PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Personalized prediction for multiple chronic diseases by developing the multi-task Cox learning model

Shuaijie Zhang[1,2], Fan Yang[1,2]*, Lijie Wang[1,2], Shucheng Si[1,2], Jianmei Zhang[3], Fuzhong Xue[1,2]*

1 Department of Epidemiology and Health Statistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China, **2** National Institute of Health Data Science of China, **3** Department of Geriatrics, Weihai Municipal Hospital Affiliated Shandong University, 76 Heping Rd, Weihai, Shandong, China

* fanyang@sdu.edu.cn (FY); xuefzh@sdu.edu.cn (FX)

Citation: Zhang S, Yang F, Wang L, Si S, Zhang J, Xue F (2023) Personalized prediction for multiple chronic diseases by developing the multi-task Cox learning model. PLoS Comput Biol 19(9): e1011396. https://doi.org/10.1371/journal.pcbi.1011396

Editor: Eric Lofgren, Washington State University, UNITED STATES

Received: June 1, 2022

Accepted: July 26, 2023

Published: September 21, 2023

Data Availability Statement: This research has been conducted using the UK Biobank Resource (Application ID: 51470)and the Weihai physical examination dataset. Researchers could acquire this data by submitting an application to the UK Biobank (https://www.ukbiobank.ac.uk/) through the UK Biobank Access Management System (https://bbams.ndph.ox.ac.uk/ams/). This URL is a dedicated URL for application data. Once logged in, researchers can follow the offcial instructions to access the data. The Weihai physical examination data is individual-level data restricted for sharing

Abstract

Personalized prediction of chronic diseases is crucial for reducing the disease burden. However, previous studies on chronic diseases have not adequately considered the relationship between chronic diseases. To explore the patient-wise risk of multiple chronic diseases, we developed a multitask learning Cox (MTL-Cox) model for personalized prediction of nine typical chronic diseases on the UK Biobank dataset. MTL-Cox employs a multitask learning framework to train semiparametric multivariable Cox models. To comprehensively estimate the performance of the MTL-Cox model, we measured it via five commonly used survival analysis metrics: concordance index, area under the curve (AUC), specificity, sensitivity, and Youden index. In addition, we verified the validity of the MTL-Cox model framework in the Weihai physical examination dataset, from Shandong province, China. The MTL-Cox model achieved a statistically significant ($p<0.05$) improvement in results compared with competing methods in the evaluation metrics of the concordance index, AUC, sensitivity, and Youden index using the paired-sample Wilcoxon signed-rank test. In particular, the MTL-Cox model improved prediction accuracy by up to 12% compared to other models. We also applied the MTL-Cox model to rank the absolute risk of nine chronic diseases in patients on the UK Biobank dataset. This was the first known study to use the multitask learning-based Cox model to predict the personalized risk of the nine chronic diseases. The study can contribute to early screening, personalized risk ranking, and diagnosing of chronic diseases.

Author summary

Chronic diseases present a significant challenge to public healthcare systems, responsible for almost 60% of global fatalities. Addressing this issue requires the implementation of preventative treatments and interventions to enable the early detection of chronic diseases in those who are at risk. However, most research in the field of chronic disease risk prediction has focused only on single-disease prediction tasks, without adequate attention paid

due to ethical requirements. Access to this data may be requested via the department of Science and Education (slyykjk@163.com) at the Weihai Municipal Hospital.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2021YFF0704100) awarded to FY and (2020YFC2003500) awarded to FX and the National Natural Science Foundation of China (Grant No. 82173625) awarded to FX and (Grant No. 82273736) awarded to FY. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

to the connections between chronic diseases. Thus, a multitask learning Cox model was developed to predict nine typical chronic diseases (lung cancer, gastric cancer, esophagus cancer, colorectal cancer, liver cancer, hypertension, diabetes, stroke, and coronary heart disease), based on data from the UK Biobank. In addition, the validity of the MTL-Cox model framework was also demonstrated with the Weihai physical examination dataset in China. Our experimental results indicate that the MTL-Cox model could offer personalized absolute risk prediction for chronic diseases.

Introduction

According to a commonly accepted definition, chronic diseases are conditions that last for at least one year and require ongoing medical attention or limit daily activities [1], which are a leading cause of death and disability worldwide [2]. In the United States, chronic diseases account for approximately 84% of the nation's healthcare spending, which totals $3.8 trillion annually (Anderson G. Chronic Care: Making the Case for Ongoing Care. Princeton, NJ: Robert Wood Johnson Foundation, 2010. Available at: http://www.rwjf.org/content/dam/farm/reports/reports/2010/rwjf54583. Accessed on Apr. 02nd, 2022.). Furthermore, chronic diseases such as stroke, diabetes, cancer, and cardiovascular diseases like hypertension and *c*oronary *h*eart *d*isease (CHD) account for nearly 60% of global fatalities [3–5]. Lung cancer, gastric cancer, liver cancer, colorectal cancer, and esophageal cancer have high morbidity rates among both males and females [6]. As a result, this study aims to research the nine chronic diseases mentioned above.

Clinical personalized risk prediction algorithms are powerful tools for healthcare management [7]. These algorithms assist clinicians in developing effective intervention strategies for patients and optimize the allocation of medical resources. In view of their significance, numerous machine learning methods have been developed for estimating the personalized risk of patients [8–12]. However, developing clinical personalized risk prediction models is challenging due to the uncertain follow-up time of subjects. Given time and cost constraints, it is difficult to track all subjects until an event of interest occurs. In other words, not all subjects experience an event of interest during the follow-up period, and some drop out for various reasons, making it impossible to observe the event of interest. However, this does not imply that such events will never occur in the future. In statistics, these drop-out subjects are referred to as *right-censored* [13, 14], accounting for a large proportion of real-world health data [15], leading to bias in survival analysis [16]. Therefore, it is essential to consider right-censored subjects while predicting disease risk. This paper addresses the problem of right-censoring by utilizing the Cox proportional hazards model.

Chronic diseases, such as cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes, are closely linked from a physiological perspective (www.cdc.gov/chronicdisease/about/index.htm. Accessed on Sep. 17th, 2022.) [17]. Based on this insight, some epidemiological evidence suggested that there exist common risk factors causing chronic diseases [18, 19]. It is worth noting that one chronic disease can often be accompanied by others, which means that the correlation between these diseases can be used to prevent additional chronic diseases. However, most studies on chronic disease prevention use single-task learning-based methods that only focus on capturing individual risk probabilities, without considering the interplay between multiple diseases [20–26]. This approach often results in weak model generalization and limited ability to capture multiple diseases. To address these challenges, this study employs a *m*ulti*t*ask *l*earning (MTL) framework that leverages the correlations among multiple diseases

to improve model performance and generalization. The MTL framework, originally proposed by Caruana [27], considers two tasks to be related if they share the same feature space during model training. By learning multiple related tasks together, models can be optimized for improved performance [28].

In this paper, we address the problem of predicting multiple chronic diseases using right-censored data by proposing an MTL-based survival analysis method called MTL-Cox, which trains semi-parametric *Cox* proportional hazards models via a *m*ulti*t*ask *l*earning framework. By sharing latent representations across feature spaces, MTL enables multiple tasks related to chronic disease risk prediction to be learned in parallel. To ensure the model's robustness, we update MTL-Cox by adding the $L_{2,1}$ norm as a regularization term to promote similar parameter sparsity patterns among multiple chronic disease predictors. Additionally, we employ the proximal gradient method to train MTL-Cox to converge at a fast rate of $\mathcal{O}(1/\varepsilon)$. To evaluate the proposed framework, we demonstrate its risk prediction performance using demographic and clinical data from the UK Biobank and verify the validity of the MTL-Cox model framework on the Weihai physical examination dataset. The experimental results show that our proposed framework outperformed competing models, and Fig 1 conveys the big picture and flow process of MTL-Cox.

Our contributions are as follows:

1. Proposed an MTL-based survival analysis method called MTL-Cox for predicting multiple chronic diseases using right-censored data. Enabled multiple tasks related to chronic disease risk prediction to be learned in parallel by sharing latent representations across feature spaces via MTL.
2. Demonstrated the superior risk prediction performance of the proposed framework using demographic and clinical data from the UK Biobank and validated the MTL-Cox model framework on the Weihai physical examination dataset.
3. Defined five measurements to identify the personalized risk.

The rest of the paper is structured as follows. The Related Works section presents the existing literature on disease risk signaling using statistical and machine learning methods. The Notations and Preliminaries section introduces the notations used in this study and provides a basic background. The Model section provides a detailed explanation of the MTL-Cox framework. In the Materials section, we showcase data from the UK Biobank and the Weihai physical examination dataset, along with details of data processing. The Experiments and Results section demonstrates the multiple chronic disease risk prediction for patients and diseases separately. The Discussion section quantitatively evaluates the performance of the proposed model using five measurement metrics, including the concordance index, AUC, specificity, sensitivity, and Youden index. We summarize the essential findings in the Conclusion section. Lastly, the Future Work section discusses potential strategies for improving MTL-Cox.

Related works

Extensive literature has been devoted to the analysis of chronic disease survival analysis in the past few decades. The approaches employed in these studies can be broadly categorized into two groups: statistical methods and machine learning techniques.

Statistical methods

In this section, we present various statistical methods utilized for survival analysis, namely nonparametric, semi-parametric, and parametric methods.

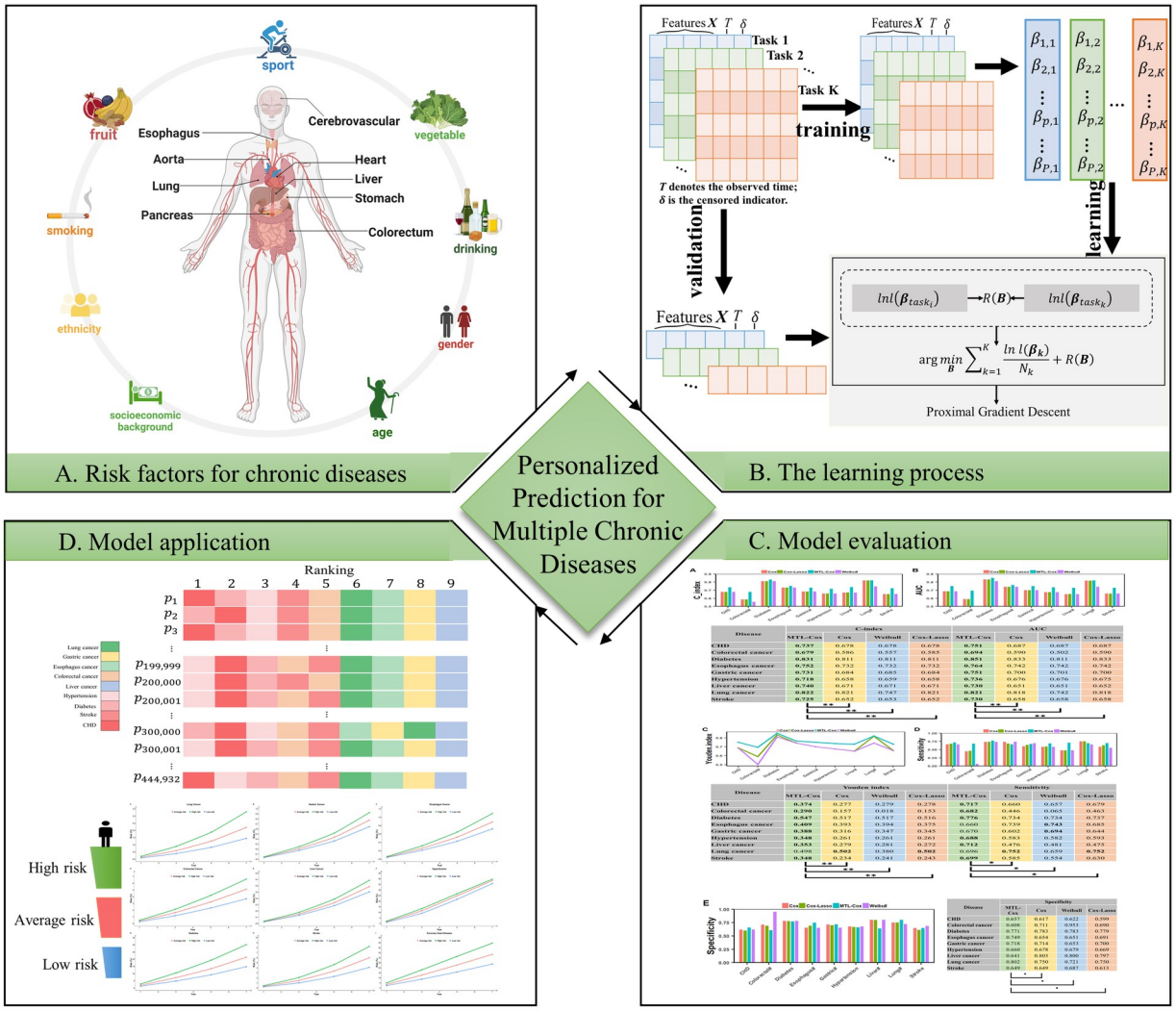


Fig 1. The framework of the MTL-Cox model. (A) The inner circle shows the main sites of the nine chronic diseases, and the outer circle demonstrates factors that affect chronic diseases in the UK Biobank data, which is detailed in the Materials Section. The figure is created with *Biorender*. (B) The flow of the MTL-Cox model construction and optimization, which detailed in the Model Section; (C) Concordance index, AUC, specificity, sensitivity, and Youden index are used to evaluate models, which detailed in the Experiments and results Section; (D) Applying the MTL-Cox model for chronic diseases personalized prediction, which detailed in the Experiments and results Section.

https://doi.org/10.1371/journal.pcbi.1011396.g001

Nonparametric methods, such as the Kaplan-Meier [29] and Life-Table methods [30], are typically employed when the data distribution is unavailable. However, these methods fail to capture individual differences in disease risk perception. Parametric methods, on the other hand, rely on a specific data distribution and are thus limited in their ability to adapt to diverse time-to-event samples for disease risk analysis [31, 32]. In the semi-parametric category, the Cox proportional hazards model is the most commonly used for survival analysis [33]. This method overcomes the shortcomings of both nonparametric and parametric methods by taking individual differences into account, and providing personalized predictions for each subject. Furthermore, the parameter estimation does not require survival times to follow a specific distribution. Therefore, we have employed the Cox model for personalized prediction of chronic disease survival.

Machine learning methods

Over the past few decades, the computing community has made significant contributions to the field of survival analysis for chronic diseases. The research in this area can be broadly classified into single- and multiple-disease prediction.

For single-disease prediction, various methods have been proposed by different researchers, such as Jiang et al. [34] who developed *s*upport *v*ector *m*achine (SVM)-based classification methods to identify gastric cancer, which showed promising results. Kumardeep et al. [35] formulated a robust survival analysis model to predict liver cancer by using autoencoder and SVM techniques. Siavash and Mohammad [36] proposed a novel decision tree-based method to identify risk factors of esophagus cancer and predict early readmission following esophagus cancer. Among the models proposed, some belong to a neural network, such as the knowledge-guided convolutional neural network model developed by Ye et al. [37] to predict the risk of mortality in critically ill patients with diabetes, which yielded a competitive AUC performance. Additionally, a Bayesian neural network was proposed to predict the survival of gastric cancer patients [38]. As AI technology evolves, researchers have started using ensemble learning methods and deep learning models for disease prediction, such as Diao et al. [39] adoption of E*x*treme *G*radient *Boost*ing (XGBoost) to predict common etiologies in patients with suspected secondary hypertension, and She et al. [40] use of the *deep* learning *surv*ival neural network model (DeepSurv) model to predict non-small cell lung cancer, which provided desirable prediction accuracy and individual survival information. However, these methods can only predict one disease at a time.

For multiple-disease prediction, unlike the methods mentioned above, an advanced machine learning technique called multitask learning has been developed to create an integrated multiple-disease predictor through parallel learning. In 2021, Feng et al. [41] proposed an MTL-based framework for predicting hypertension and type-2 diabetes, where a two-branch network was developed for each disease. Each branch consisted of two convolutional layers and two fully connected layers. The results demonstrated that the MTL model achieved better performance than single-task models. Additionally, Anusha et al. [42] utilized the deep belief network and recurrent neural network for multi-disease prediction, which showed superior performance. However, these models do not consider the right censoring.

The studies mentioned above have demonstrated promising results in disease prediction. However, there is still potential for further improvement. Most of the relevant methods only focus on predicting survival rates for single diseases and do not consider the correlation between chronic diseases. Additionally, while some studies have proposed models for predicting multiple diseases, these models have limitations in handling right-censored data. It is essential to address these issues from a perspective of robustness and generalization for multiple-disease survival analysis. In this study, we propose the use of semi-parametric Cox models for multiple diseases in the multitask learning framework. The MTL-Cox model is specifically designed for right-censored data and achieves excellent performance by simultaneously learning multiple related tasks.

Notations and preliminaries

The notations and preliminary definitions used in this study that has been presented in this section.

Notations

All vectors are denoted by bold lowercase letters (e.g., $\boldsymbol{x}_i$), while matrices are represented by bold uppercase letters (e.g., $\boldsymbol{X}$). A summary of the notations used in this paper is presented in Table 1.

Table 1. Notations of symbols that used in this paper.

Notation	Description
N	Number of subjects.
P	Number of features.
K	Number of tasks.
T	Survival/failure time, where the dimension is $\mathbb{R}^{N\times 1}$.
δ	Binary vector for event status, where the dimension is $N \times 1$.
$R(T_i)$	The risk set at T_i.
$\boldsymbol{X}$	The feature matrix, where the dimension is $\mathbb{R}^{N\times P}$.
$\boldsymbol{x}_p$	The feature vector of p-th feature for N subjects, where the dimension is $\mathbb{R}^{N\times 1}$.
$\boldsymbol{\beta}$	The coefficient vector for single task learning, where the dimension is $\mathbb{R}^{P\times 1}$.
β_p	The coefficient of p-th feature for single-task learning.
$\boldsymbol{\beta}_k$	The coefficient vector of k-th task's features for multitask learning, where the dimension is $\mathbb{R}^{P\times 1}$.
$\boldsymbol{B}$	The coefficients matrix of all tasks' features, where the dimension is $\mathbb{R}^{P\times K}$.

https://doi.org/10.1371/journal.pcbi.1011396.t001

Preliminaries

Cox proportional hazards model. The Cox model is a widely used method in survival analysis that effectively utilizes survival or failure time (T_i). A complete survival dataset consists of three elements: start time, end time, and end event. In this study on the UK Biobank dataset, the start time denotes the point at which subjects entered the cohort. The end time is either March 31, 2017 or July 31, 2019, with more details provided in Section 4.1. The end event refers to the occurrence of the target disease. The parameters of the Cox model are defined as follows.

*D*efinition 1 *Survival time*. Survival time defines the period between the starting and ending points. The starting time varies by the cohort study type. In static cohorts, the starting time is the same for all participants and is defined as the beginning of the study. In contrast, in dynamic cohorts, each subject enrolls at a different time, which is defined as the starting time. The ending time represents the occurrence of the event of interest, such as the onset, recurrence, or death. In this study, data from both the UK Biobank's dynamic cohort and the Weihai physical examination dynamic cohort are utilized. Therefore, survival time is defined as the time from the cohort's sequential enrollment to the event of interest. To clarify the definition of survival time, an example is illustrated in the Fig 2.

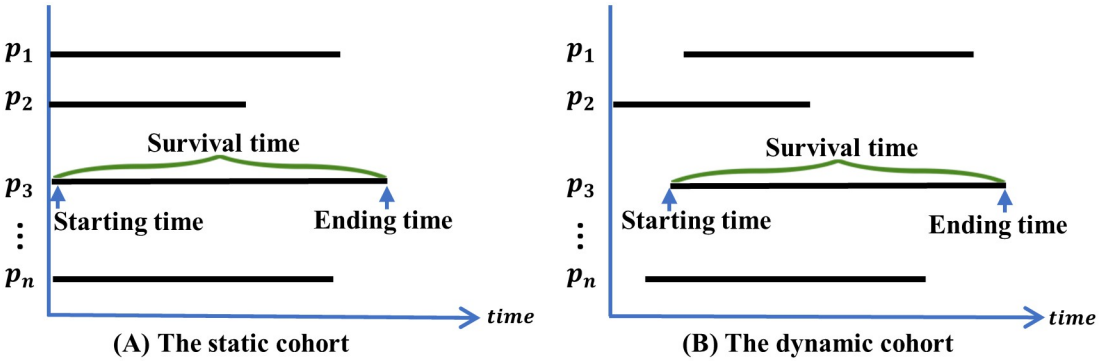


Fig 2. Examples of the static and dynamic cohort survival times. The static cohorts have a unified starting time; Subjects in dynamic cohorts with sequential enrolling time. For the notation of $P_{i_{i\in\mathbb{R}^+}}$ denotes the i-th specific subject.

https://doi.org/10.1371/journal.pcbi.1011396.g002

*D*efinition 2 *Censored data*. Subjects are considered censored when information about the event is not available at the time of occurrence due to loss to follow-up or non-occurrence of the outcome event before the trial ends [43]. Censored data can be categorized into three groups: left-censored, right-censored, and interval-censored. The left-censored group includes subjects who were at risk of developing the disease before entering the cohort. Since the Cox model can only handle right-censored data, these left-censored subjects are excluded at the beginning of the study. The right-censored group includes subjects who did not experience the event of interest by the endpoint of the trial. This study did not involve interval-censored data. An example is shown in Fig 3 to clarify the definition of censoring.

*D*efinition 3 *Uncensored data*. Uncensored data refer to events of interest that occurred during the trial period, and we know the specific point in time. In this study, uncensored data represent new cases during the trial period. We introduced a censoring indicator $\delta_i \in \{0, 1\}$ to indicate the censored type of data (i.e., $\delta_i = 1$ stands for an uncensored instance, and $\delta_i = 0$ means a censored instance).

*D*efinition 4 *Features*. Features represent attributes, properties, or characteristics of objectives [44].

*D*efinition 5 *Survival and hazard function*. The survival function $S(t)$ represents the probability that a subject's survival time that T is not earlier than a specific time t. Mathematically, the survival function can be expressed as follows:

$$S(t) = P(T > t). \tag{1}$$

Another commonly used function in survival analysis is the hazard function $h(t)$, which can be interpreted as the probability that a subject survives at time t and dies in $t+\Delta t$ ($t > 0$). The hazard function is given by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}. \tag{2}$$

*D*efinition 6 *Baseline*. The baseline is denoted as the beginning time of conducting the related research [45].

Multitask learning. *D*efinition 7 *Multitask Learning*. Given a set of K learning tasks, either all or a subset of which may be related, the goal of multitask learning is to train a model to learn all K tasks simultaneously, while utilizing the knowledge and information present in other tasks, to enhance the learning of each task's model [28].

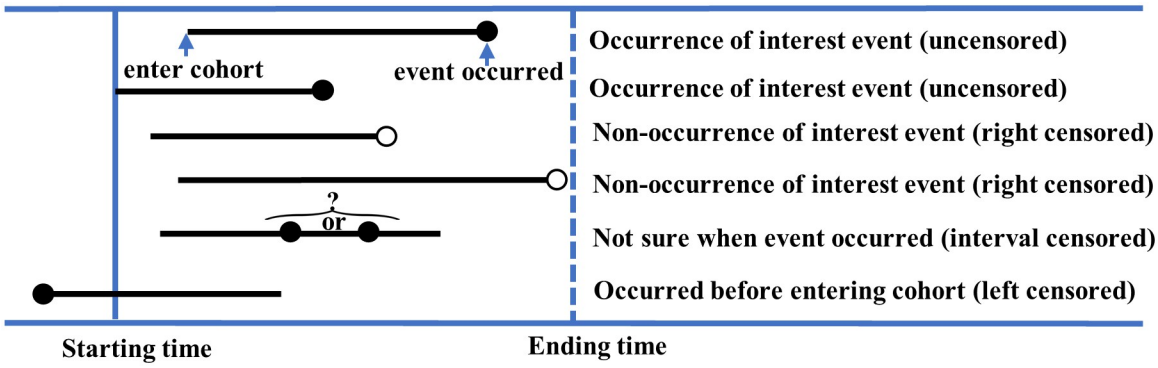


Fig 3. An illustration demonstrating the censoring definition.

https://doi.org/10.1371/journal.pcbi.1011396.g003

Multitask learning, a form of inductive transfer proposed by Caruana [27], is a powerful branch of machine learning that involves jointly training multiple tasks simultaneously, to utilize the correlation among different tasks. This approach can benefit the learning of model parameters by leveraging the knowledge contained across all tasks, resulting in improved learning efficiency and acting as a regularizer to prevent overfitting [28].

Disease risk metrics for personalized prediction. The individual risks are measured by the following metrics:

*D*efinition 8 *Basis risk* is the incidence rate of each age group within a certain period of time (e.g., 5 years, 10 years) representing the average risk level of disease in the same population and forms an important basis for classifying the risk level.

*D*efinition 9 *Relative risk (RR)* is an indicator of the strength of the association between exposure and morbidity (mortality), which cannot be used as a predictor of disease risk.

*D*efinition 10 *Absolute risk (AR)* denotes the probability that an individual with specific risk factors does not have the studied outcome (e.g., stroke) at age a, but the outcome occurs at age $(a + \mathcal{T})$, where $\mathcal{T}$ is the artificially defined follow-up time.

*D*efinition 11 *Relative absolute risk (RAR)* is the ratio of the absolute risk of an individual to the average absolute risk of the same age group for given risk factors, reflecting that the absolute risk of each individual is a multiple of the average absolute risk of the population of the same age group.

*D*efinition 12 *Excess absolute risk (EAR)* is the difference between the absolute risk of an individual and the average absolute risk (i.e., the population average basis risk) of the same age group for given risk factors, reflecting the difference between the absolute risk of each individual and the average absolute risk of the same age group.

$L_{2,1}$ norm-based regularizer. Overfitting is a common problem for model training [46], which can be diluted by utilizing the $L_{2,1}$ norm regularizer to alleviate the negative impact. The $L_{2,1}$ norm *de facto* was proposed based on the L_1 norm and L_2 norm. In machine learning, L_1 norm enforces sparsity in models, and L_2 norm penalties in some sense discourage sparsity. For a greater level of understanding of $L_{2,1}$ norm-based regularizer, we introduce the L_1 norm and L_2 norm at the beginning of the following section.

L_1 norm. The L_1 norm is defined as the sum of the absolute values of the elements of the parameter vector. The mathematical formula is as follows:

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{I} |x_i|, \tag{3}$$

where $\boldsymbol{x}$ denotes the parameter vector, x_i represents the *i-th* element of the parameter vector, and I is the total number of elements of vector $\boldsymbol{x}$.

L_2 norm. The L_2 norm is the square root of the inner product of the vector. The mathematical formula is given as follows:

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{I} x_i^2}. \tag{4}$$

Definition 13 *$L_{2,1}$ norm*. The $L_{2,1}$ norm is the sum of the L_2 norm of all rows in matrix $\boldsymbol{X}$, defined by

$$\|\boldsymbol{X}\|_{2,1} = \sum_{p=1}^{P} \|\boldsymbol{x}^p\|_2, \tag{5}$$

where p is the *p-th* row of $\boldsymbol{X}$.

The $L_{2,1}$ norm first computes the L_2 norm of the matrix $\boldsymbol{X}$ row vector and then computes the L_1 norm of all L_2 norm. The L_2 norm promotes a dense (non-zero) solution within each row, and the outer L_1 norm forces the L_2 norm of some rows to be zero, which means that all values in that row are zero. The $L_{2,1}$ norm regularization, therefore, reduces the weights of some rows to zero, which serves the purpose of feature selection [47].

The model

Here, we introduce our model framework in detail. We employed the model architecture from a unified multitask survival analysis framework to develop MTL-Cox [48]. MTL-Cox consists of two parts: 1) The Multiple Cox model is used for multiple chronic diseases and 2) regularization terms make multiple diseases share parameters.

Cox model

The Cox model is expressed by the hazard function $h(t)$. The Cox model can be represented with the feature matrix $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_P\}$ as

$$h(t; \boldsymbol{X}) = h_0(t)exp(\beta_1 \boldsymbol{x}_1 + \cdots + \beta_P \boldsymbol{x}_P). \tag{6}$$

In the above equation, $h_0(t)$ represents the baseline hazard function when the feature vector is 0. The exponential function is denoted by exp ($exp(x) = e^x$). The coefficients $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_P\}$ measure the impact, or effect size, of each feature. $h_0(t)$ does not need to follow a specific distribution, and it cannot be estimated [49]. Therefore, $h_0(t)$ is considered nonparametric. However, the exponential part (i.e., $exp(\beta_1 \boldsymbol{x}_1 + \cdots + \beta_P \boldsymbol{x}_P)$) takes the form of a parametric model, where the parameters can be estimated from the observed feature values of the subjects [49].

Parameter estimates in the Cox model are obtained by maximizing the partial likelihood. First, let $T_1 \leqslant T_2 \leqslant \cdots \leqslant T_N$ be the ordered time to the event of interest for N subjects. For each survival time T_i, all subjects with a survival time greater than T_i constitute a risk set, denoted as $R(T_i)$. Subjects in $R(T_i)$ survived before T_i, but were still at risk of experiencing the event of interest. Thus, the partial likelihood function is formalized in Eq (7):

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{N} \left[\frac{exp(\sum_{p=1}^{P} x_{ip}\beta_p)}{\sum_{s \in R(T_i)} exp(\sum_{p=1}^{P} x_{sp}\beta_p)} \right]^{\delta_i}, \tag{7}$$

where $\delta_i = 1$ represents that a subject falls ill at time T_i, $\delta_i = 0$ represents a censored subject, P is the number of features, and s denotes subjects in the risk set $R(T_i)$. Finally, the coefficients can be estimated by minimizing the negative log partial likelihood:

$$\ln l(\boldsymbol{\beta}) = -\sum_{i=1}^{N} \delta_i \left[\left(\sum_{p=1}^{P} x_{ip}\beta_p \right) - ln \sum_{s \in R(T_i)} exp\left(\sum_{p=1}^{P} x_{sp}\beta_p \right) \right]. \tag{8}$$

The first derivative function is

$$\ln' l(\boldsymbol{\beta}) = -\sum_{i:\ \delta_i=1} (\boldsymbol{X}_i - \frac{\sum_{s \in R(T_i)} exp(\boldsymbol{X}_s\boldsymbol{\beta})\boldsymbol{X}_s}{\sum_{s \in R(T_i)} exp(\boldsymbol{X}_s\boldsymbol{\beta})}), \tag{9}$$

and the Hessian matrix of the negative log partial likelihood is

$$\ln'' l(\boldsymbol{\beta}) \quad = -\sum_{i:\ \delta_i=1} \left(\frac{\sum_{s \in R(T_i)} exp(\boldsymbol{X}_s\boldsymbol{\beta})\boldsymbol{X}_s\boldsymbol{X}_s^T}{\sum_{s \in R(T_i)} exp(\boldsymbol{X}_s\boldsymbol{\beta})} - \frac{[\sum_{s \in R(T_i)} exp(\boldsymbol{X}_s\boldsymbol{\beta})\boldsymbol{X}_s][\sum_{s \in R(T_i)} exp(\boldsymbol{X}_s\boldsymbol{\beta})\boldsymbol{X}_s^T]}{[\sum_{s \in R(T_i)} exp\boldsymbol{X}_s\boldsymbol{\beta})]^2} \right), \tag{10}$$

where $\boldsymbol{X}_s \in \mathbb{R}^{1\times P}$ denotes the feature values of the *s-th* subject, and $\boldsymbol{X}_s^T$ is the transpose of $\boldsymbol{X}_s$. Using the first derivative function and Hessian matrix, we can minimize the negative log partial likelihood.

The traditional Cox model is a well-recognized statistical technique for exploring the relationship between the survival of a subject and several features. However, the traditional Cox model belongs to the category of single-task learning and ignores the relationship between multiple diseases. To solve the above issues, we propose the multitask Cox model, which learns the survival analysis problems of multiple related chronic diseases in parallel.

Multitask Cox model

Framework. The chronic diseases studied in this research share common risk factors, and therefore, we used parameter-based multitask learning with a low-rank assumption to estimate the coefficients of different chronic disease predictors that share a low-dimensional subspace [28].

Parameter-based multitask learning can be classified into two sub-categories: hard parameter-sharing and soft parameter-sharing mechanisms. The hard parameter-sharing mechanism shares the hidden layers between all tasks while keeping several task-specific output layers [50], as shown in Fig 4. However, since all tasks need to use the same set of parameters on shared-bottom layers, the hard parameter-sharing mechanism has optimization conflicts caused by task differences [51]. On the other hand, the soft parameter-sharing mechanism is associated with the identity parameters in each task. The differences between the parameters of the model are then regularized to encourage sharing of knowledge among the parameters [50]. For example, Obozinski et al. [52] used the $L_{2,1}$ norm for regularization, and the trace norm is also a common regularized term suitable for MTL [53]. The soft parameter-sharing mechanism is regularized by sharing prior parameters to train stabilities while forgetting overly strong task-related constraints. Therefore, we developed the multitask Cox model using the soft parameter-sharing mechanism as the learning strategy.

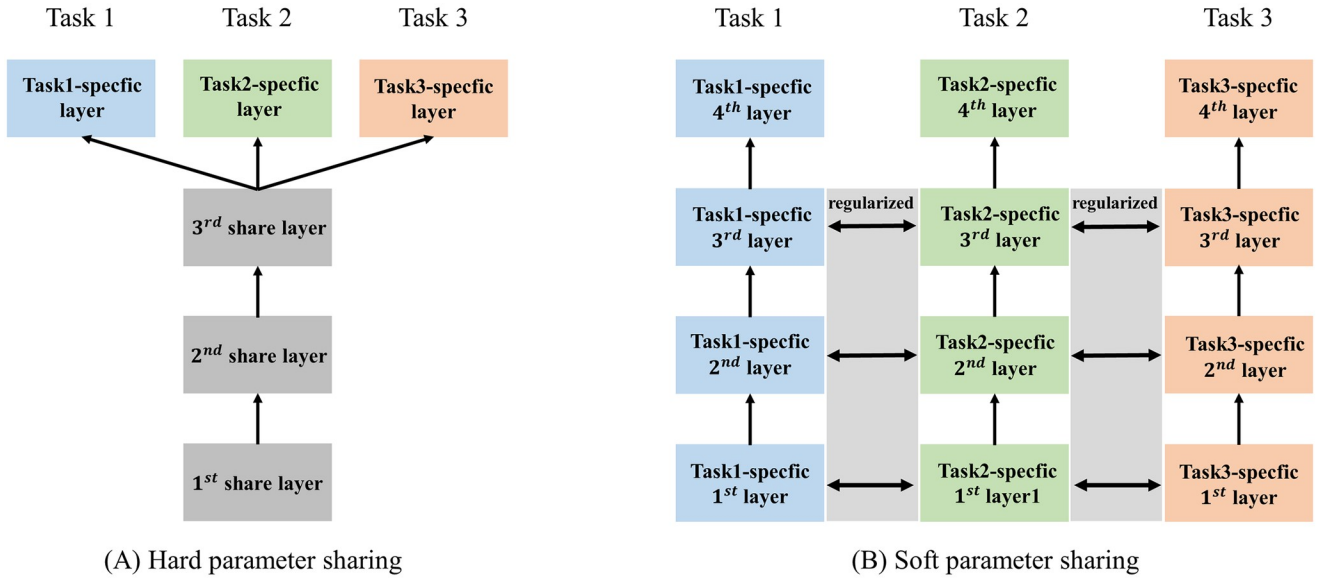


Fig 4. Strategies of parameter sharing in the multitask learning framework. (A) Hard parameter-sharing mechanism transfers the parameter training representations from hidden layers across all tasks. (B) The soft parameter sharing mechanism shares penalized information between tasks by regularization.

https://doi.org/10.1371/journal.pcbi.1011396.g004

A general regularized empirical loss paradigm for MTL can be formulated as

$$\arg\min_{\boldsymbol{B}} L(\boldsymbol{B}) + R(\boldsymbol{B}). \tag{11}$$

In this paper, a chronic disease is a task. Cox models of multiple chronic diseases are denoted by $L(\boldsymbol{B})$; that is, $L(\boldsymbol{B}) = \sum_{k=1}^{K} \frac{\ln\ l(\boldsymbol{\beta}_k)}{N_k}$. $\boldsymbol{\beta}_k$, N_k, and $\ln\ l(\boldsymbol{\beta}_k)$ represent the parameters to be estimated, the number of training instances, and the empirical loss on the training set with respect to the *k-th* task, respectively. $\boldsymbol{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ \ldots, \ \boldsymbol{\beta}_K]\mathbb{R}^{P\times K}$, the *k-th* column of $\boldsymbol{B}$ (i.e., $\boldsymbol{\beta}_k\mathbb{R}^{P\times 1}$) denotes the parameters to be estimated for the *k-th* task. $R(\boldsymbol{B})$ is the regularization term to encode the task relatedness.

The $L_{2,1}$ norm combines the strengths of the L_1 and L_2 norms, allowing the model to generalize in both sparse and dense spaces. And the $L_{2,1}$ norm can be expressed using the Eq 12:

$$R(\boldsymbol{B}) = \|\boldsymbol{B}\|_{2,1} = \sum_{p=1}^{P} \|\boldsymbol{\beta}^p\|_2 = \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \beta_{p,k}^2} \quad , \tag{12}$$

where $\boldsymbol{\beta}^p$, the *p-th* feature of all tasks, is the *p-th* row of $\boldsymbol{B}$.

When the number of tasks is one, matrix $\boldsymbol{B}$ only has a single column. In this scenario, Eq (12) can be expressed as the L_1 norm regularization optimization problem, where it equals the sum of absolute values in vector $\boldsymbol{\beta}$. On the other hand, if there are multiple tasks, the $L_{2,1}$ norm of $\boldsymbol{\beta}^p$, combines the parameters for the *p-th* feature of all tasks. This type of regularization takes into account the relationship between multiple tasks to select features instead of solely relying on the strength of a single input variable, as single-task learning does. Therefore, the $L_{2,1}$ norm regularization can learn multiple tasks simultaneously and synthesize the information from various tasks to enhance the model's performance.

To maintain a balance between $L(\boldsymbol{B})$ and $R(\boldsymbol{B})$, a positive regularization parameter, λ, is introduced. Consequently, the loss function of the proposed multitask Cox model can be expressed as:

$$\min_{\boldsymbol{B}} \sum_{k=1}^{K} \frac{\ln l(\boldsymbol{\beta}_k)}{N_k} + \lambda \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} \beta_{p,k}^2} \quad , \tag{13}$$

where $\ln l(\boldsymbol{\beta}_k)$ is the loss function of the Cox model for the *k-th* task; that is,

$$\ln l(\boldsymbol{\beta}_k) = -\sum_{i=1}^{N} \delta_{i,k} \left[\left(\sum_{p=1}^{P} \boldsymbol{x}_{ip,k} \boldsymbol{\beta}_{p,k} \right) - \ln \sum_{s\in R(T_i)} exp\left(\sum_{p=1}^{P} \boldsymbol{x}_{sp,k} \boldsymbol{\beta}_{p,k} \right) \right]. \tag{14}$$

Optimization. In this section, we present the details of learning and optimizing the model.

As the Hessian matrix of the negative log partial likelihood is nonnegative and the $L_{2,1}$ regularization norm is convex [54], the objective function Eq (11) is guaranteed to be convexity. Furthermore, since the $L_{2,1}$ norm is non-smooth [55], the proximal gradient has been employed to optimize the model along with converging at a fast rate $\mathcal{O}(1/\varepsilon)$ [56].

Choosing the initial $\boldsymbol{B}^{(0)}$ and repeating for $j = 1, 2, 3, \ldots$

$$\boldsymbol{B}^{(j)} = prox_{t_j}(\boldsymbol{B}^{(j-1)} - t_j \nabla L(\boldsymbol{B}^{(j-1)})), \tag{15}$$

where

$$\text{prox}_t = \arg\min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{B} - G\|_2^2 + \lambda\|\boldsymbol{B}\|_{2,1}, \tag{16}$$

where $\boldsymbol{G}$ is gradient step; that is, $\boldsymbol{G} = \boldsymbol{S} - \frac{1}{\gamma}\nabla L(\boldsymbol{S})$. In addition, γ is the step size, and $\boldsymbol{S}$ is the current search point. $\boldsymbol{S}$ can be written as

$$\boldsymbol{S}^{(j)} = \boldsymbol{B}^{(j)} + t_j(\boldsymbol{B}^{(j)} - \boldsymbol{B}^{(j-1)}), \tag{17}$$

t_j denotes the combination scalar of previous points. $\nabla L(\boldsymbol{S})$ is the gradient of the loss function at search point $\boldsymbol{S}$. For all K tasks, $\nabla L(\boldsymbol{S})$ is formulated as

$$\nabla L(\boldsymbol{S}) = \left[\frac{\ln' l(\boldsymbol{S}_1)}{N_1}, \frac{\ln' l(\boldsymbol{S}_2)}{N_2}, \ldots, \frac{\ln' l(\boldsymbol{S}_K)}{N_K}\right], \tag{18}$$

$\boldsymbol{S}_i$ is the parameter to be estimated for the *i-th* task, and is given in the *i-th* column of $\boldsymbol{S}$. N_i denotes the number of training instances of the *i-th* task. $\ln' l(\boldsymbol{S}_i)$ is the derivative of the negative log partial likelihood, and all tasks $\ln' l(\boldsymbol{S}_i)$ share the same formulation:

$$\ln' l(\boldsymbol{\beta}) = -\sum_{i=1}^{N}\delta_i\left[\sum_{p=1}^{P}\boldsymbol{x}_{ip} - \frac{\sum_{s\in R(T_i)}\boldsymbol{X}_s exp(\boldsymbol{X}_s\boldsymbol{\beta})}{\sum_{s\in R(T_i)} exp(\boldsymbol{X}_s\boldsymbol{\beta})}\right]. \tag{19}$$

In citing from [57], we derived the following theorem refereed the convergence of proximal gradient descent:

Theorem 1 Let $f(\boldsymbol{B}) = \min_{\boldsymbol{B}} L(\boldsymbol{B}) + \lambda R(\boldsymbol{B})$. Proximal gradient descent with the fixed step size $t \leq 1/L$ satisfies the following Eq (20)

$$f\left(\boldsymbol{B}^{(j)}\right) - f* \leq \frac{\|\boldsymbol{B}^{(0)} - \boldsymbol{B}*\|_2^2}{2tj}. \tag{20}$$

The proximal gradient descent converge in the rate of $\mathcal{O}(1/j)$ inference from Eq (20). The detailed proof procedure is available in [57]. The pseudo-code of the MTL-Cox model is presented in Algorithm 1.

Algorithm 1: Proximal gradient algorithm used for model learning.

Input: Feature matrix $\boldsymbol{X}$, initial coefficient matrix $\boldsymbol{B}^{(0)}$, and corresponding regularization scales.

Output: $\hat{\boldsymbol{B}}$

Initialization: $\boldsymbol{B}^{(1)} = \boldsymbol{B}^{(0)}$, $q_{-1} = 0$, $q_0 = 1$, $\gamma_0 = 1$, and $j = 1$;

repeat

  Set $t_j = \frac{q_{i-2}-1}{q_{i-1}}$, $\boldsymbol{S}^{(j)} = \boldsymbol{B}^{(j)} + t_j(\boldsymbol{B}^{(j)} - \boldsymbol{B}^{(j-1)})$; // the current search point is defined as a combination of the previous two search points

  for $i = 1, 2, \ldots$ **do**

    Set $\gamma = 2^i\gamma_{j-1}$; // select the optimal step size by the line search strategy

    $\boldsymbol{B}^{(j)} = prox_{t_j}(\boldsymbol{B}^{(j-1)} - t_j\nabla L(\boldsymbol{B}^{(j-1)}))$; // Proximal gradient descent

    if $(\boldsymbol{B}^{(j+1)}) \leq \prod_{\gamma,\boldsymbol{S}^{(i)}}(\boldsymbol{B}^{(i+1)})$ **then**

      $\gamma_j = \gamma$, **break**

    end if

  end for

  $q_j = \frac{1+\sqrt{1+4q_{j-1}^2}}{2}$;

  j++;

until Convergence of $\boldsymbol{B}^{(j)}$;

14: $\hat{\boldsymbol{B}} = \boldsymbol{B}^{(j)}$

Materials

In this paper, we first constructed an MTL-Cox model for nine chronic diseases based on the UK Biobank dataset. Subsequently, we built another MTL-Cox model for five cancers within these chronic diseases using the Weihai physical examination dataset to validate the effectiveness of the MTL-Cox model framework. Here, we first describe the dataset used in our experiment and demonstrate the process of feature selection based on the UK Biobank dataset. More details about the Weihai physical examination dataset can be found in the Supplementary file S1 File.

Ethics statement

The study protocol was approved by the North West Multi-center Research Ethics Committee, the Patient Information Advisory Group in England and Wales, and the Community Health Index Advisory Group in Scotland. All participants in the surveys gave their informed consent. The patients/participants provided their written informed consent to participate in this study.

Datasets

The data utilized in this study were obtained from the UK Biobank, a dataset for cohort studies that was used to evaluate the MTL-Cox model's ability to predict chronic diseases. The UK Biobank recruited 500,000 participants between 2006 and 2010, ranging in age from 37 to 73 years old [58–60]. Participants attended their closest of 22 assessment centers across England, Wales, and Scotland during a baseline assessment visit, which included information on their socioeconomic and demographic characteristics, general health and medical history, lifestyle, and diet. All participants provided written informed consent, and the study protocol was approved by the North West Multi-Centre Research Ethics Committee.

Following the 2020 World Health Organization report on chronic diseases [61], we selected nine typical chronic diseases for our study, including lung cancer, gastric cancer, esophageal cancer, colorectal cancer, liver cancer, hypertension, diabetes, stroke, and *c*oronary *h*eart *d*isease (CHD). The cancer records for the UK Biobank cohort were mostly completed by March 2017 (https://epi.grants.cancer.gov/cohort-consortium/members/ukb.html. Accessed on Mar. 29th, 2022; https://www.nature.com/collections/bpthhnywqk. Accessed on Mar. 29th, 2022.). Therefore, to improve prediction accuracy, we defined the follow-up time as follows: participant follow-up began at the time of inclusion in the UK Biobank study. We categorized the end date of follow-up into two cases: for cancers, participant records were available until either March 31, 2017, or until the patient was diagnosed with cancer. However, for other diseases (i.e., hypertension, diabetes, stroke, and CHD), participant follow-up ended on July 31, 2019. We excluded participants who had been diagnosed with the target disease before entering the cohort. Follow-up was censored if events of interest were not observed for various reasons. The censored statistics of the nine chronic diseases are shown in Fig 5. We defined the nine chronic disease codes using the International Classification of Diseases, 10th edition (ICD-10), as follows: lung cancer, C34; gastric cancer, C16; esophageal cancer, C15; colorectal cancer, C18-C20; liver cancer, C22; hypertension, I10-I15; diabetes, E10-E14; stroke, I60-I64; and CHD, I20-I25.

Feature selection

Based on expert knowledge, we acquired 78 features from the UK Biobank dataset, comprising 58 clinical features and 20 demographic features. However, the original encoding of categorical

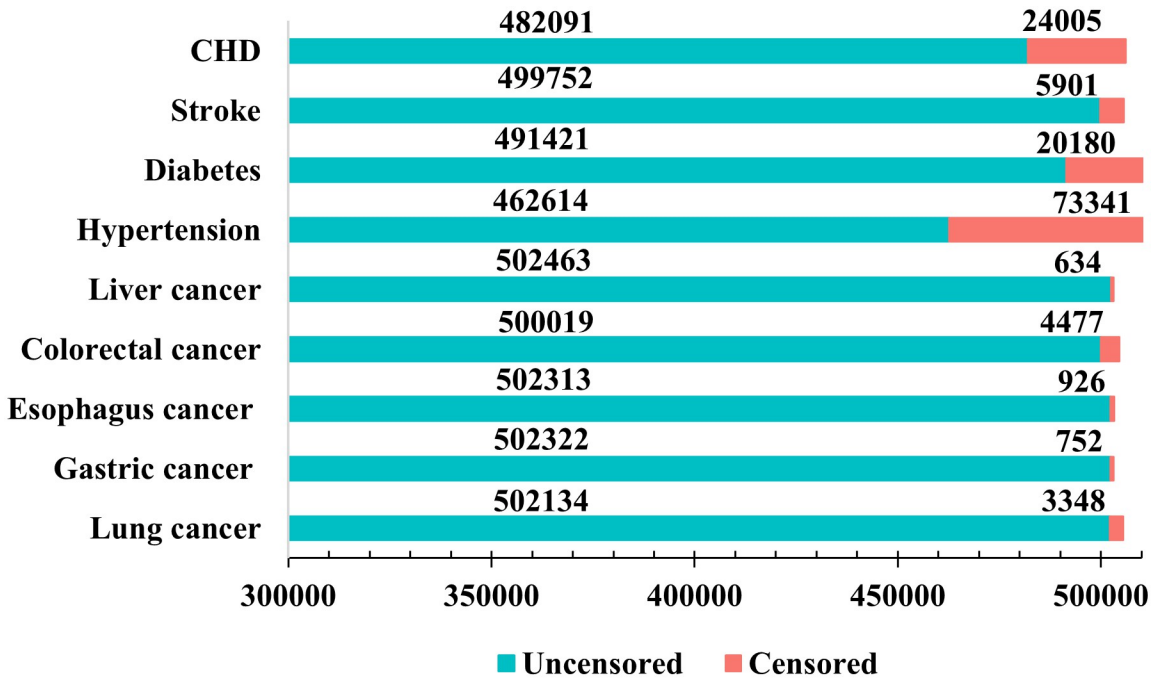


Fig 5. The censored statistics of the nine chronic diseases. The definitions of censored and uncensored are presented in the section Preliminaries.

https://doi.org/10.1371/journal.pcbi.1011396.g005

variables was not suitable for data analysis, thus we recoded them (as listed in S15 Table). Additionally, the continuous variables were standardized. Detailed information on demographic and clinical features can be found in the Supplementary file S1 Table, while baseline information for the nine chronic diseases is presented in the Supplementary file S2 Table.

It is important to note that when it comes to disease prediction models, having more features does not necessarily mean better performance. We prefer using a limited number of features that produce the best prediction outcomes. This approach not only reduces the cost and effort required to collect subject information but also reduces computational requirements and time. Hence, we used mono-factor analysis and forward regression sequentially to select 36 features related to chronic diseases from the 78 features (as shown in Fig 6). Furthermore, the summary of selected features is given in Table 2. The union of the nine disease variables was used as input features for MTL-Cox.

Experiments and results

Here, we introduce the main performance metrics in the survival analysis and compare the performance with other survival analysis models.

Performance metrics

Because of the existence of right-censored data in the survival data, the standard evaluation metrics used for regression analysis, such as a sum of the mean square errors R^2 and root mean squared error, are not used for measuring the prediction performance of survival analysis models. Therefore, we used five metrics (i.e., concordance, AUC, specificity, sensitivity, and Youden index) to comprehensively evaluate the performance of the model. These metrics are commonly used in survival analysis.

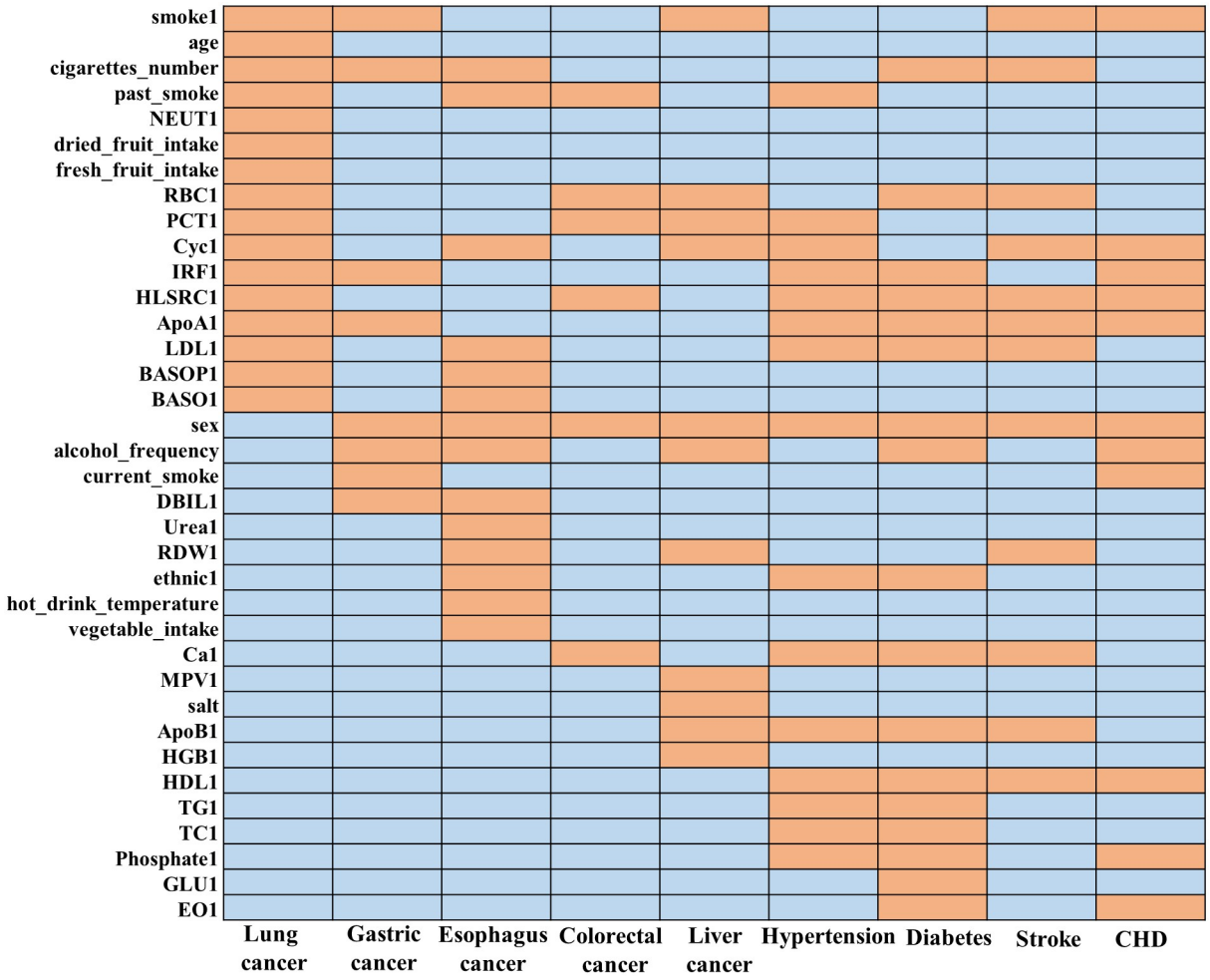


Fig 6. The selected features of nine chronic diseases. The color of orange represents the selected features of the corresponding disease; other features are represented in the color of blue.

https://doi.org/10.1371/journal.pcbi.1011396.g006

Concordance index. The concordance index (C-index), proposed by Franke Harrell Jr., is an evaluation metric commonly used in survival analysis [62, 63]. It measures the proportion of pairs of patients for whom the predicted outcomes are concordant with the actual outcomes. In other words, it estimates the probability that the predicted outcome will match the actual observed outcome. The C-index can be mathematically expressed as follows:

$$C = P(\hat{y}_1 > \hat{y}_2 | y_1 > y_2), \tag{21}$$

where $\hat{y}_i (i = 1, 2)$ is the predicted value, and $y_i (i = 1, 2)$ is the actual observation value. In the survival analysis model, subjects with a lower risk should survive longer, so the C-index can be computed using

$$\hat{C} = \frac{1}{n} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I\left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}} > \boldsymbol{x}_j \hat{\boldsymbol{\beta}}\right), \tag{22}$$

where n represents the number of comparable pairs, I denotes the indicator function, and $\hat{\boldsymbol{\beta}}$ represents the vector of estimated parameters. The calculation example of n is illustrated in detail in Fig 7. The C-index ranges from 0.5 to 1, where a C-index of 0.5 represents random

Table 2. A summary description of 36 chronic diseases related selected features.

Name	Description	Name	Description
smoke1	Smoking status	sex	Gender
age	Age at recruitment	alcohol_frequency	Alcohol intake frequency
cigarettes_number	Number of cigarettes currently smoked daily	current_smoke	Current smoking status
past_smoke	Past tobacco smoking	DBIL1	Direct bilirubin
NEUT1	Neutrophill count	Urea1	Urea
dried_fruit_intake	Dried fruit intake	RDW1	Red blood cell distribution width
fresh_fruit_intake	Fresh fruit intake	ethnic1	Ethnic background
RBC1	Red blood cell (erythrocyte) count	hot_drink_temperature	Hot drink temperature
PCT1	Platelet crit	vegetable_intake	Vegetable intake
Cyc1	Cystatin C	Ca1	Calcium
IRF1	Immature reticulocyte fraction	MPV1	Mean platelet (thrombocyte) volume
HLSRC1	High light scatter reticulocyte count	salt	Salt added to food
ApoA1	Apolipoprotein A	ApoB1	Apolipoprotein B
LDL1	LDL direct	HGB1	Hemoglobin concentration
BASOP1	Basophill percentage	HDL1	HDL cholesterol
BASO1	Basophill count	TG1	Triglycerides
TC1	Cholesterol	Phosphate1	Phosphate
GLU1	Glucose	EO1	Eosinophill count

https://doi.org/10.1371/journal.pcbi.1011396.t002

prediction and implies that the model has no predictive effect. On the other hand, a C-index of 1 indicates perfect consistency, suggesting that the model's predictions match the actual outcomes completely.

AUC. *A*rea *U*nder the *C*urve (AUC) is a commonly used evaluation metric for measuring the performance of binary classification models. It is calculated as the area under the *R*eceiver *O*perating *C*haracteristic (ROC) curve. The ROC curve plots the true positive rate (sensitivity) on the y-axis and the false positive rate (1-specificity) on the x-axis. AUC is preferred as an evaluation criterion for models because sometimes the ROC curve does not clearly indicate which classifier performs better. AUC can be calculated by determining the area under the ROC curve. The value range for AUC is between 0.5 and 1, where a higher AUC indicates better model performance.

Specificity, sensitivity, and Youden index. The performance of models can also be evaluated using sensitivity, specificity, and the Youden index:

- Specificity: refers to the fraction of those without the disease who will have a negative test result. According to the definition of specificity, the value of specificity is expressed as

$$Specificity = \frac{TN}{FP + TN}, \tag{23}$$

  where TN denotes the number of true-negative cases, FP is the number of false-positive cases. In addition, 1-*Specificity* represents the misdiagnosis rate.

- Sensitivity: refers to the proportion of patients who are actually sick that the model can correctly determine as patients. The sensitivity can be calculated via the following formula:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{24}$$

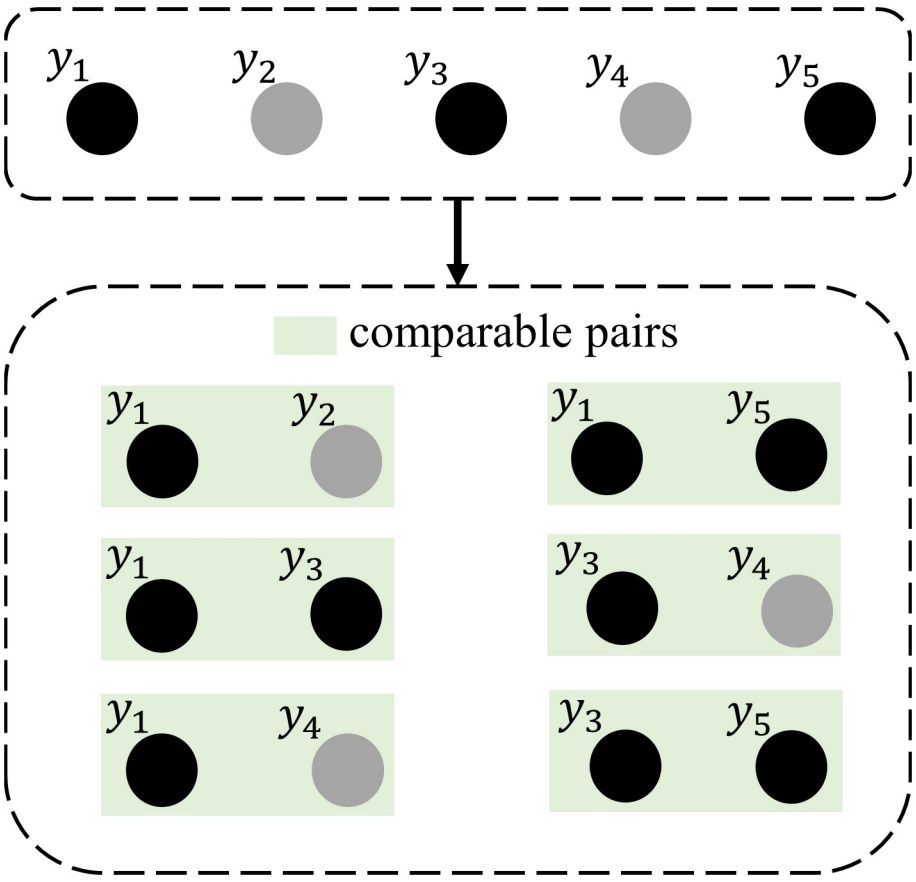


Fig 7. An example of calculating the number of comparable pairs for the C-index ($y_1 \preccurlyeq y_2 \preccurlyeq y_3 \preccurlyeq y_4 \preccurlyeq y_5$). $\preccurlyeq$ based on the timeline. Black circles denote uncensored observations, and gray circles indicate censored observations.

https://doi.org/10.1371/journal.pcbi.1011396.g007

where *TP* is truly positive, and *FN* denotes false negative. Missed diagnosis rate can be formulated as 1-*Sensitivity*.

- Youden index: represents the model's total ability to discover real patients and non-patients. An excellent model should keep both missed and misdiagnosed rates as low as possible, and the Youden index takes these two aspects into account. The Youden index can be computed as

$$Youden = Specificity + Sensitivity - 1. \tag{25}$$

The higher the Youden index, the better the performance of the model, and the greater the degree of authenticity.

Competing methods

The performance of the MTL-Cox model was compared to several single-task learning survival analysis models, including Cox, the Weibull regression model, and Cox-LASSO. This section provides the details of the Weibull regression and Cox-LASSO models.

A. Weibull regression model

The Weibull regression model is a type of parametric survival analysis model [64]. In this model, the hazard function in proportional hazards can be expressed by [65]

$$\begin{aligned} h(t, \boldsymbol{x}, \boldsymbol{\beta}, \lambda) &= \lambda t^{\lambda-1} e^{-\lambda(\beta_0 + \boldsymbol{x}\boldsymbol{\beta})} \\ &= \lambda t^{\lambda-1} e^{-\lambda\beta_0} e^{-\lambda\boldsymbol{\beta}'\boldsymbol{x}'} \\ &= \lambda\gamma t^{\lambda-1} e^{-\lambda\boldsymbol{\beta}'\boldsymbol{x}'} \\ &= h_0(t) e^{\boldsymbol{\theta}\boldsymbol{x}'}, \end{aligned} \tag{26}$$

where $\boldsymbol{x}$ is a vector with dimensions $1 \times p$, $\boldsymbol{\beta}$ is a vector with dimensions $p \times 1$, and p represents the p-th feature. The baseline hazard function is $h_0(t) = \lambda\gamma t^{\lambda-1}$, while the scale parameter is represented by $\gamma = \frac{1}{\sigma}$ and the shape parameter by λ. The parameter $\boldsymbol{\theta}_{1\times p}$ has a hazard ratio interpretation for subjects. The $\boldsymbol{\beta}$ vector is estimated by maximum likelihood estimation.

B. Cox-LASSO

Cox-LASSO regression can refine the model by constructing a penalty function [66, 67]. Specifically, the LASSO technique minimizes the residual sum of squares while imposing a constraint on the sum of the absolute coefficients. This approach helps to address the issue of multicollinearity and leads to the better predictive performance of the model [66]. The expressions for the Cox-LASSO and Cox models are the same, as shown in Eq (6), but the estimated coefficients, $\boldsymbol{\beta}$, differ between the two models. In the Cox model, maximum likelihood estimation is used to estimate the coefficients $\boldsymbol{\beta}$. On the other hand, in the Cox-LASSO model, $\boldsymbol{\beta}$ is estimated by introducing lasso term constraints proposed by Tibshirani in 1997 [66], as shown in Eq (27).

$$\hat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{B}} \ln l(\boldsymbol{\beta}), \text{subject to} \sum |\beta_p| \leq s, \tag{27}$$

where s is a specified parameter following the constraint of $s > 0$ [66], $l(\boldsymbol{\beta})$ stands for the log partial likelihood (Eq (7)), and $\ln l(\boldsymbol{\beta})$ denotes the partial likelihood function that is denoted as:

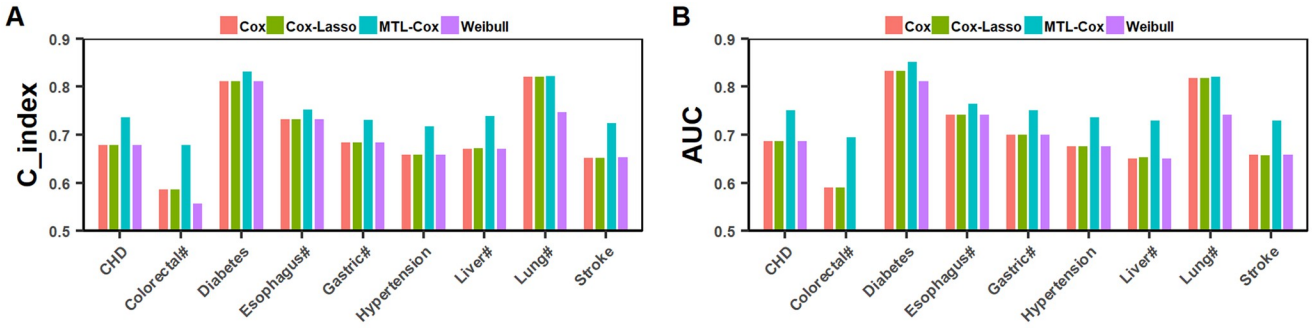
$$\ln l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \delta_i \left[\left(\sum_{p=1}^{P} x_{ip}\beta_p \right) - \ln \sum_{s\in R(T_i)} exp\left(\sum_{p=1}^{P} x_{sp}\beta_p \right) \right]. \tag{28}$$

We divided the complete dataset into five non-intersecting subsets and applied the 5-fold cross-validation technique to train both MTL-Cox and competing methods. Specifically, we used one subset for validation and the other four subsets for training. To assess the performance of MTL-Cox against other single-task learning approaches, we conducted a paired-sample Wilcoxon signed-rank test to quantitatively evaluate the concordance index, AUC, specificity, sensitivity, and Youden index.

The MTL-Cox model has three hyper-parameters: the number of searching parameters, the rate of the smallest search parameter compared to the largest one, and the scale of the first searching point; these were set as 50, 0.01, and 1, respectively. We used the personalized disease evaluation system proposed in this paper using the MTL-Cox model to achieve personalized disease risk prediction. In our experiment, we implemented the MTL-Cox model via Matlab and the personalized prediction in R (The code can be accessed through the URL https://github.com/Shuaijiea/paper-code.).

Results

The results obtained from the UK Biobank dataset are presented below, while those obtained from the Weihai physical examination dataset can be found in Supplementary file S1 File.



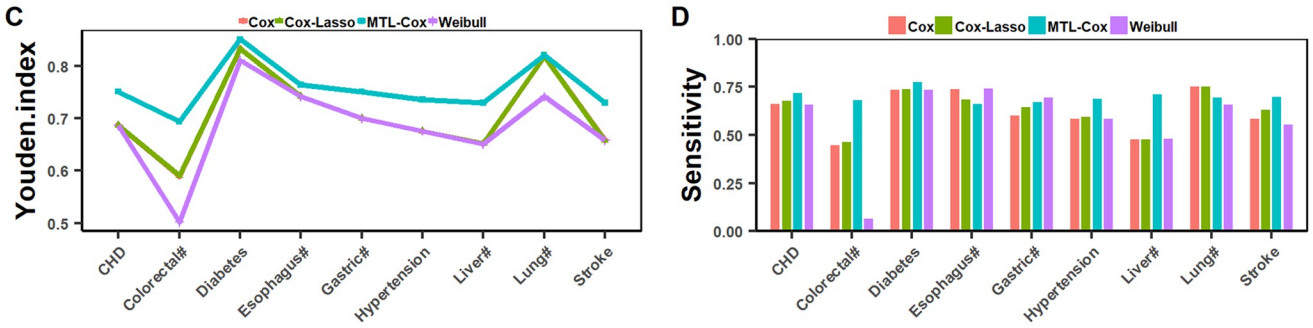
Disease	C-index				AUC			
	MTL-Cox	Cox	Weibull	Cox-Lasso	MTL-Cox	Cox	Weibull	Cox-Lasso
CHD	**0.737**	0.678	0.678	0.678	**0.751**	0.687	0.687	0.687
Colorectal cancer	**0.679**	0.586	0.557	0.585	**0.694**	0.590	0.502	0.590
Diabetes	**0.831**	0.811	0.811	0.811	**0.851**	0.833	0.811	0.833
Esophagus cancer	**0.752**	0.732	0.732	0.732	**0.764**	0.742	0.742	0.742
Gastric cancer	**0.731**	0.684	0.685	0.684	**0.751**	0.700	0.701	0.700
Hypertension	**0.718**	0.658	0.659	0.658	**0.736**	0.676	0.676	0.675
Liver cancer	**0.740**	0.671	0.671	0.671	**0.730**	0.651	0.651	0.652
Lung cancer	**0.822**	0.821	0.747	0.821	**0.821**	0.818	0.742	0.818
Stroke	**0.725**	0.652	0.653	0.652	**0.730**	0.658	0.658	0.658

Fig 8. Performances of MTL-Cox and competing methods were evaluated using C-index and AUC. Notes: '#' stands for the word "cancer", which facilitates the layout of the x-axis labels. '*' denotes $p < 0.05$, '**' denotes $p < 0.01$, and '***' denotes $p < 0.001$.

https://doi.org/10.1371/journal.pcbi.1011396.g008

Experimental results are depicted in Figs 8, 9 and 10, with further details recorded in Supplementary file S14 Table. Our MTL-Cox model outperformed competing methods, demonstrating superior performance in terms of C-index and AUC ($p < 0.01$), as well as sensitivity ($p < 0.05$). Although there was no statistically significant difference based on specificity, it is worth noting that a trade-off exists between sensitivity (missed rates) and specificity (misdiagnosis rates). Therefore, we used the Youden index as a measure to balance sensitivity and specificity when evaluating the overall diagnostic test performance. The results presented in Fig 9 indicate that our MTL-Cox model outperforms competing methods in terms of Youden index ($p < 0.01$).

As MTL-Cox incorporates the Cox model, it is possible to determine the absolute risk of individual developing diseases using Eq (6). The probability of each subject developing nine chronic diseases in the 1st, 3rd, 5th, and 7th year was calculated and presented in Supplementary file S4, S5 and S6 Tables. For example, using the 5th year, we ranked the absolute risk of the nine chronic diseases for each subject, which is shown in Supplementary file S13 Table and visualized in Fig 11. However, solely relying on absolute risk to judge a subject's risk level is not enough. For instance, in Fig 12A, the subject's risk of coronary heart disease is approximately 10.0%, while the risk of lung cancer is 2.5% as per the model. Although the numerical risk level of coronary heart disease is higher than that of lung cancer, the risk of coronary heart disease is still lower than the population's baseline risk, and it does not indicate the risk of a

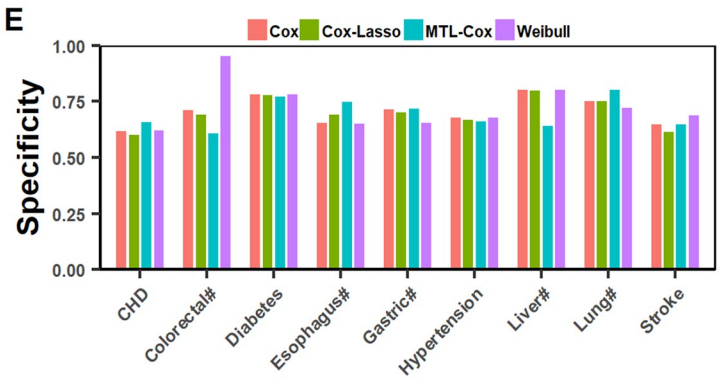


Disease	Youden index				Sensitivity			
	MTL-Cox	Cox	Weibull	Cox-Lasso	MTL-Cox	Cox	Weibull	Cox-Lasso
CHD	**0.374**	0.277	0.279	0.278	**0.717**	0.660	0.657	0.679
Colorectal cancer	**0.290**	0.157	0.018	0.153	**0.682**	0.446	0.065	0.463
Diabetes	**0.547**	0.517	0.517	0.516	**0.776**	0.734	0.734	0.737
Esophagus cancer	**0.409**	0.393	0.394	0.375	0.660	0.739	**0.743**	0.685
Gastric cancer	**0.388**	0.316	0.347	0.345	0.670	0.602	**0.694**	0.644
Hypertension	**0.348**	0.261	0.261	0.261	**0.688**	0.583	0.582	0.593
Liver cancer	**0.353**	0.279	0.281	0.272	**0.712**	0.476	0.481	0.475
Lung cancer	0.498	**0.502**	0.380	**0.502**	0.696	**0.752**	0.659	**0.752**
Stroke	**0.348**	0.234	0.241	0.243	**0.699**	0.585	0.554	0.630

Fig 9. Performances of MTL-Cox and competing methods were evaluated using Youden index and Sensitivity. Notes: '#' stands for the word "cancer", which facilitates the layout of the x-axis labels. '*' denotes $p < 0.05$, '**' denotes $p < 0.01$, and '***' denotes $p < 0.001$.

https://doi.org/10.1371/journal.pcbi.1011396.g009

high-risk individual with coronary heart disease. On the other hand, the risk of lung cancer is higher than the population's baseline risk and refers to the risk of a high-risk individual with lung cancer. Therefore, mapping the absolute risk to the population's baseline risk is necessary to better measure the relative absolute risk and excess absolute risk of each subject.



Disease	Specificity			
	MTL-Cox	Cox	Weibull	Cox-Lasso
CHD	0.657	0.617	0.622	0.599
Colorectal cancer	0.608	0.711	0.953	0.690
Diabetes	0.771	0.783	0.783	0.779
Esophagus cancer	0.749	0.654	0.651	0.691
Gastric cancer	0.718	0.714	0.653	0.700
Hypertension	0.660	0.678	0.679	0.669
Liver cancer	0.641	0.803	0.800	0.797
Lung cancer	0.802	0.750	0.721	0.750
Stroke	0.649	0.649	0.687	0.613

Fig 10. Performances of MTL-Cox and competing methods were evaluated using Specificity. Notes: '#' stands for the word "cancer", which facilitates the layout of the x-axis labels. '.' denotes $p > 0.05$, '*' denotes $p < 0.05$, '**' denotes $p < 0.01$, and '***' denotes $p < 0.001$.

https://doi.org/10.1371/journal.pcbi.1011396.g010

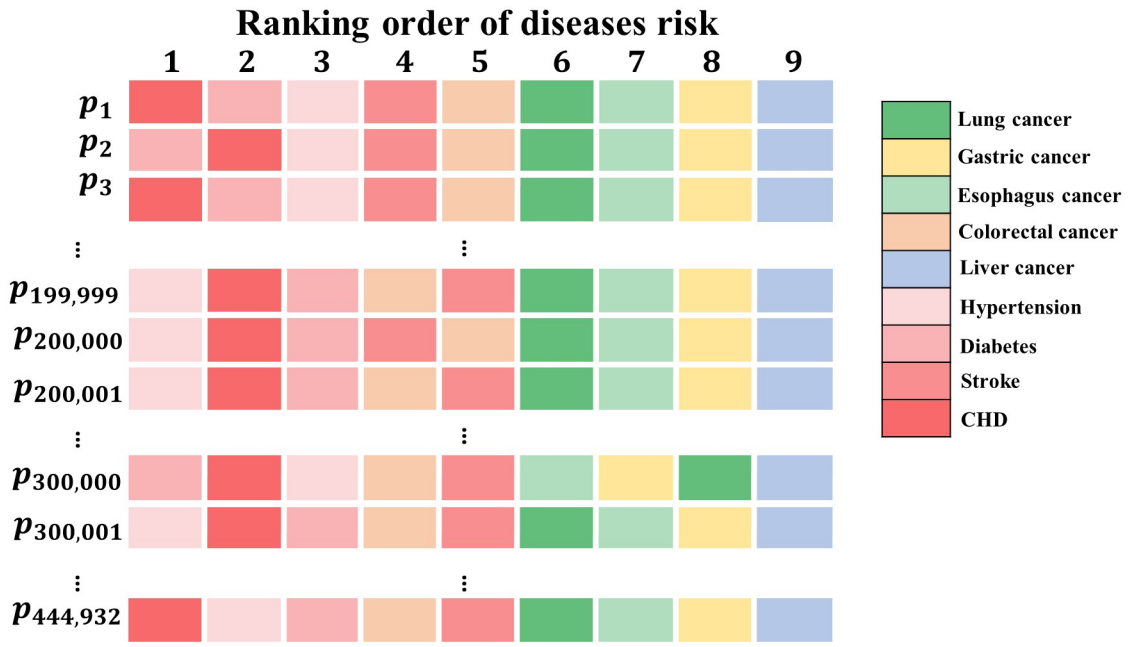


Fig 11. Personalized absolute risk ranking of nine chronic diseases in the fifth year. p_i denotes the *i–th* subjects.

https://doi.org/10.1371/journal.pcbi.1011396.g011

We initially categorized the population basis risk output from the MTL-Cox model into three categories: low-risk, average-risk, and high-risk. The absolute risk of each subject was then sorted into the 33%, 50%, and 66% quantiles to establish the upper bound value of the low-risk group, the median value of the average-risk group, and the lower value of the high-risk group. Next, we created population basis lines for low-risk, average-risk, and high-risk groups for nine chronic diseases, as shown in Figs 13, 14 and 15. Finally, we illustrated the risk of lung cancer for a sample subject after one year and three years. As depicted in Fig 12B, this subject's risk of developing lung cancer after one year and three years was greater than that of

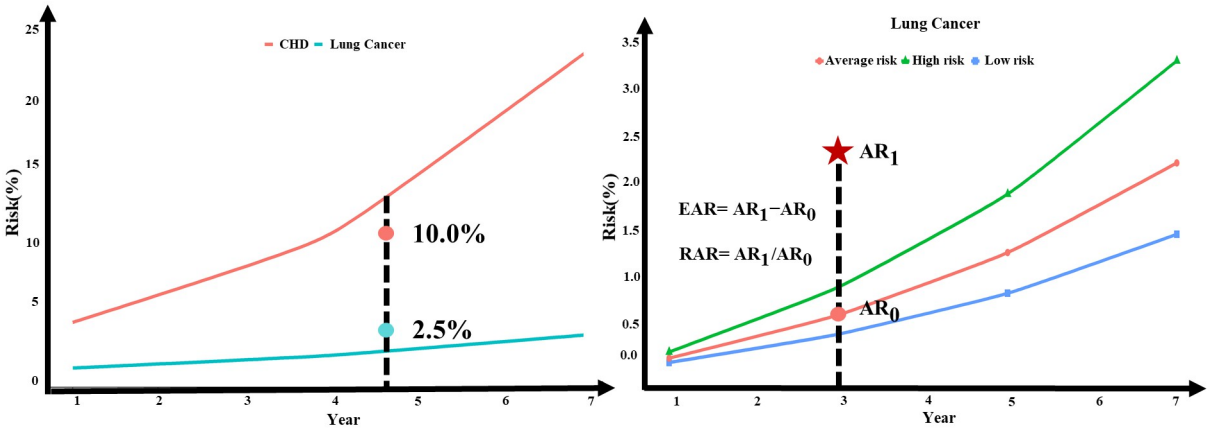


Fig 12. (A) Illustration of Absolute and Basis Risk for a Population: The red line in the figure on the left represents the average basis risk for coronary heart disease (CHD), while the green line in the figure on the left represents the average basis risk for lung cancer. The subject's absolute risk for CHD is approximately 10.0%, and their absolute risk for lung cancer is 2.5%. **(B) Personalized Risk Level Prediction Example**: In the third year, AR_1 represents the subject's absolute risk of lung cancer, while AR_0 is the basis average risk of lung cancer in the same year. The excess absolute risk is denoted by EAR, and the relative absolute risk is RAR.

https://doi.org/10.1371/journal.pcbi.1011396.g012

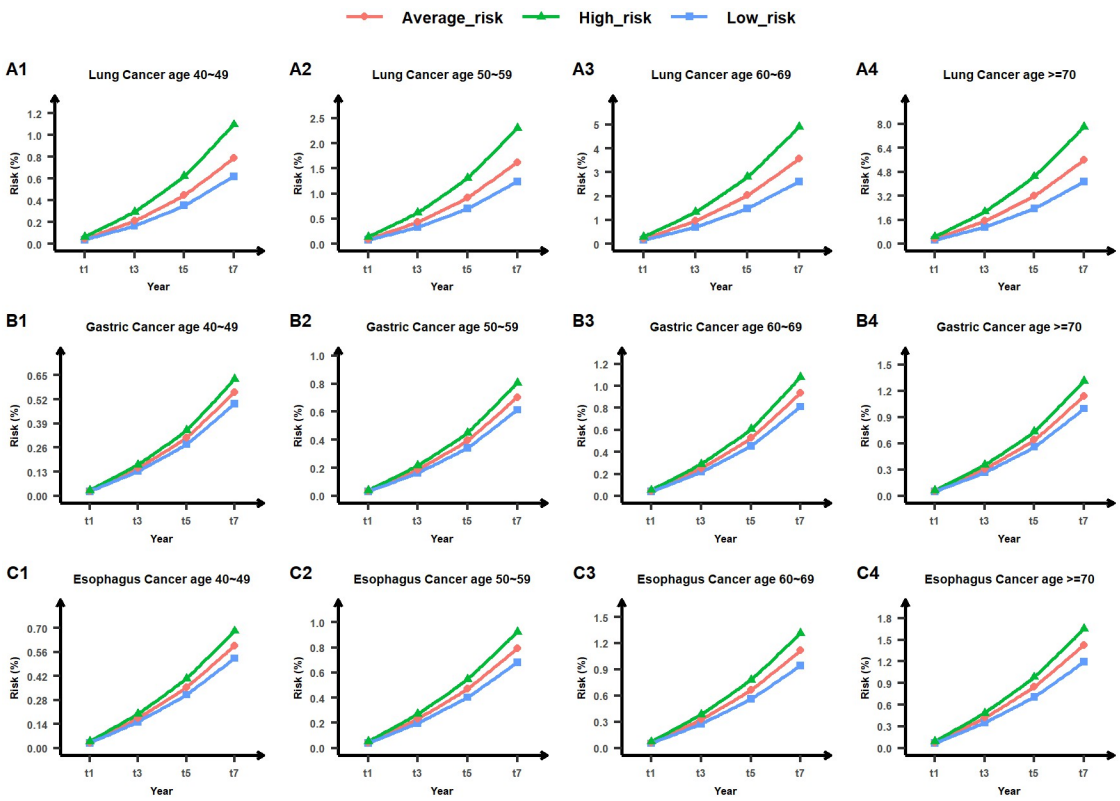


Fig 13. Low-risk, average-risk, and high-risk levels for lung cancer, gastric cancer, and esophagus cancer across four age groups (40 ~ 49, 50 ~ 59, 60 ~ 69, and ≥70). The horizontal axis denotes the timeline in years, while the vertical axis represents the risk value measured in percentage. The area below the blue line indicates low risk, the area between the blue and green lines denotes average risk, and the area above the green line signifies high risk. Comprehensive outcomes are provided in Supplementary file S3 Table.

https://doi.org/10.1371/journal.pcbi.1011396.g013

the high-risk population. Additionally, their risk of lung cancer after three years was 4.02 times higher than the population's average absolute risk, with an excess absolute risk of 1.79340% after three years. Therefore, this subject is considered a high-risk individual for lung cancer, and this finding can help remind both the subject and their physician to take appropriate treatment measures as soon as possible. Relative absolute risk for all subjects can be found in Supplementary files S7, S8 and S9 Tables. Additionally, the excess absolute risk for all subjects can be found in Supplementary files S10, S11 and S12 Tables.

Discussion

In the experimental results section, we present the evaluation of the MTL-Cox model from the perspective of five quantitative evaluation metrics. Our experimental results show that the MTL-Cox model outperforms other competing single-task learning methods in terms of concordance index, AUC, sensitivity, and the Youden index. However, there was no clear advantage in terms of specificity. Regarding the concordance index, we observe that MTL-Cox performed with the highest values in all diseases, indicating its superior precision. Additionally, the higher AUC value of MTL-Cox compared to single-task learning methods suggests its better performance in discriminating between diseased and healthy populations. We also observe that the MTL-Cox model outperforms other methods in terms of sensitivity, which is crucial in detecting disease in patients and reducing missed diagnoses. With regards to

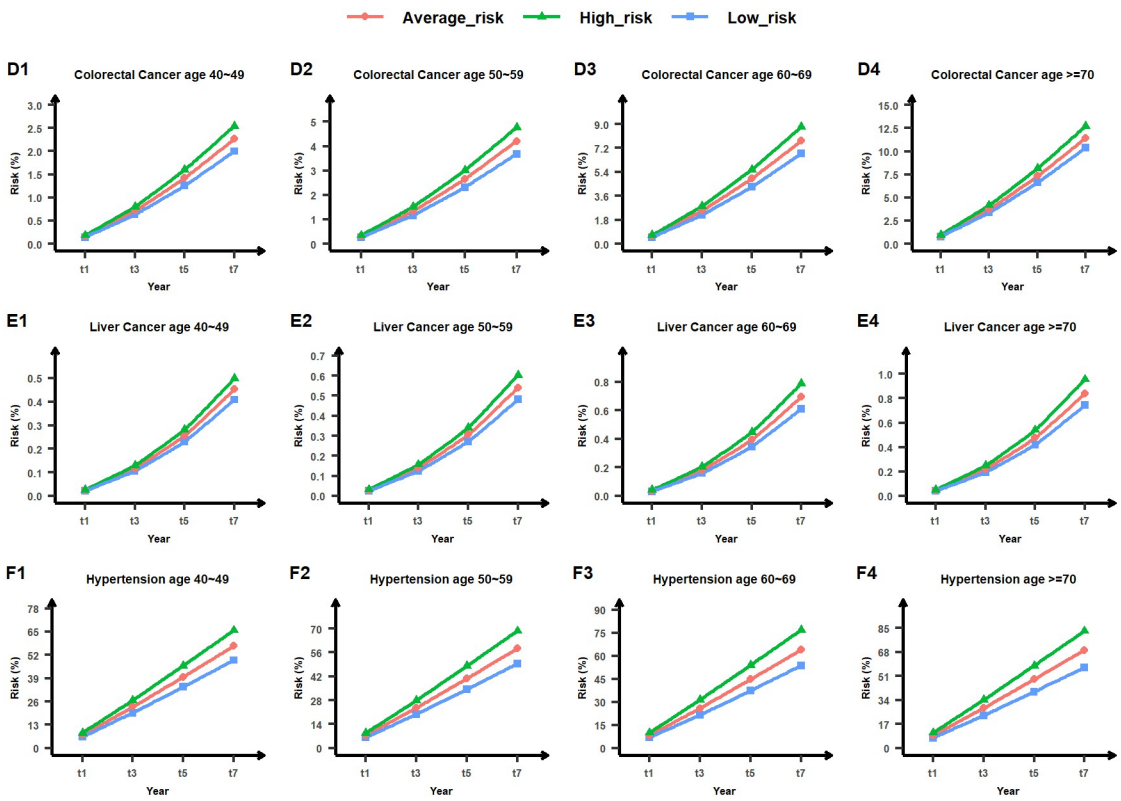


Fig 14. Low-risk, average-risk, and high-risk levels for colorectal cancer, liver cancer, and hypertension across four age groups (40 ~ 49, 50 ~ 59, 60 ~ 69, and ≥70). The horizontal axis denotes the timeline in years, while the vertical axis represents the risk value measured in percentage. The area below the blue line indicates low risk, the area between the blue and green lines denotes average risk, and the area above the green line signifies high risk. Comprehensive outcomes are provided in Supplementary file S3 Table.

https://doi.org/10.1371/journal.pcbi.1011396.g014

specificity, we noted that there was no significant difference between the MTL-Cox model and other single-task learning methods on the UK Biobank. This is because sensitivity and specificity are inversely related when the population prevalence remains constant [68]. However, the Youden index of the MTL-Cox model was significantly higher than competing methods, indicating its advantage in precision measurement when considering both the missed diagnosis rate and the misdiagnosis rate.

Moreover, we take gastric cancer prediction in the UK Biobank dataset as an example to demonstrate the accuracy of the MTL-Cox model. We predict the risk probability of developing gastric cancer in the fifth year and convert it into binary classification (0 for not developing gastric cancer, 1 for developing) using the cut-off value obtained from the AUC. Among those who developed gastric cancer, 108 individuals were predicted accurately by the MTL model while other competing methods failed to do so. This accurate prediction is attributed to the multitask learning framework employed in the MTL-Cox model. The multitask learning framework of the model allows for shared training information and knowledge, leading to an effective increase in sample size and improving the generalization ability of the model [27]. The advantages of multitask learning include data amplification, eavesdropping, attention focusing, representation bias, and overfitting prevention [27]. Data amplification refers to an effective increase in sample size due to extra information in the training signals of related tasks. Eavesdropping allows a task to learn features that are difficult for another task to learn.

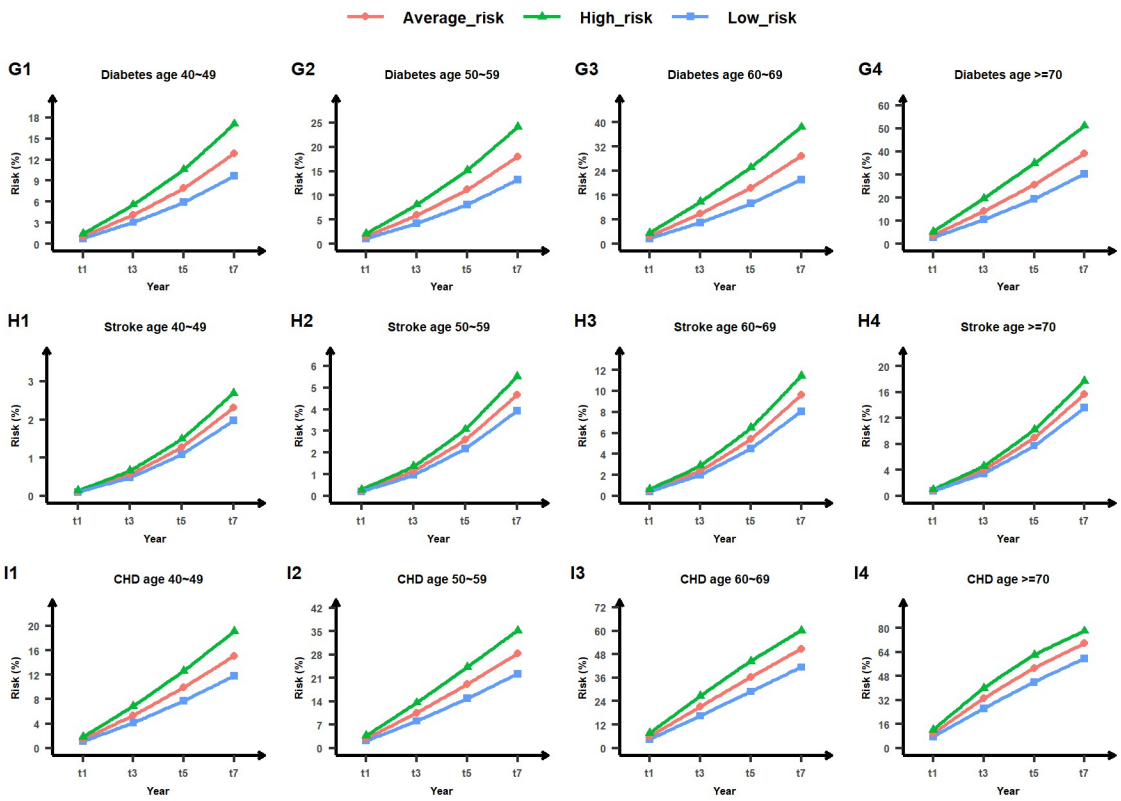


Fig 15. Low-risk, average-risk, and high-risk levels for diabetes, stroke, and coronary heart disease (CHD) across four age groups (40 ~ 49, 50 ~ 59, 60 ~ 69, and ≥70). The horizontal axis denotes the timeline in years, while the vertical axis represents the risk value measured in percentage. The area below the blue line indicates low risk, the area between the blue and green lines denotes average risk, and the area above the green line signifies high risk. Comprehensive outcomes are provided in Supplementary file S3 Table.

https://doi.org/10.1371/journal.pcbi.1011396.g015

Attention focusing helps the model distinguish between relevant and irrelevant features. Representation bias prefers to learn a class of models that also emphasizes other tasks, which helps the model demonstrate generalization to new tasks. Overfitting prevention is achieved by the shared module in multitask learning, which takes into account all tasks and avoids overfitting to a single task's training set. However, the specific mechanism of the superior performance of the multitask learning framework cannot be proved theoretically, and future work is needed to analyze the internal mechanism of the MTL-Cox model.

Furthermore, in terms of feature selection, our goal is to predict diseases using a precise and minimal feature set, considering both the clinical practicality and computational efficiency of the MTL-Cox model in high-dimensional data. To achieve this, we conduct feature selection based on a single disease, as each chronic disease has specific biological characteristics and pathological mechanisms. This approach helps us more accurately screen variables related to the target disease. However, considering the possibility of missing variables, we take into account the correlation between chronic diseases and use the union of all variables preliminarily screened from chronic diseases as the input for the MTL-Cox model. It is worth noting that, as described in the paper, we have incorporated a regularization term, namely the $L_{2,1}$ norm, into the MTL-Cox model. This regularization term takes into account the relationships between multiple chronic diseases for feature selection while conducting feature selection for multiple diseases, rather than selecting features based on a single disease as in single-task

learning. This also highlights the advantages of the MTL-Cox model in feature selection, as it can simultaneously consider the correlations and differences between multiple diseases, and find the optimal feature sets with predictive power for multiple diseases. However, this feature selection strategy is not perfect for high-dimensional data. Therefore, in future work, researchers can further improve the MTL-Cox model to enable rapid variable selection and prediction in high-dimensional data.

Conclusion

Currently, chronic disease prevention and control are limited, and the incidence and mortality rates of chronic diseases remain high and are expected to keep rising [69–71]. This is due to the unclear etiology and pathogenesis of most chronic diseases, as well as the poor treatment effect. Therefore, preventing the occurrence of chronic diseases is of practical significance. In this study, we designed a risk prediction model for nine chronic diseases based on the MTL-Cox model using UK Biobank data. We also validated the MTL-Cox model's effectiveness using the Weihai physical examination database. The MTL-Cox model can handle survival data with the right censoring by embedding the Cox proportional hazards model and consider the task-relatedness of chronic diseases by using multitask learning. Furthermore, we applied five performance metrics to evaluate the survival analysis model: C-index, AUC, specificity, sensitivity, and Youden index, to measure the model's prediction performance. Compared with other models, the MTL-Cox model leveraged more data from different learning tasks than single-task learning, leading to better knowledge sharing between tasks, better performance of each task, and a low risk of overfitting for each task [28]. As a result, the MTL-Cox model was found to be a reliable early diagnosis method for clinical medicine. Additionally, we calculated the absolute risk of subjects from different years based on the MTL-Cox model and ranked the risk. Subsequently, the risks of the nine chronic diseases in the UK Biobank were graded to achieve hierarchical early warning. Finally, to accurately quantify the risk of subjects, we mapped individuals' absolute risk to the population basis risk, which reflected the relative absolute risk and excess absolute risk. The MTL-Cox model can guide the lives of individuals based on input features, especially in high-risk groups, thereby reducing the risk of chronic diseases. Overall, this study contributes to the prevention and early detection of chronic diseases, which is of great importance for public health.

Future work

There are several limitations to this study that will be addressed in future research. Firstly, the age range of the UK Biobank dataset used in this paper was limited to individuals aged between 37 to 73 years, which may have introduced selection bias as chronic diseases are more prevalent in the elderly population (https://www.cdc.gov/nchs/health_policy/adult_chronic_conditions.htm. Accessed on Spe. 17^{th}, 2022.). Secondly, future studies will consider incorporating multimodal data related to chronic diseases to improve the generalizability of the model. Thirdly, developing feature selection methods of high-dimensional data for the MTL-Cox model will be a subject of future work.

Supporting information

S1 Table. Details of demographic and clinical features. The details of demographic and clinical features, including the names of features, features type, and the description of features. (XLSX)

S2 Table. Baseline information for nine chronic diseases. The information for the nine chronic diseases we studied on the UK Biobank dataset.
(XLSX)

S3 Table. Low risk, average risk, and high risk of nine chronic diseases. The cut-off level of low risk, average risk, and high risk for nine chronic diseases.
(XLSX)

S4 Table. Personalized predicted probabilities of lung cancer, gastric cancer, and esophagus cancer_Absolute Risk. The absolute risk of each subject developing lung cancer, gastric cancer, and esophagus cancer in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively.
(XLSX)

S5 Table. Personalized predicted probabilities of colorectal cancer, liver cancer, and hypertension_Absolute Risk. The absolute risk of each subject developing colorectal cancer, liver cancer, and hypertension in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively.
(XLSX)

S6 Table. Personalized predicted probabilities of diabetes, stroke, and CHD_Absolute Risk. The absolute risk of each subject developing diabetes, stroke, and CHD in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively.
(XLSX)

S7 Table. Personalized predicted probabilities of lung cancer, gastric cancer, and esophagus cancer_Relative Absolute Risk. The relative absolute risk of each subject developing lung cancer, gastric cancer, and esophagus cancer in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively, which reflects that the absolute risk of each individual is a multiple of the average absolute risk of the population.
(XLSX)

S8 Table. Personalized predicted probabilities of colorectal cancer, liver cancer, and hypertension_Relative Absolute Risk. The relative absolute risk of each subject developing colorectal cancer, liver cancer, and hypertension in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively, which reflects that the absolute risk of each individual is a multiple of the average absolute risk of the population.
(XLSX)

S9 Table. Personalized predicted probabilities of diabetes, stroke, and CHD_Relative Absolute Risk. The relative absolute risk of each subject developing diabetes, stroke, and CHD in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively, which reflects that the absolute risk of each individual is a multiple of the average absolute risk of the population.
(XLSX)

S10 Table. Personalized predicted probabilities of lung cancer, gastric cancer, and esophagus cancer_Excess Absolute Risk. The excess absolute risk of each subject developing lung cancer, gastric cancer, and esophagus cancer in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively, which reflects the difference between the absolute risk of each individual and the average absolute risk of everyone.
(XLSX)

S11 Table. Personalized predicted probabilities of colorectal cancer, liver cancer, and hypertension_Excess Absolute Risk. The excess absolute risk of each subject developing colorectal cancer, liver cancer, and hypertension in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively,

which reflects the difference between the absolute risk of each individual and the average absolute risk of everyone.
(XLSX)

S12 Table. Personalized predicted probabilities of diabetes, stroke, and CHD_Excess Absolute Risk. The excess absolute risk of each subject developing diabetes, stroke, and CHD in the 1-*th*, 3-*th*, 5-*th*, and 7-*th* year, respectively, which reflects the difference between the absolute risk of each individual and the average absolute risk of everyone.
(XLSX)

S13 Table. Absolute risk ranking of nine chronic diseases in the fifth year for each subject. Taking the fifth year as an example, the absolute risk of nine chronic diseases per subject was ranked. The values in the table are absolute risk, and the darker the color, the higher the risk.
(XLSX)

S14 Table. Five evaluation metrics for four models. The experimental results between different methods are reported in this table.
(XLSX)

S15 Table. Categorical feature coding details. The details of categorical features coding.
(XLSX)

S16 Table. The features details of the Weihai physical examination dataset. The details of features on the Weihai physical examination dataset, including the names of features, the description of features, and features initially selected for each target disease.
(XLSX)

S17 Table. Baseline information for five cancers of the Weihai physical examination dataset. The information for the nine chronic diseases we studied on the Weihai physical examination dataset.
(XLSX)

S1 File. Validating the MTL-Cox model framework in the Weihai physical examination dataset. The details of materials and experiment results are based on the Weihai physical examination dataset.
(DOCX)

Acknowledgments

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: Shuaijie Zhang, Fan Yang, Fuzhong Xue.

Data curation: Shuaijie Zhang, Lijie Wang, Shucheng Si, Jianmei Zhang.

Methodology: Shuaijie Zhang, Fan Yang.

Software: Shuaijie Zhang.

Visualization: Shuaijie Zhang.

Writing – original draft: Shuaijie Zhang, Fan Yang.

Writing – review & editing: Shuaijie Zhang, Fan Yang, Fuzhong Xue.

References

1. Beaglehole R, Ebrahim S, Reddy S, Voûte J, Leeder S, et al. Prevention of chronic diseases: a call to action. The Lancet. 2007; 370(9605):2152–2157. https://doi.org/10.1016/S0140-6736(07)61700-0 PMID: 18063026
2. Abegunde DO, Mathers CD, Adam T, Ortegon M, Strong K. The burden and costs of chronic diseases in low-income and middle-income countries. The Lancet. 2007; 370(9603):1929–1938. https://doi.org/10.1016/S0140-6736(07)61696-1 PMID: 18063029
3. Unwin N, Alberti K. Chronic non-communicable diseases. Annals of Tropical Medicine & Parasitology. 2006; 100(5-6):455–464. https://doi.org/10.1179/136485906X97453 PMID: 16899148
4. Field AE, Coakley EH, Must A, Spadano JL, Laird N, Dietz WH, et al. Impact of overweight on the risk of developing common chronic diseases during a 10-year period. Archives of internal medicine. 2001; 161 (13):1581–1586. https://doi.org/10.1001/archinte.161.13.1581 PMID: 11434789
5. Booth FW, Roberts CK, Laye MJ. Lack of exercise is a major cause of chronic diseases. Comprehensive physiology. 2012; 2(2):1143. https://doi.org/10.1002/cphy.c110025 PMID: 23798298
6. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021; 71(3):209–249. PMID: 33538338
7. V Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Medical Informatics Assoc. 2019; 26(12):1651–1654. https://doi.org/10.1093/jamia/ocz130 PMID: 31373357
8. Jayanthi P. Machine learning and deep learning algorithms in disease prediction. Deep Learning for Medical Applications with Unique Data. 2022; p. 123–152. https://doi.org/10.1016/B978-0-12-824145-5.00009-5
9. Jothi KA, Subburam S, Umadevi V, Hemavathy K. Heart disease prediction system using machine learning. Materials Today: Proceedings. 2021.
10. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology. 2013; 66(4):398–407. https://doi.org/10.1016/j.jclinepi.2012.11.008 PMID: 23384592
11. Surya BS, Singh NK, Rekha SS. Implementation of Liver Disease Prediction Using Machine Learning. Technoscience Academy. 2021;(2).
12. Symum H, Zayas-Castro JL. Prediction of Chronic Disease-Related Inpatient Prolonged Length of Stay Using Machine Learning Algorithms. Healthcare Informatics Research. 2020; 26(1):20. https://doi.org/10.4258/hir.2020.26.1.20 PMID: 32082697
13. Royston P. Visualizing and analyzing time to event data: Lifting the veil of censoring. In: United Kingdom Stata Users Group Meetings; 2006.
14. Lee Elisa T, M M, et al. A computer program for comparing K samples with right-censored data. Computer Programs in Biomedicine. 1972.
15. Moore KL, Laan MJVD. RCTs with Time-to-Event Outcomes. Springer New York. 2011.
16. Barrajón E, Barrajón L. Effect of right censoring bias on survival analysis; 2020.
17. Lenz G. Chronic diseases. Encyclopedia of Public Health; 2008.
18. Walter F M. Lay Understanding of Familial Risk of Common Chronic Diseases: A Systematic Review and Synthesis of Qualitative Research. The Annals of Family Medicine. 2004; 2(6):583–594. https://doi.org/10.1370/afm.242 PMID: 15576545
19. Wu F, Guo Y, Chatterji S, Zheng Y, Naidoo N, Jiang Y, et al. Common risk factors for chronic non-communicable diseases among older adults in China, Ghana, Mexico, India, Russia and South Africa: the study on global AGEing and adult health (SAGE) wave 1. BMC public health. 2015; 15(1):1–13. https://doi.org/10.1186/s12889-015-1407-0
20. Shi Y, Zhang J, Huang Y. Prediction of cardiovascular risk in patients with chronic obstructive pulmonary disease: a study of the National Health and Nutrition Examination Survey database. BMC Cardiovascular Disorders. 2021; 21(1):1–10. https://doi.org/10.1186/s12872-021-02225-w PMID: 34470611
21. Nenova Z, Shang J. Chronic disease progression prediction: Leveraging case-based reasoning and big data analytics. Production and Operations Management. 2022; 31(1):259–280. https://doi.org/10.1111/poms.13532
22. Yun J, Cho YH, Lee SM, Hwang J, Lee JS, Oh YM, et al. Deep radiomics-based survival prediction in patients with chronic obstructive pulmonary disease. Scientific reports. 2021; 11(1):1–9. https://doi.org/10.1038/s41598-021-94535-4 PMID: 34312450

23. Chen J, Yang Z, Yuan Q, Xiong Dx, Guo Lq. Prediction models for pulmonary function during acute exacerbation of chronic obstructive pulmonary disease. Physiological Measurement. 2020; 41 (12):125010.
24. Adam MG, Beyer G, Christiansen N, Kamlage B, Pilarsky C, Distler M, et al. Identification and validation of a multivariable prediction model based on blood plasma and serum metabolomics for the distinction of chronic pancreatitis subjects from non-pancreas disease control subjects. Gut. 2021; 70(11):2150–2158. https://doi.org/10.1136/gutjnl-2020-320723 PMID: 33541865
25. Yildirim P. Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). vol. 2. IEEE; 2017. p. 193–198.
26. Lee HA, Kim SU, Seo YS, Lee YS, Kang SH, Jung YK, et al. Prediction of the varices needing treatment with non-invasive tests in patients with compensated advanced chronic liver disease. Liver International. 2019; 39(6):1071–1079. https://doi.org/10.1111/liv.14036 PMID: 30589490
27. Caruana R. Multitask learning. Machine learning. 1997; 28(1):41–75. https://doi.org/10.1023/A:1007379606734
28. Zhang Y, Yang Q. A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering. 2021.
29. Jager KJ, Van Dijk PC, Zoccali C, Dekker FW. The analysis of survival data: the Kaplan–Meier method. Kidney international. 2008; 74(5):560–565. https://doi.org/10.1038/ki.2008.217 PMID: 18596735
30. Cutler SJ, Ederer F. Maximum utilization of the life table method in analyzing survival. Journal of chronic diseases. 1958; 8(6):699–712. https://doi.org/10.1016/0021-9681(58)90126-7 PMID: 13598782
31. Cox C, Chu H, Schneider MF, Munoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in medicine. 2007; 26(23):4352–4374. https://doi.org/10.1002/sim.2836 PMID: 17342754
32. Putzel P, Smyth P, Yu J, Zhong H. Dynamic Survival Analysis with Individualized Truncated Parametric Distributions. In: Greiner R, Kumar N, Gerds TA, van der Schaar M, editors. Proceedings of AAAI Symposium on Survival Prediction—Algorithms, Challenges and Applications, SPACA 2021, Stanford University, Palo Alto, CA, USA, March 22-24, 2021. vol. 146 of Proceedings of Machine Learning Research. PMLR; 2021. p. 159–170. Available from: https://proceedings.mlr.press/v146/putzel21a.html.
33. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. Journal of the American statistical Association. 1989; 84(408):1074–1078. https://doi.org/10.1080/01621459.1989.10478874
34. Jiang Y, Xie J, Han Z, Liu W, Xi S, Huang L, et al. Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit. Clinical Cancer Research. 2018; 24(22):5574–5584. https://doi.org/10.1158/1078-0432.CCR-18-0848 PMID: 30042208
35. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research. 2018; 24(6):1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853 PMID: 28982688
36. Bolourani S, Tayebi MA, Diao L, Wang P, Patel V, Manetta F, et al. Using machine learning to predict early readmission following esophagectomy. The Journal of Thoracic and Cardiovascular Surgery. 2021; 161(6):1926–1939. https://doi.org/10.1016/j.jtcvs.2020.04.172 PMID: 32711985
37. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Medical Informatics and Decision Making. 2020; 20 (11):1–7. https://doi.org/10.1186/s12911-020-01318-4 PMID: 33380338
38. Kangi AK, Bahrampour A. Predicting the survival of gastric cancer patients using artificial and bayesian neural networks. Asian Pacific journal of cancer prevention: APJCP. 2018; 19(2):487.
39. Diao X, Huo Y, Yan Z, Wang H, Yuan J, Wang Y, et al. An application of machine learning to etiological diagnosis of secondary hypertension: retrospective study using electronic medical records. JMIR Medical Informatics. 2021; 9(1):e19739. https://doi.org/10.2196/19739 PMID: 33492233
40. She Y, Jin Z, Wu J, Deng J, Zhang L, Su H, et al. Development and validation of a deep learning model for non–small cell lung cancer survival. JAMA network open. 2020; 3(6):e205842–e205842. https://doi.org/10.1001/jamanetworkopen.2020.5842 PMID: 32492161
41. Feng R, Cao Y, Liu X, Chen T, Chen J, Chen DZ, et al. ChroNet: A multi-task learning based approach for prediction of multiple chronic diseases. Multimedia Tools and Applications. 2021; p. 1–15.
42. Ampavathi A, Saradhi TV. Multi disease-prediction framework using hybrid deep learning: an optimal prediction model. Computer Methods in Biomechanics and Biomedical Engineering. 2021; 24 (10):1146–1168. https://doi.org/10.1080/10255842.2020.1869726 PMID: 33427480
43. Prinja S, Gupta N, Verma R. Censoring in clinical trials: review of survival analysis techniques. Indian journal of community medicine: offcial publication of Indian Association of Preventive & Social Medicine. 2010; 35(2):217. https://doi.org/10.4103/0970-0218.66859 PMID: 20922095

44. Liu H, Motoda H. Feature selection for knowledge discovery and data mining. vol. 454. Springer Science & Business Media; 2012.
45. Carneiro I. EBOOK: Introduction to Epidemiology. McGraw-Hill Education (UK); 2018.
46. Dietterich T. Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR). 1995; 27(3):326–327. https://doi.org/10.1145/212094.212114
47. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. vol. 2. Springer; 2009.
48. Wang L, Li Y, Zhou J, Zhu D, Ye J. Multi-task survival analysis. In: 2017 IEEE International Conference on Data Mining (ICDM). IEEE; 2017. p. 485–494.
49. Therneau TM, Grambsch PM. The cox model. In: Modeling survival data: extending the Cox model. Springer; 2000. p. 39–77.
50. Ruder S. An Overview of Multi-Task Learning in Deep Neural Networks. 2017;.
51. Ma J, Zhao Z, Yi X, Chen J, Hong L, Chi EH. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018. p. 1930–1939.
52. Obozinski G, Taskar B, Jordan M. Multi-task feature selection. Statistics Department, UC Berkeley, Tech Rep. 2006; 2(2.2):2.
53. Yang Y, Hospedales TM. Trace norm regularised deep multi-task learning. arXiv preprint arXiv:160604038. 2016;.
54. Sahinalp SC. Research in Computational Molecular Biology: 21st Annual International Conference, RECOMB 2017, Hong Kong, China, May 3-7, 2017, Proceedings. vol. 10229. Springer; 2017.
55. Ceci M, Hollmén J, Todorovski L, Vens C, Džeroski S. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II. vol. 10535. Springer; 2017.
56. Tanabe H, Fukuda EH, Yamashita N. Convergence rates analysis of a multiobjective proximal gradient method. Optimization Letters. 2022; p. 1–18.
57. Karimi H, Nutini J, Schmidt M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2016. p. 795–811.
58. Petermann-Rocha F, Lyall DM, Gray SR, Esteban-Cornejo I, Quinn TJ, Ho FK, et al. Associations between physical frailty and dementia incidence: a prospective study from UK Biobank. The Lancet Healthy Longevity. 2020; 1(2):e58–e68. https://doi.org/10.1016/S2666-7568(20)30007-6 PMID: 36094146
59. Palmer LJ. UK Biobank: bank on it. The Lancet. 2007; 369(9578):1980–1982. https://doi.org/10.1016/S0140-6736(07)60924-6
60. Collins R. What makes UK Biobank special? Lancet (London, England). 2012; 379(9822):1173–1174. https://doi.org/10.1016/S0140-6736(12)60404-8 PMID: 22463865
61. Organization WH, et al. WHO library cataloguing-in-publication data World Health Statistics 2010. World health. 2010; 3.
62. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982; 247(18):2543–2546. https://doi.org/10.1001/jama.1982.03320430047030 PMID: 7069920
63. Harrell F, Lee K, Califf R, Pryor D, Rosati R. Regression modelling strategies for improved prognostic prediction. Statistics in medicine. 1984; 3(2):143–152. https://doi.org/10.1002/sim.4780030207 PMID: 6463451
64. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. Controlled clinical trials. 2003; 24(6):682–701. https://doi.org/10.1016/S0197-2456(03)00072-2 PMID: 14662274
65. Zhang Z. Parametric regression model for survival data: Weibull regression model as an example. Annals of translational medicine. 2016; 4(24). https://doi.org/10.21037/atm.2016.08.45 PMID: 28149846
66. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine. 1997; 16 (4):385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4%3C385::AID-SIM380%3E3.0.CO;2-3 PMID: 9044528
67. Zhou D, Liu X, Wang X, Yan F, Wang P, Yan H, et al. A prognostic nomogram based on LASSO Cox regression in patients with alpha-fetoprotein-negative hepatocellular carcinoma following non-surgical therapy. BMC cancer. 2021; 21(1):1–16. https://doi.org/10.1186/s12885-021-07916-3 PMID: 33685417
68. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian pediatrics. 2011; 48(4):277–287. https://doi.org/10.1007/s13312-011-0055-4 PMID: 21532099

69. Petersen PE, Ogawa H. The global burden of periodontal disease: towards integration with chronic disease prevention and control. Periodontology 2000. 2012; 60(1):15–39. https://doi.org/10.1111/j.1600-0757.2011.00425.x PMID: 22909104

70. Vermunt NP, Harmsen M, Westert GP, Olde Rikkert MG, Faber MJ. Collaborative goal setting with elderly patients with chronic disease or multimorbidity: a systematic review. BMC geriatrics. 2017; 17 (1):1–12. https://doi.org/10.1186/s12877-017-0534-0 PMID: 28760149

71. Airhihenbuwa CO, Tseng TS, Sutton VD, Price L. Non–Peer Reviewed: Global Perspectives on Improving Chronic Disease Prevention and Management in Diverse Settings. Preventing chronic disease. 2021; 18. https://doi.org/10.5888/pcd18.210055