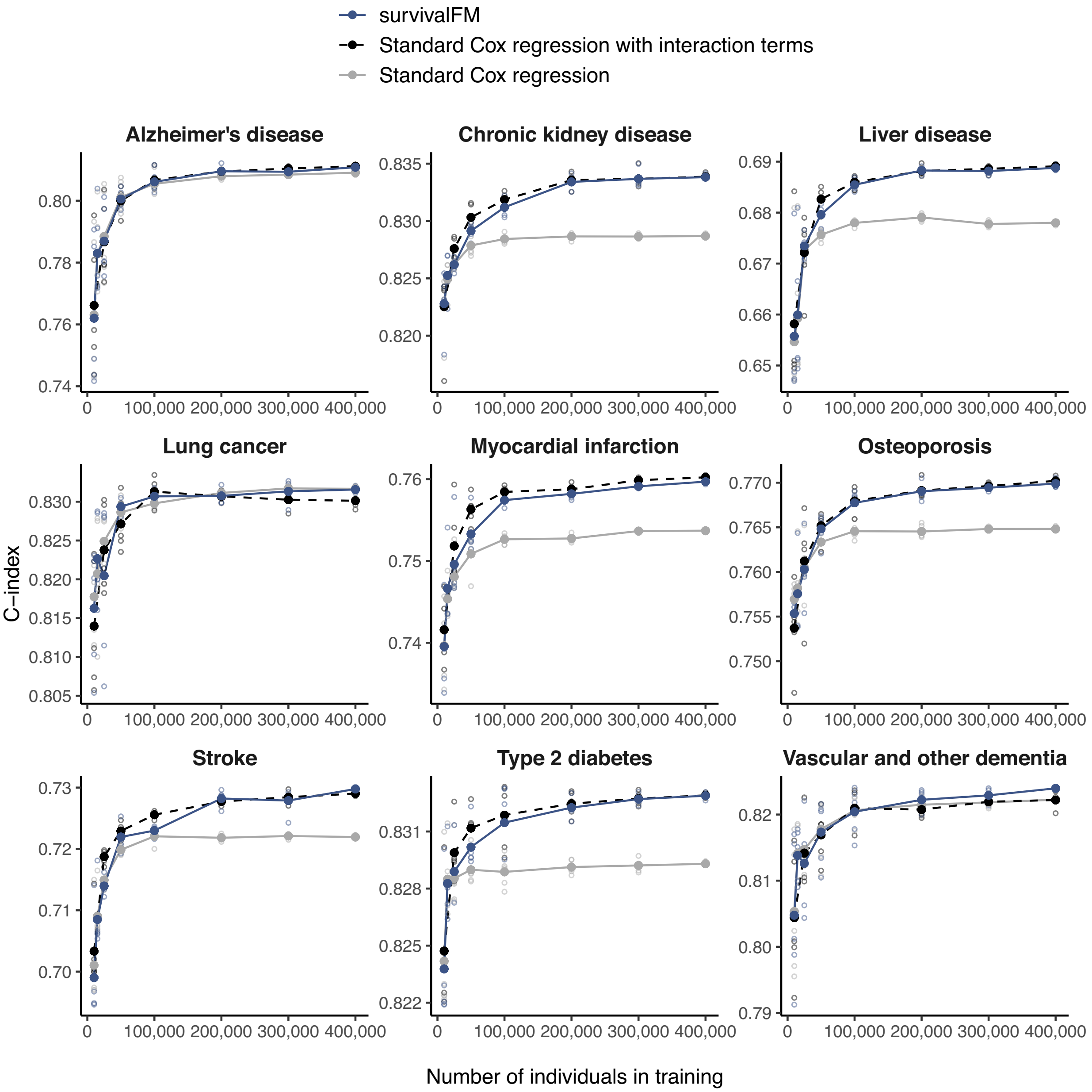

**Fig. 4 | Comprehensive interaction modeling using survivalFM benefits from large training data sizes.** Impact of the size of the training dataset (x-axis) on the discrimination performance in the Scotland test, measured by concordance index (C-index; y-axis), comparing survivalFM (blue, solid line) to standard Cox regression (gray, solid line) and standard Cox regression incorporating all pairwise interactions (black, dashed line). Results are shown for the input dataset consisting of standard risk factors. For each training dataset size, the analysis was repeated five times by randomly sampling with replacement from the full training set. Hollow points indicate the performance from each individual iteration, while filled points (connected by lines) show the average performance. Sample sizes and event counts for each example are provided in Supplementary Data 3. Source data are provided as a Source Data file.

We aimed to determine if comprehensive modeling of the interaction terms among the QRISK3 risk factors using survivalFM could improve the model's ability to predict cardiovascular risk. The endpoint was defined as 10-year incidence of composite CVD, including coronary heart disease, ischemic stroke, and transient ischemic attack, and including both fatal and non-fatal events (Supplementary Data 7, Methods). Following the exclusion criteria from the QRISK3 derivation study, we excluded participants with prior CVD diagnoses and those on a cholesterol-lowering medication at the study entry. The baseline characteristics of the study population in this clinical prediction scenario are detailed in Supplementary Data 8.

To ensure a fair comparison of the models, we retrained the QRISK3 model in the UK Biobank considering the same set of risk factors (Methods). As prior research has shown QRISK3 to systematically overestimate CVD risk in the UK Biobank[34], retraining the model ensures an accurate calibration for this cohort. We evaluated three models of increasing complexity: (1) *Cox regression*: a standard Cox regression model with linear terms only, (2) *Cox regression with age interactions*: a Cox regression model incorporating the linear terms and age interaction terms from QRISK3, and (3) survivalFM: a survivalFM model including the linear terms and all potential factorized pairwise interaction terms.

In terms of discrimination performance measured by C-index, survivalFM showed statistically significant improvements in the Scotland test set (Table 1). Specifically, it improved the discrimination performance by ΔC-index of 0.0019 (95% CI 0.0002–0.0038) over the

**Table 1 | Predictive performance of survivalFM in a practical clinical cardiovascular risk prediction scenario using predictors from QRISK3**

| Performance metric | Cox regression[a] | Cox regression with age interactions[b] | survivalFM[c] |
|---|---|---|---|
| C-index | 0.7476 (0.7369, 0.7593) | 0.7476 (0.7369, 0.7588) | 0.7495 (0.7386, 0.7605) |
| $R^2$ | 42.89% (39.58%, 46.46%) | 43.32% (39.98%, 46.93%) | 44.24% (40.86%, 47.69%) |
| *Relative performance* | | | |
| ΔC-index | | 0.0000 (−0.0014, 0.0015) | 0.0019 (0.0002, 0.0038) |
| $\Delta R^2$ | | 0.43% (−0.04%, 0.90%) | 1.35% (0.57%, 2.11%) |
| Categorical NRI | | | |
| Overall | | 0.0038 (−0.0037, 0.0110) | 0.0168 (0.0061, 0.0279) |
| Events | | 0.0115 (0.0042, 0.0189) | 0.0340 (0.0231, 0.0445) |
| Non-events | | −0.0077(−0.0094, −0.0061) | −0.0172(−0.0194, −0.0152) |

Performance of models trained with QRISK3 predictors for composite cardiovascular disease prediction in the Scotland test set ($N$ = 25,572, 1583 events). Three models are evaluated: [a] standard Cox regression, [b] Cox regression incorporating also age interaction terms from QRISK3, and [c] survivalFM. Performance is reported using absolute measures (C-index and Royston's $R^2$) and relative improvements compared to the standard Cox regression without interactions, including changes in C-index (ΔC-index) and Royston's $R^2$ ($\Delta R^2$), as well as categorical net reclassification improvement (NRI) at a 10% absolute risk threshold, with NRI further separated for events and non-events. 95% confidence intervals are estimated via bootstrapping over 1000 resamples.

standard Cox regression model. In contrast, incorporating the age interaction terms from QRISK3 provided no measurable improvement (ΔC-index = 0.0000, 95% CI −0.0014–0.0014). In terms of explained variation, assessed using Royston's $R^2$, survivalFM increased $R^2$ by 1.35 percentage points (95% CI 0.57–2.11 percentage points) over the standard Cox model, whereas adding the age interactions yielded a smaller, statistically non-significant improvement of 0.43 percentage points (95% CI −0.04–0.90 percentage points). Thus, modeling comprehensive interactions using survivalFM more than four times improved the discrimination performance gains compared to only incorporating the currently included age interactions. Additionally, it led to more than a threefold increase in explained variation compared to incorporating only the pre-specified age interactions.

To further assess how well the models reclassified individuals into appropriate risk categories, we computed categorical net reclassification improvements (NRI) at the guideline recommended 10% absolute risk threshold[35]. Incorporating the age interaction terms from QRISK3 resulted in an overall NRI of 0.0038 (95% CI 0.0037–0.0110) compared to the standard Cox model (Table 1). survivalFM showed a greater overall NRI of 0.0168 (95% CI 0.0061–0.0279), again demonstrating further gains beyond the currently included age interaction terms. survivalFM accurately reclassified 3.40% of individuals who experienced an event into the high-risk category, while it inappropriately reclassified a smaller portion of 1.72% of non-events as high-risk (Table 1). These improvements are also visible in the reclassification plots (Fig. 5a, b) showing how the individual predictions change with the inclusion of interaction terms. All models were well calibrated in the Scotland test set (Supplementary Fig. 17a) and exhibited broadly similar distributions across the risk spectrum (Supplementary Fig. 17b).

To further assess model performance, we conducted a Kaplan–Meier analysis, evaluating observed CVD event risk over 10 years in patients stratified by the guideline recommended 10% risk threshold (Fig. 5c, d). As expected, cumulative event rates in these groups were consistent across models, reflecting both the shared 10% threshold and the fact that all models were well-calibrated (Supplementary Fig. 17a). Specifically, event rates were 15% in the high-risk group and 4% in the low-risk group across all models, with no significant differences in Kaplan–Meier curves (log-rank test: $p = 0.88$ for predicted risk ≥10%, $p = 0.63$ for predicted risk <10%). However, models incorporating interaction terms identified larger high-risk groups while maintaining comparable absolute risk thresholds, thereby capturing more events. By 10 years, survivalFM identified 844 events in the high-risk group (predicted risk ≥10%), a 6.7% increase over the 791 events captured by the standard Cox model. Adding QRISK3's age interactions resulted in 809 events in the high-risk group, a smaller 2.3% increase over the standard Cox model.

Analysis of the model coefficients from survivalFM revealed a broad array of interactions contributing to the CVD predictions. The ratio of total cholesterol to HDL cholesterol demonstrated one of the most pronounced interaction profiles among all predictor variables (Fig. 6). This suggests that the effect of the cholesterol ratio on CVD risk is influenced by the presence of other risk factors. For example, the interaction weight for the cholesterol ratio with prevalent atrial fibrillation was negative, despite both factors having positive individual weights. This suggests that these variables capture partly overlapping aspects of cardiovascular risk. Atrial fibrillation is often associated with a broader cardiovascular risk[36], which could already be reflected in the elevated cholesterol ratio. This may thus imply that when both risk factors are present, they do not independently add to the risk. Comparing the estimated effects for the model terms overlapping between survivalFM and the standard Cox regression model with linear and age interaction terms from QRISK3, the shared terms exhibited very similar weights, with correlation of 0.95 between the estimated effects by the two methods (Supplementary Fig. 18). This shows that despite the introduction of complex interactions, the fundamental risk associations remain broadly consistent.

## Discussion

Accurate prediction of disease onset and prognosis is essential to realize preventative medicine. In this study, we have introduced survivalFM, a machine learning method for multivariable time-to-event prediction. The method extends the Cox proportional hazards model by estimating all potential pairwise interaction effects among predictor variables, allowing for more expressive modeling of time-to-event outcomes, such as disease onset. We have shown that estimating these comprehensive interaction effects improves risk prediction and refines individual risk predictions across a range of common diseases, providing more nuanced insights into the interplay among factors underlying disease risk. Since survivalFM generalizes to other use cases, we expect this method to find applications in precision medicine and benefit survival modeling in large studies involving many predictors.

Our results from UK Biobank revealed that survivalFM can identify predictive interaction terms, which are missed when using standard Cox proportional hazards regression. This ability to uncover predictive interaction terms extended across various disease outcomes and data modalities. Importantly, survivalFM consistently matched or surpassed the performance of the standard Cox regression model. This robustness is by design, as survivalFM separates linear effects from interaction effects and, by appropriate tuning of model hyperparametrs, can assign negligible weight to non-contributory interaction effects while emphasizing predictive ones. These findings highlight the

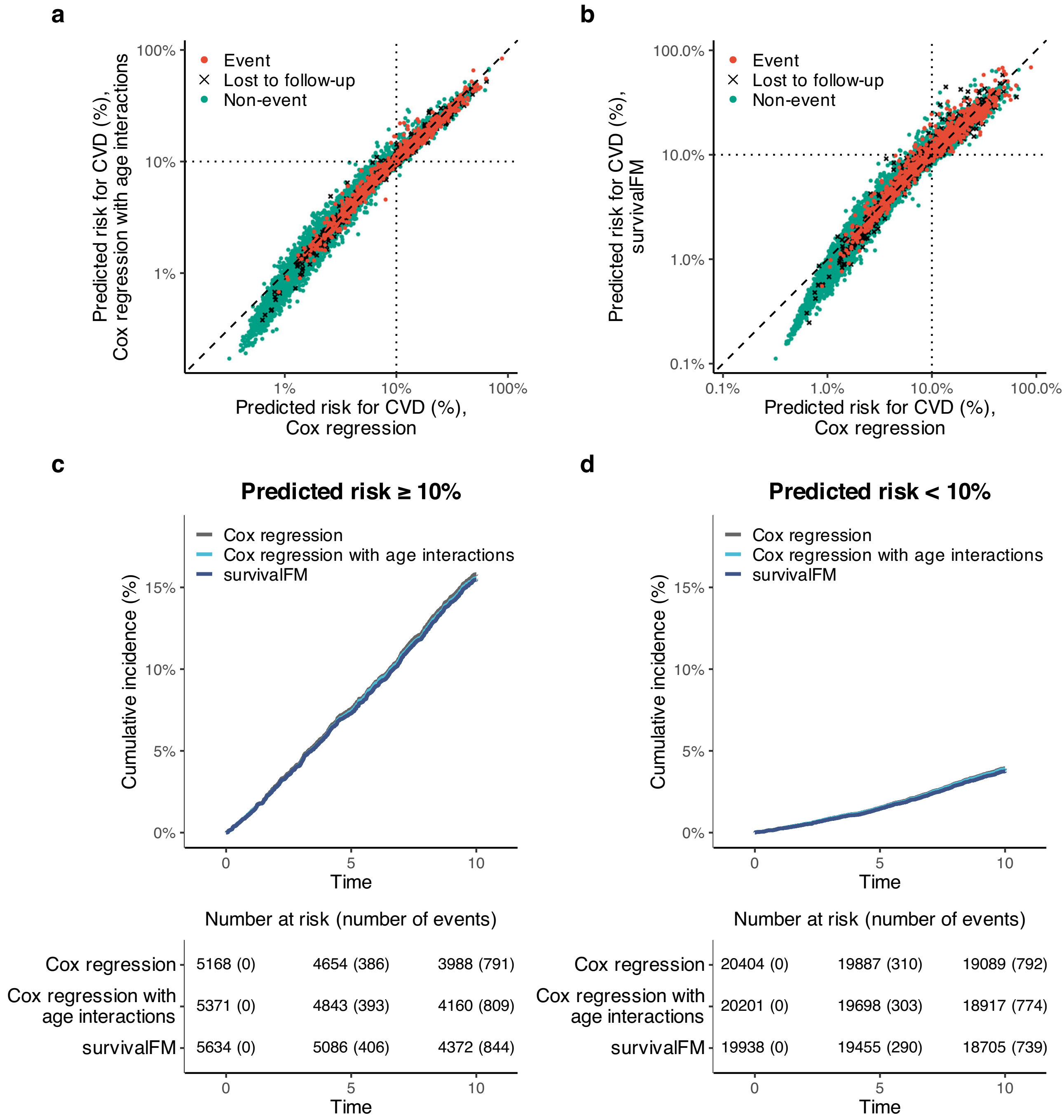

**Fig. 5 | Reclassification and risk stratification in a clinical cardiovascular risk prediction scenario using QRISK3 predictors.** **a** Reclassification plot comparing individual risk predictions when adding QRISK3 age interaction terms to a standard Cox regression model (y-axis) versus a standard Cox model (x-axis). **b** Reclassification plot comparing risk predictions from survivalFM (y-axis) to the standard Cox model (x-axis). Both axes are on a log scale. Black dotted vertical and horizontal lines show the 10% absolute risk threshold for the high risk category. **c** Kaplan–Meier curves for cumulative CVD incidence in the high-risk group (≥10% predicted risk) across three models: (1) Cox regression (dark gray), (2) Cox regression with QRISK3 age interactions (light blue), and (3) survivalFM (blue), with a risk table below. **d** Kaplan–Meier analysis for the low-risk group (<10% predicted risk) using the same models and color coding. Results are based on the Scotland test set (*N* = 25,572, 1583 events). Source data are provided as a Source Data file.

utility of survivalFM in refining risk prediction models across various prediction scenarios, including models derived from traditional clinical predictors and modern omics data types.

While the absolute improvements in population-level metrics such as the C-index and $R^2$ were moderate, this is consistent with findings particularly from genetic risk prediction studies, where interaction effects, such as gene-gene or gene-environment interactions, have been shown to explain relatively little additional variance beyond main effects[37–39]. Nevertheless, including interaction terms yielded notable gains in net reclassification improvement (NRI), highlighting the value of modeling interactions for improving individual-level risk predictions. This supports the broader view that even modest improvements at the population level can yield meaningful benefits in personalized risk prediction[26,27].

Our results further showed that survivalFM can add predictive value in practical clinical risk prediction scenarios, such as in CVD risk

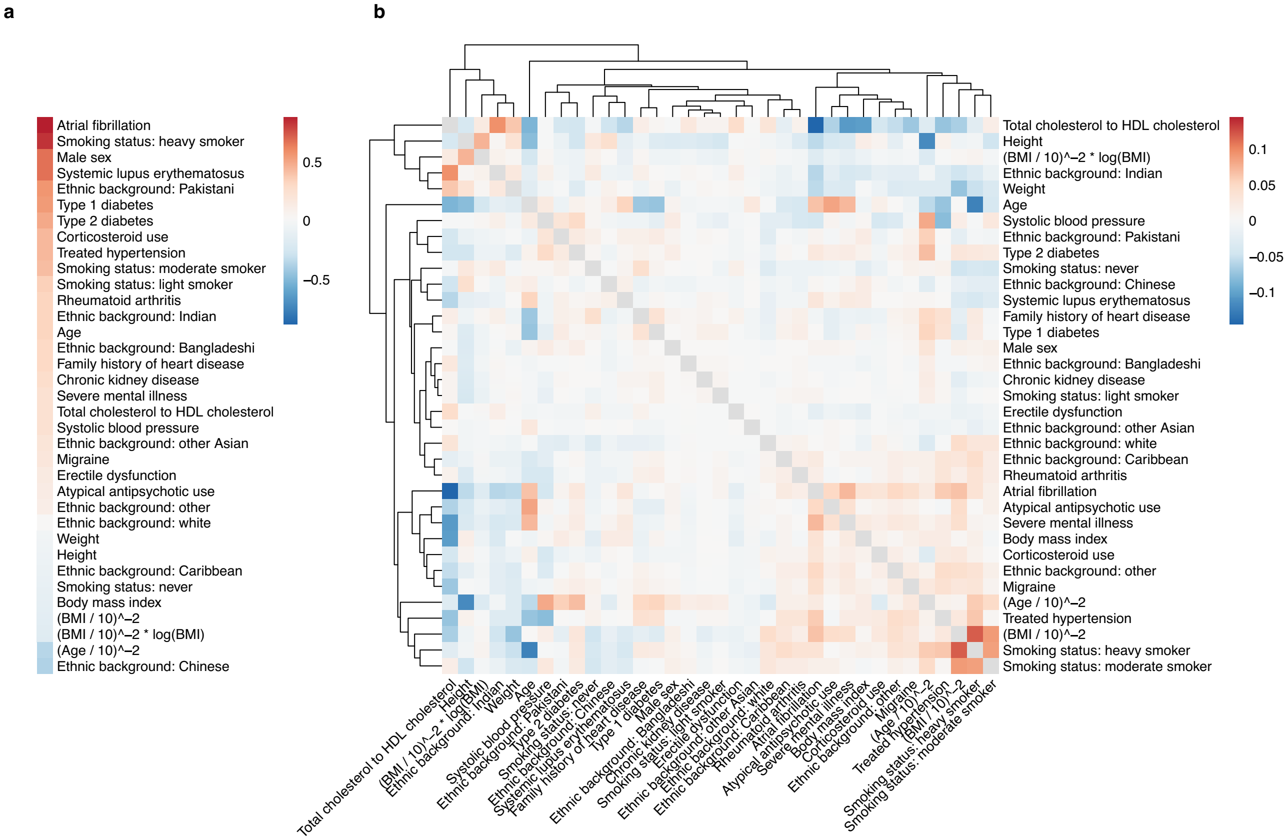

**Fig. 6 | Estimated model coefficients from survivalFM model, trained on the England and Wales training set considering the risk factors from QRISK3 for cardiovascular disease risk prediction.** Estimated coefficients for **a** the linear effects $\boldsymbol{\beta}$ and **b** the interaction effects given by the inner product of the factor vectors $\beta_{i,j} = \langle \mathbf{p}_i, \mathbf{p}_j \rangle$. The dendrogram shows a hierarchical clustering of the interaction profiles, using Euclidean distance as the measure of similarity. Source data are provided as a Source Data file.

prediction using predictors from the established QRISK3 model. CVD remains as the leading cause of mortality worldwide[40], making accurate risk stratification critical for healthcare providers to allocate preventive measures effectively. Applying survivalFM to QRISK3 risk factors improved both discrimination and reclassification at the clinically recommended 10% risk threshold, more than doubling the performance gains obtained from the current model's age-related interaction terms alone. For context, while a recent study[22] reported a net reclassification improvement of 0.013 by adding a polygenic risk score to a CVD prediction model in a similar scenario involving QRISK3 risk factors, survivalFM achieved a comparable improvement by optimizing the use existing clinical variables.

A key strength of survivalFM is that despite introducing nonlinearity through the comprehensive interaction terms, it maintains interpretability by providing the estimated effects for both the individual terms and the approximated interactions. This is unlike many other advanced machine learning techniques, which often lack transparency. Another advantage of survivalFM is a straightforward training process, which only involves optimizing the regularization parameters and setting the rank for factorizing the interaction parameters. We anticipate the accompanying R package will facilitate rapid adoption of the method in other prediction studies.

Interpretation of the trained models suggested that in many cases numerous small interaction effects collectively enhanced the prediction accuracy, highlighting the importance of modeling the entire interaction landscape. However, we also found that learning these interaction effects typically requires a substantial sample size. This can limit the method's applicability in smaller cohorts and settings with lower sample sizes. Therefore, future studies in adequately powered cohorts are needed to assess the consistency of the identified interactions and gains in prediction accuracy across diverse populations. Whilst large sample size is needed, many biobank initiatives are emerging with clinical and omics data at scale. Our results indicate such initiatives could be used as a base for discovering and replicating comprehensive risk factor interactions that are missed by conventional statistical methods.

survivalFM currently employs L2 regularization, which effectively controls the magnitude of model parameters but does not enforce sparsity. Future research could explore alternative or additional forms of regularization that promote sparsity in the interactions. However, achieving sparsity in the interactions is non-trivial, as it would necessitate enforcing orthogonality among certain factor vectors (to achieve zero inner product), but retaining some of them non-zero. This requires special regularization techniques to be developed[41] and we see this an interesting avenue for future work.

The generalizable nature of survivalFM makes it applicable also to other data modalities than those highlighted in this paper. For instance, comprehensive modeling of interactions across omics data modalities could provide valuable insights into the molecular interplay behind disease risk. Another use case could be studies of protein interaction patterns in relation to disease onset. Recent studies in UK Biobank have demonstrated the strong promise proteomics data in predicting various diseases[42–44]. Given that proteomics data in UK Biobank comprises around 3000 measured proteins, the number of potential interactions is in millions. While the current sample size of 50,000 with proteomics data in UK Biobank is at the lower limit for comprehensive interaction modeling, with a sufficient sample size, survivalFM has the potential to uncover protein interactions predictive of disease onset and potentially provide further insights for personalized treatment strategies.

In conclusion, survivalFM provides an advancement in survival analysis, enhancing disease risk prediction by effectively incorporating comprehensive interaction terms. Our findings provide a foundation for future research and translation of risk prediction models, emphasizing the importance of interaction effects in understanding disease development and refining risk prediction models.

# Methods

## Study population

This study used data from the UK Biobank. The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee and all participants provided written informed consent. Further details of the study protocol and data collection are available online (https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf) and in the literature[45].

The UK Biobank is a comprehensive prospective cohort study serving as a major globally available health research resource. It includes data from approximately half a million participants aged 37–73, representing a sample from the general UK population. The participants were recruited through 22 assessment centers throughout England, Wales, and Scotland between 2006 and 2010. The follow-up is still ongoing. In this study, the data was accessed under UK Biobank project ID 147811. Sex was not considered in the study design.

## survivalFM: Extending the proportional hazards model with factorized interaction terms

**Survival data**. Throughout this paper, we assume right-censored survival data. This means that the outcome consists of two components: the event of interest (here, the onset of a disease) and the duration from the beginning of the study until either to the occurrence of the event, patient loss to follow-up, or end of the duration of follow-up (i.e., right censoring). The survival dataset $\mathcal{D}$ consists of tuples $\mathcal{D} = \{(\mathbf{x}_i, t_i, \delta_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ represents a vector of predictor variables for the individual $i$, $t_i$ marks the observed time to the event of interest or to the point of censoring, and $\delta_i$ is an indicator function which denotes whether $t_i$ corresponds to an actual event occurrence ($\delta_i = 1$) or censored observation ($\delta_i = 0$).

**Model formulation**. We base survivalFM on the widely utilized proportional hazards model[1] which associates time-to-event outcomes with a set of predictor variables using a hazard function of the form:

$$h(t|\mathbf{x}) = h_0(t)\exp(f(\mathbf{x})) \tag{3}$$

where $h_0(t)$ is a shared baseline hazard function that varies over time, and $\exp(f(\mathbf{x}))$ is a partial hazard that describes the effects of the predictor variables on the baseline hazard. In the standard formulation of the Cox proportional hazards model, the partial hazard $\exp(f(\mathbf{x}))$ is assumed to be parametrized by a linear combination of the variables of the individual, $f(\mathbf{x}) = \boldsymbol{\beta}^{\top}\mathbf{x}$, with $\boldsymbol{\beta}$ representing the coefficients or parameters of the model assigning weights to the individual variables $\mathbf{x}_i$.

In this study, in addition to the individual effects of the variables, we propose to add an approximation of all pairwise interaction terms using a factorized parametrization of the coefficients, following the approach originally introduced along with factorization machines[11] in the context of recommender systems:

$$f(\mathbf{x}) = \boldsymbol{\beta}^{\top}\mathbf{x} + \sum_{1 \le i \ne j \le d} \widetilde{\beta}_{i,j} x_i x_j = \boldsymbol{\beta}^{\top}\mathbf{x} + \sum_{1 \le i \ne j \le d} \langle \mathbf{p}_i, \mathbf{p}_j \rangle x_i x_j \tag{4}$$

where $\langle \cdot , \cdot \rangle$ denotes the inner product. The first part contains the linear effects of the predictor variables in the same way as in the standard formulation of the Cox proportional hazards model. The second part contains all pairwise interactions between the predictor variables $x_i$ and $x_j$. However, instead of directly estimating the interaction weights $\beta_{i,j}$, the factorized parametrization approximates the coefficients using an inner product between two latent vectors $\widetilde{\beta}_{i,j} = \langle \mathbf{p}_i, \mathbf{p}_j \rangle$. The low-rank factor vectors $\mathbf{p}_i \in \mathbb{R}^k$ and $\mathbf{p}_j \in \mathbb{R}^k$ are collected into a parameter matrix $\mathbf{P} \in \mathbb{R}^{d \times k}$ (Fig. 1a). Rank $k$ is a hyperparameter that determines the dimensionality of the latent factor vectors, and usually the optimal rank of the factorization is substantially lower than the number of input predictors ($k \ll d$).

The rank parameter ($k$) is a central hyperparameter in survivalFM, as it defines the dimensionality of the latent space in which interactions between predictors are captured. A higher $k$ allows for greater expressiveness, while a lower k enforces a more compact representation. In this study, to keep the computational complexity manageable, the rank parameter was fixed to 10. To illustrate the impact of the selection of rank hyperparameter $k$, we demonstrated that setting k=0 (i.e., modeling only individual effects without interactions) reduces survivalFM to a standard Cox regression model, yielding comparable performance (Supplementary Figs. 19–22). As $k$ increases, performance initially improves, particularly between $k = 0$ and $k = 5$, but typically plateaus beyond $k = 5$ or $k = 10$. In most cases, no clear signs of overfitting are observed at higher values of $k$, with the exception of liver disease and type 2 diabetes models incorporating polygenic risk scores (Supplementary Fig. 22). However, higher $k$ values increase the computational complexity, reinforcing the need for a conservative choice that balances predictive accuracy and efficiency (Supplementary Figs. 19–22). While the rank parameter can be optimized via cross-validation, this approach can become computationally prohibitive in large datasets involving many predictor variables.

As survivalFM bases on the assumption that the effects of pairwise feature interactions can be represented by low-rank factor vectors, it allows it to estimate interaction effects even under highly sparse data. Hence, the model can learn interaction effects ( $\beta_{i,\,j}$) without requiring direct co-occurence of predictor variables $x_i$ and $x_j$ in the training data. Instead, the factor vectors $\mathbf{p}_i$ and $\mathbf{p}_j$ can be learned through interactions with other predictor variables and their dot product still gives $\beta_{i,\,j}$. This property is particularly useful, for instance, when working with categorical variables that have limited joint occurrences in the training data. However, the model's ability to generalize interactions beyond observed data also means that it may assign nonzero interaction effects to feature pairs that never actually co-occur in real-world data (e.g., "moderate smoker" and "heavy smoker" in Fig. 6). This is a natural consequence of low-rank factorization, where latent factors capture generalized interaction patterns. Importantly, while the model does not inherently enforce real-world constraints such as mutual exclusivity among certain categorical variables, these cases do not affect model predictions as such feature combinations never appear in real input data.

**Parameter estimation**. Following the standard Cox proportional hazards regression, we estimate the model parameters $\theta$ using a partial likelihood function $L(\theta|\mathcal{D})$. For each individual who experiences an event at time $t$, their likelihood contribution is the ratio of the hazard of that individual to the cumulative hazard of all other individuals at risk at the same time point, multiplied across all individuals with event occurrence. Formally, this can be expressed as follows:

$$L(\theta|\mathcal{D}) = \prod_{i:\delta_i=1} \frac{h_0(t)\exp(f(\mathbf{x}_i))}{\sum_{j\in R(t_i)} h_0(t)\exp(f(\mathbf{x}_j))} = \prod_{i:\delta_i=1} \frac{\exp(f(\mathbf{x}_i))}{\sum_{j\in R(t_i)} \exp(f(\mathbf{x}_j))} \tag{5}$$

where $\mathbf{x}_i$ denotes the vector of predictor variables for an individual $i$, $t_i$ is the observed event time for individual $i$ and $R(t_i)$ denotes the risk set at time $t_i$. Being in the risk set essentially means that the individual has not had an event yet or that their censoring date has not passed yet.

Here, $f(\mathbf{x})$ corresponds to the log-risk function from Eq. (4) containing the individual effects and all pairwise interaction terms in a factorized form. As the baseline hazard function $h_0(t)$ is assumed to be shared across all individuals, it cancels out when calculating the partial likelihood, hence eliminating the need for its specification, a key feature of the Cox proportional hazards model rendering it semi-parametric.

To find the optimal parameters $\theta = \{\boldsymbol{\beta}, \mathbf{P}\}$, instead of maximizing the partial likelihood, one can equivalently minimize the negative log-likelihood to obtain a more convenient formulation. Taking the logarithm of the partial likelihood function yields a log-likelihood function of the form:

$$l(\theta|\mathcal{D}) = \log\left(\prod_{i:\delta_i=1} \frac{\exp(f(\mathbf{x}_i))}{\sum_{j\in R(t_i)} \exp(f(\mathbf{x}_j))}\right) = \sum_{i:\delta_i=1}\left(f(\mathbf{x}_i) - \log\left(\sum_{j\in R(t_i)} \exp(f(\mathbf{x}_j))\right)\right) \tag{6}$$

To overcome overfitting in scenarios involving many predictor variables, one can include regularization terms. Here, we consider L2 regularization (Ridge). Hence, the regularized learning problem is given by:

$$\arg\min_{\boldsymbol{\beta},\mathbf{P}} -\frac{2}{n} l(\theta|\mathcal{D}) + \lambda_1||\boldsymbol{\beta}||_2^2 + \lambda_2||\mathbf{P}||_2^2 \tag{7}$$

where $\lambda_1$ and $\lambda_2$ are the regularization parameters for the individual effects and the factorized interactions, respectively. Using separate regularization parameters for the individual effects and the interactions allows for individual penalization of these two parts. The log-likelihood is scaled by a factor of $2/n$ for convenience and to follow the definition from the popular *glmnet* R package[46], used for the standard Cox regression comparison in this study.

Gradient of the negative log-likelihood function $l(\theta|\mathcal{D})$ with respect to the model parameters $\theta = \{\boldsymbol{\beta}, \mathbf{P}\}$ is given by:

$$\frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{n:\delta_n=1} \frac{\partial}{\partial\boldsymbol{\theta}}\left(f(\mathbf{x}_i|\theta) - \log\left(\sum_{j\in R(t_i)} \exp(f(\mathbf{x}_j|\theta))\right)\right) \tag{8}$$

$$= \sum_{n:\delta_n=1}\left(\frac{\partial f(\mathbf{x}_i|\theta)}{\partial\boldsymbol{\theta}} - \frac{\sum_{j\in R(t_i)} \frac{\partial f(\mathbf{x}_j|\theta)}{\partial\boldsymbol{\theta}} \exp(f(\mathbf{x}_j|\theta))}{\sum_{j\in R(t_i)} \exp(f(\mathbf{x}_j|\theta))}\right) \tag{9}$$

where

$$\frac{\partial f(\mathbf{x}|\theta)}{\partial\boldsymbol{\theta}} = \begin{cases} x_i & \text{if } \theta \text{ is } \beta_i \\ x_i \sum_{j=1}^{n} p_{j,f} x_j - p_{i,f} x_i^2 & \text{if } \theta \text{ is } p_{i,f} \end{cases} \tag{10}$$

The sum $\sum_{j=1}^{n} p_{j,f} x_j$ is independent of $i$ and thus can be precomputed[11]. In addition, the gradients of the L2 regularization terms are given by $\frac{\partial}{\partial\boldsymbol{\theta}}\lambda||\theta||_2^2 = 2\lambda\theta$.

To solve (7), we use an efficient BFGS (Broyden-Fletcher-Goldfarb-Shanno) quasi-Newton algorithm[47–50], as implemented in the base R stats package[51]. In contrast to the standard Newton-Raphson method, the BFGS algorithm uses an approximation of the Hessian to determine the search direction. Due to the factorization of the interaction parameters, the number of estimated parameters remains moderate even in the presence of many predictor variables, making the computation of the Hessian approximation feasible. Empirical evidence from our analyses indicated that alternative stochastic gradient descent (SGD)-based optimization methods, commonly employed in machine learning, were not as effective here.

## Predictor variable sets

**Standard risk factors.** As standard risk factors, we included predictors from the ASCVD risk estimator plus[16,52,53], which are also commonly featured in other primary prevention tools. These demographic and cardiovascular risk factors have been shown to be predictive of diseases beyond CVD[17–19]. These included age, sex (self-reported), ethnic background, systolic and diastolic blood pressure, total, HDL, and LDL cholesterol, smoking status, prevalent type 2 diabetes (excluded from the analyses related to type 2 diabetes), hypertension, and cholesterol-lowering treatment, further detailed in Supplementary Data 9. This data was extracted from the data collected at the study's initial recruitment visit. Prevalent diabetes status was extracted from primary care records, hospital episode statistics, and self-reported conditions during the initial assessment. These standard risk factors were included in all models trained.

**Clinical biochemistry and blood counts.** A comprehensive set of clinical biochemistry measures were provided by UK Biobank for blood samples taken at the initial recruitment visit and have been previously described in the literature[54,55]. These included hematologic markers (complete blood counts, white blood cell populations and reticulocytes) and a wide range of blood biochemistry measures covering established risk factors, diagnostic biomarkers and other chracterisation of phenotypes, such as measures for renal and liver function. Nucleated blood cell counts were excluded from our analyses due to over 99% of the cohort having these recorded as missing or zero. We also excluded estradiol, rheumatoid factor and lipoprotein (a), due to a large portion of the cohort (>20%) having these recorded as missing or under the limit of detection. The blood sample handling and storage protocol has been previously described in the literature[56]. A complete list of the included variables is provided in Supplementary Data 10.

**Metabolomics biomarkers.** The metabolomics data included 168 lipids and metabolites from a high-throughput NMR metabolomics assay, available for the baseline blood samples from approximately 275,000 individuals in the UK Biobank. The metabolite data covers a wide range of small molecules, such as amino acids, inflammation markers and ketones, as well as lipids, lipoproteins and fatty acids. Percentage ratios calculated from these 168 original measures were excluded from our analyses. Details of the metabolite data have been previously described[20]. A complete list of the included metabolomics biomarkers is included in Supplementary Data 11.

**Polygenic risk scores.** The polygenic risk score data included 53 polygenic risk scores (PRS) released by the UK Biobank and described in ref. 21. These included scores for both disease traits and quantitative traits. In our analyses, we included only the standard PRS set obtained entirely from external genome-wide association study (GWAS) data. As provided in the UK Biobank, the score distributions were already centered at zero across all ancestries using a principal component-based ancestry centering step. A complete list of the included variables is provided in Supplementary Data 12.

**QRISK3 risk factors.** We matched the risk factors from the QRISK3 model with the corresponding variables available in the UK Biobank. These variables were gathered during the baseline assessment visit and included cholesterol levels measured from blood samples and prevalent disease diagnoses obtained from linked hospital records, primary care data, and self-reported conditions. In instances where an exact match for a QRISK3 model risk factor was unavailable in the UK Biobank, the closest equivalent field was utilized. A complete list of the predictors and their corresponding UK Biobank fields is provided in Supplementary Data 13.

### Disease endpoint definitions

For the highlighted examples across different data modalities and 10-year incidence of nine different disease outcomes, each of the outcomes was defined by the earliest occurrence in primary care (available for a subset of individuals), hospital episode statistics or death records, using the first occurrences data field from UK Biobank (category 1712). For lung cancer, we additionally included data from the cancer registry. The endpoints were defined based on 3-character ICD-10 codes: Alzheimer's disease (F00, G30), chronic kidney disease (N18), liver disease (K70-K77), lung cancer (C33-C34), myocardial infarction (I21-I22), osteoporosis (M80-M81), stroke (I60-I61, I63-I64), type 2 diabetes (E11, E14), and vascular and other dementia (F01-F03) (Supplementary Data 2). Participants with a recorded diagnosis of the disease before the baseline assessment visit were excluded from the analysis of the corresponding disease endpoint.

The analysis of QRISK3 predictors focused on a 10-year composite CVD outcome, defined according to the original QRISK3 derivation study[5], including coronary heart disease (I20-I25), ischemic stroke (I63-I64), and transient ischemic attack (G45) (Supplementary Data 5). We used the earliest recorded date of cardiovascular disease on any of the three data sources (primary care, hospital episode statistics and death records) as the outcome date, using the first occurrences data field from UK Biobank (category 1712). Participants with a prior CVD diagnosis and those on a cholesterol lowering medication at the start of the study were excluded from the analyses, following the exclusion criteria from the original QRISK derivation study[5].

### Data partitions and preprocessing

The models were trained using data from participants enrolled in England and Wales (93% of the entire UK Biobank cohort) and subsequently tested using data from participants enrolled in Scotland (7% of the cohort). Within the training set, model hyperparameter tuning was performed using 10-fold cross-validation.

For data preprocessing, log-normally distributed continuous variables (concerning clinical biochemistry markers, blood counts and metabolomics biomarkers) were log1p-transformed (i.e., taking the logarithm of the given value plus one). Outliers exceeding 4 standard deviations from the mean were winsorized. Continuous variables were scaled to zero mean and unit variance and categorical variables were one-hot encoded. The means and standard deviations used for scaling were calculated from the training set and subsequently applied to the Scotland test set. To maximize sample size for model training, missing values were imputed within the training set using k-nearest neighbors (kNN) imputation ($k = 10$). To ensure no data leakage, imputation was exclusively performed within the training set. Hence, for the Scotland test set, only individuals with complete data were included.

### Model hyperparameter tuning

In both the standard Cox regression and our proposed survivalFM method, we employed L2 (ridge) regularization to control model complexity and prevent overfitting. This requires tuning the regularization parameter $\lambda$. For survivalFM, we allowed differing regularization strengths for the linear ($\lambda_1$) and the interaction part ($\lambda_2$), to separately control the influence of main effects and interaction effects. Similarly, in standard Cox regression models with all pairwise interactions explicitly enumerated, we adopted a similar strategy of allowing differing regularization strengths for the main and interaction effects. All regularization parameters were optimized using 10-fold cross-validation by considering a series of equally spaced values on a logarithmic scale between $\{1, 1^{-4}\}$, choosing those that maximized the concordance index (C-index) across the validation folds. In addition, survivalFM requires setting the rank of the factorization ($k$) for the interaction parameters, which was here set to $k = 10$.

### Analysis of model performance

Confidence intervals for all metrics were calculated with 1000 bootstrapping iterations. Statistical inferences about differences were based on the distributions of bootstrapped performance difference metrics by considering performances statistically significantly different when the 95% confidence intervals did not overlap zero. All analyses were performed using R version 4.3.1[51].

The standard Cox regression models used for the comparisons were trained using the R package *glmnet* (version 4.1.8)[46,57]. Concordance indices (Harrel's C-index) were computed using the R package *survival* (version 3.8.3)[58], Royston's $R^2$ using the R package *survMisc* (version 0.5.6)[59] and net reclassification improvements using the R package *nricens* (version 1.6)[60].

### Statistics and reproducibility

This research was conducted as an observational cohort study using data obtained from the UK Biobank. No statistical method was used to predetermine sample size. Sample inclusion and exclusion criteria are detailed in the "Disease endpoint definitions" and "Data partitions and preprocessing" sections. The study design did not involve any intervention, and therefore, traditional experimental procedures such as randomization and blinding are not applicable.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Individual-level UK Biobank data are available under restricted access due to participant confidentiality and data privacy regulations. Researchers can apply for access through the UK Biobank Access Management System: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access. Access is granted to bona fide researchers for health-related research, following a formal application and approval process. Source data are provided with this paper.

## Code availability

The method developed in this study has been made available as an R package and can be installed from: https://github.com/aalto-ics-kepaco/survivalfm[61]. The scripts used for conducting the analyses presented in this manuscript are available on GitHub: https://github.com/aalto-ics-kepaco/survivalfm-analysis[62].

## References

1. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc.: Ser. B* **34**, 187–202 (1972).
2. Harrell Jr, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
3. Corraini, P., Olsen, M., Pedersen, L., Dekkers, O. M. & Vandenbroucke, J. P. Effect modification, interaction and mediation: an overview of theoretical insights for clinical investigators. *Clin. Epidemiol.* **9**, 331–338 (2017).
4. Lewington, S. et al. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* **360**, 1903–1913 (2002).
5. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* **357**, j2099 (2017).
6. Hageman, S. et al. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur. Heart J.* **42**, 2439–2454 (2021).

7. Kaptoge, S. et al. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *Lancet Glob. Health* **7**, e1332–e1345 (2019).
8. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
9. Katzman, J. L. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 1–12 (2018).
10. Nagpal, C., Li, X. & Dubrawski, A. Deep survival machines: fully parametric survival regression and representation learning for censored data with competing risks. *IEEE J. Biomed. Health Inform.* **25**, 3163–3175 (2021).
11. Rendle, S. Factorization machines. In *2010 IEEE International conference on data mining*, 995–1000 (IEEE, 2010).
12. Zhang, S. et al. A metabolomic profile of biological aging in 250,341 individuals from the uk biobank. *Nat. Commun.* **15**, 8081 (2024).
13. McCartney, G. et al. Explaining the excess mortality in scotland compared with england: pooling of 18 cohort studies. *J. Epidemiol. Community Health* **69**, 20–27 (2015).
14. Argentieri, M. A. et al. Integrating the environmental and genetic architectures of aging and mortality. *Nat. Med.* **31**, 1016–1025 (2025).
15. Ganna, A. & Ingelsson, E. 5 year mortality predictors in 498 103 uk biobank participants: a prospective population-based study. *Lancet* **386**, 533–540 (2015).
16. American College of Cardiology. ASCVD Risk Estimator Plus. Available: https://tools.acc.org/ASCVD-Risk-Estimator-Plus. Date accessed 30 April 2024.
17. Buergel, T. et al. Metabolomic profiles predict individual multidisease outcomes. *Nat. Med.* **28**, 2309–2320 (2022).
18. de Bruijn, R. F. & Ikram, M. A. Cardiovascular risk factors and future risk of Alzheimer's disease. *BMC Med.* **12**, 1–9 (2014).
19. Koene, R. J., Prizment, A. E., Blaes, A. & Konety, S. H. Shared risk factors in cardiovascular disease and cancer. *Circulation* **133**, 1104–1114 (2016).
20. Julkunen, H. et al. Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nat. Commun.* **14**, 604 (2023).
21. Thompson, D. J. et al. A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. *PLoS ONE* **19**, e0307270 (2024).
22. Elliott, J. et al. Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA* **323**, 636–645 (2020).
23. Mars, N. et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* **26**, 549–557 (2020).
24. Inouye, M. et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
25. Royston, P. Explained variation for survival models. *Stata J.* **6**, 83–96 (2006).
26. Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B. & Vasan, R. S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157–172 (2008).
27. Pencina, M. J., D'Agostino Sr, R. B. & Steyerberg, E. W. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* **30**, 11–21 (2011).
28. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
29. Altamirano, J. & Bataller, R. Cigarette smoking and chronic liver diseases. *Gut* **59**, 1159–1162 (2010).
30. Sato, S. et al. Elevated serum tyrosine concentration is associated with a poor prognosis among patients with liver cirrhosis. *Hepatol. Res.* **51**, 786–795 (2021).
31. Morgan, M. Y., Marshall, A., Milsom, J. P. & Sherlock, S. Plasma amino-acid patterns in liver disease. *Gut* **23**, 362–370 (1982).
32. Sunami, Y. NASH, fibrosis and hepatocellular carcinoma: lipid synthesis and glutamine/acetate signaling. *Int. J. Mol. Sci.* **21**, 6799 (2020).
33. Zakhari, S. Overview: how is alcohol metabolized by the body? *Alcohol Res. health* **29**, 245 (2006).
34. Parsons, R. E. et al. Independent external validation of the QRISK3 cardiovascular disease risk prediction model using UK Biobank. *Heart* **109**, 1690–1697 (2023).
35. National Institute for Health and Care Excellence. Cardiovascular disease: risk assessment and reduction, including lipid modification, NICE guideline NG238 https://www.ncbi.nlm.nih.gov/books/NBK554923/ (2023).
36. Odutayo, A. et al. Atrial fibrillation and risks of cardiovascular disease, renal disease, and death: systematic review and meta-analysis. *BMJ* **354**, i4482 (2016).
37. Aschard, H. et al. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am. J. Hum. Genet.* **90**, 962–972 (2012).
38. Ye, Y. et al. Interactions between enhanced polygenic risk scores and lifestyle for cardiovascular disease, diabetes, and lipid levels. *Circulation: Genom. Precis. Med.* **14**, e003128 (2021).
39. Surakka, I. et al. Sex-specific survival bias and interaction modeling in coronary artery disease risk prediction. *Circulation: Genom. Precis. Med.* **16**, e003542 (2023).
40. World Health Organization. WHO reveals leading causes of death and disability worldwide: 2000-2019. https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019 (2020).
41. Atarashi, K., Oyama, S. & Kurihara, M. Factorization machines with regularization for sparse feature interactions. *J. Mach. Learn. Res.* **22**, 1–50 (2021).
42. Gadd, D. A. et al. Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat. Aging* **4**, 939–948 (2024).
43. You, J. et al. Plasma proteomic profiles predict individual future health risk. *Nat. Commun.* **14**, 7817 (2023).
44. Sun, B. B. et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
45. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
46. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
47. Broyden, C. G. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA J. Appl. Math.* **6**, 76–90 (1970).
48. Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **13**, 317–322 (1970).
49. Goldfarb, D. A family of variable-metric methods derived by variational means. *Math. Comput.* **24**, 23–26 (1970).
50. Shanno, D. F. Conditioning of quasi-Newton methods for function minimization. *Math. Comput.* **24**, 647–656 (1970).
51. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2023).
52. Goff, D. C. et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *J. Am. Coll. Cardiol.* **63**, 2935–2959 (2014).

53. Arnett, D. K. et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **140**, e596–e646 (2019).
54. Allen, N. E. et al. Approaches to minimising the epidemiological impact of sources of systematic and random variation that may affect biochemistry assay data in UK Biobank. *Wellcome Open Res.* **5**, 222 (2020).
55. Watts, E. L. et al. Hematologic markers and prostate cancer risk: a prospective analysis in UK Biobank. *Cancer Epidemiol. Biomark. Prev.* **29**, 1615–1626 (2020).
56. Elliott, P. & Peakman, T. C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, 234–244 (2008).
57. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
58. Therneau, T. M. A package for survival analysis in R R package version 3.7-0 (2024).
59. Dardis, C. survmisc: Miscellaneous functions for survival data R package version 0.5.6 (2022).
60. Inoue, E. nricens: NRI for risk prediction models with time to event and binary response data. R package version 1.6 (2018).
61. Julkunen, H. & Rousu, J. survivalfm: Efficient modelling of interaction effects in proportional hazards survival models. R package version 1.0.0. https://doi.org/10.5281/zenodo.15355070 (2025).
62. Julkunen, H. & Rousu, J. Initial release of the analysis codes for “Machine learning for comprehensive interaction modelling improves disease risk prediction in UK Biobank" (2025) https://doi.org/10.5281/zenodo.15355122 (2025).
63. Julkunen, H. *Machine learning for precision medicine* (Doctoral thesis, Aalto University, 2025).

## Acknowledgements

This research has been conducted using data from UK Biobank via project ID 147811. The authors are grateful to all participants as well as the investigators and members of UK biobank for making these data open and accessible to the research community. This work was supported by the Research Council of Finland grants 339421 (Machine Learning for Systems Pharmacology, MASF, 2021-2025) and 345802 (AI technologies for interaction prediction in biomedicine, AIB, 2022-2024) to J.R. and Technology Industries of Finland Centennial Foundation funding (Aalto University House of AI) to H.J. The authors acknowledge the computational resources provided by the Aalto Science-IT project. Parts of this work appear in the doctoral thesis of Heli Julkunen, titled “Machine Learning for Precision Medicine", Aalto University, 2025[63].

## Author contributions

H.J. conceived the idea and designed the method with input from J.R. H.J. processed the data, wrote the code, performed the analyses, prepared the figures and wrote the manuscript. J.R. supervised the work and contributed to writing the manuscript. Both authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-61891-y.

**Correspondence** and requests for materials should be addressed to Heli Julkunen.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.