Supplementary Information for "UKB-MDRMF: A Multi-Disease Risk and Multimorbidity Framework Based on UK Biobank Data"

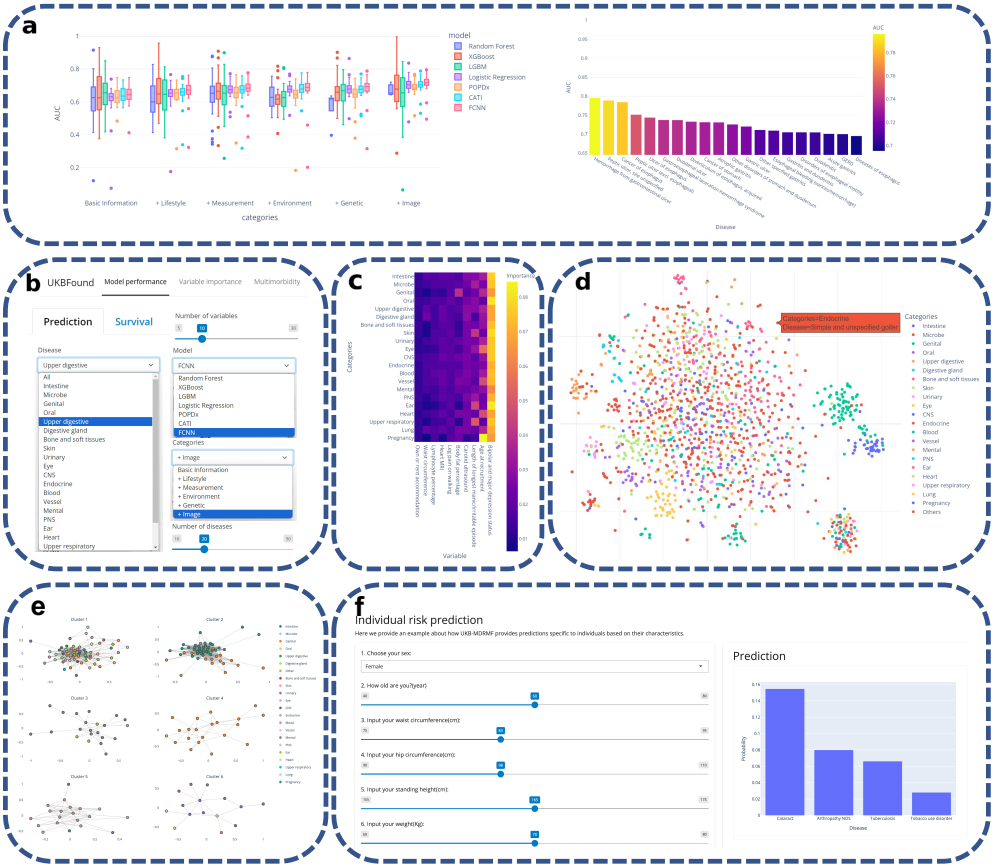
Yukang Jiang[1†], Bingxin Zhao[2†], Xiaopu Wang[3†], Borui Tang[4†], Huiyang Peng[3], Zidan Luo[3], Yue Shen[5], Zheng Wang[6], Zhiwen Jiang[4], Jie Wang[5], Jieping Ye[6*], Xueqin Wang[3*], Hongtu Zhu[4,7,8,9,10,11*]

[1]Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA.
[2]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, 19104, PA, USA.
[3]School of Management, University of Science and Technology of China, Hefei, 230026, AH, China.
[4]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA.
[5]CAS Key Laboratory of Technology in GIPAS, University of Science and Technology of China, Hefei, 230026, AH, China.
[6]Alibaba Group, Hangzhou, 310030, ZJ, China.
[7]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA.
[8]Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA.
[9]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA.
[10]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA.

*Corresponding author(s). E-mail(s): yejieping.ye@alibaba-inc.com; wangxq20@ustc.edu.cn; htzhu@email.unc.edu;
†These authors contributed equally to this work.

Supplementary Note 1: Platform of UKB-MDRMF

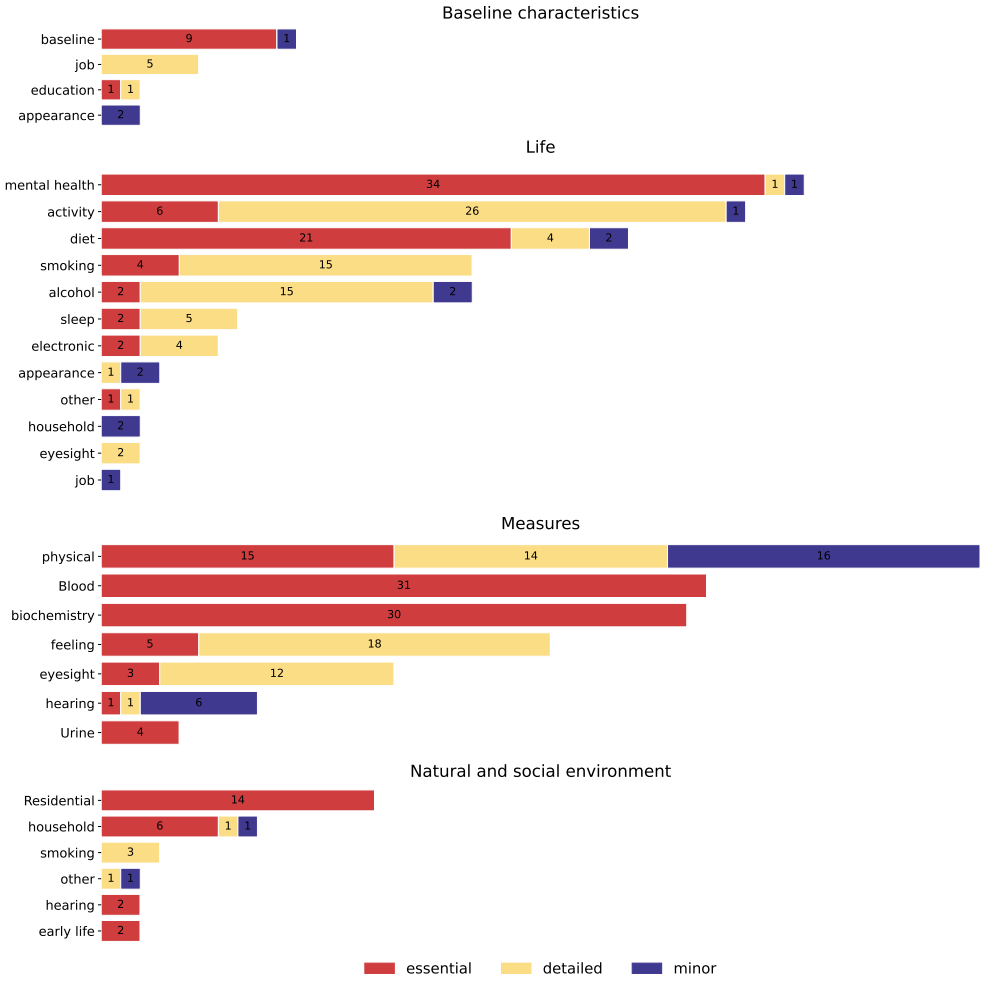
We developed an interactive platform (https://luminite.shinyapps.io/ukb-mdrmf/) to showcase detailed results of UKB-MDRMF, allowing exploration of disease predictions, variable importance, and comorbidities for specific risk factors and disease types. Platform functionalities are demonstrated in Supplementary Fig. 1.



Supplementary Fig. 1. Overview of the UKB-MDRMF interactive platform. (a). Disease prediction performance across different data categories using various models, evaluated by AUC. (b). User interface for selecting disease types, models, and variables for prediction and survival analysis. (c). Heatmap displaying variable importance for different disease types. (d). Visualization of comorbidity patterns across various disease types. (e). Interactive visualization of comorbidity network clusters. (f). Simplified online demo for individual disease prediction and risk assessment.

Supplementary Note 2: Data processing

2.1 Summary of selected variables



Supplementary Fig. 2. Summary of selected variables. The variables are divided into several subcategories and priorities. Generic and imaging data are omitted from the figure. Genetic data contain only two fields, and for imaging data, we use principal components extracted from 884 Brain MRI and 27 Heart MRI data fields.

2.2 Three subclasses of features

To enhance user accessibility and examine the varying impact of different features on predictions, we organized the data into three subclasses: “essential”, “detailed”,

Supplementary Table 1. Data fields and categories with essential information.

Category	Subcategory	Number of data fields
Basic Information	baseline	9
	education	1
Genetic	genetic	2
Life	activity	6
	alcohol	2
	diet	21
	electronic	2
	mental health	34
	other	1
	sleep	2
	smoking	4
Measures	biochemistry	30
	Blood	31
	eyesight	3
	feeling	5
	hearing	1
	physical	15
	Urine	4
Natural and social environment	early life	2
	hearing	2
	household	6
	Residential	14

and "minor", as shown in Supplementary Tables 1-3. These subclasses increase in detail with "essential" containing fewer but more important variables and "minor" encompassing a wider array of less critical variables. This flexible approach allows us to tailor feature selection for disease prediction and assessment to the specific needs of data collection, analysis, and disease characteristics. ("S3_allofus_data_dictionary.csv")

Supplementary Fig. 3 presents the model results for the three subclasses of features, which we define as different priority levels. We incrementally added information from various categories, using the best-performing models—FCNN for disease prediction and DeepSurv for risk assessment—for evaluation. Using only essential information yields reasonable predictive capability, and adding detailed information further improves performance. However, adding minor information does not significantly enhance model performance. Thus, in practical applications, variable selection should align with task requirements. If higher accuracy is desired, essential and detailed information should be used. If computational speed is prioritized over precision, using essential information alone is recommended.

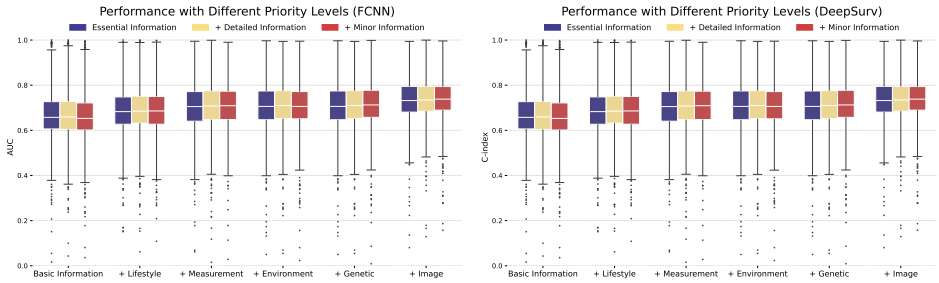
[1]For the imaging data, we use principal components instead of original data fields

Supplementary Table 2. Data fields and categories with detailed information.

Category	Subcategory	Number of data fields
Basic Information	education	1
	job	5
Life	activity	26
	alcohol	15
	appearance	1
	diet	4
	electronic	4
	eyesight	2
	mental health	1
	other	1
	sleep	5
	smoking	15
Measures	eyesight	12
	feeling	18
	hearing	1
	physical	14
Natural and social environment	household	1
	other	1
	smoking	3

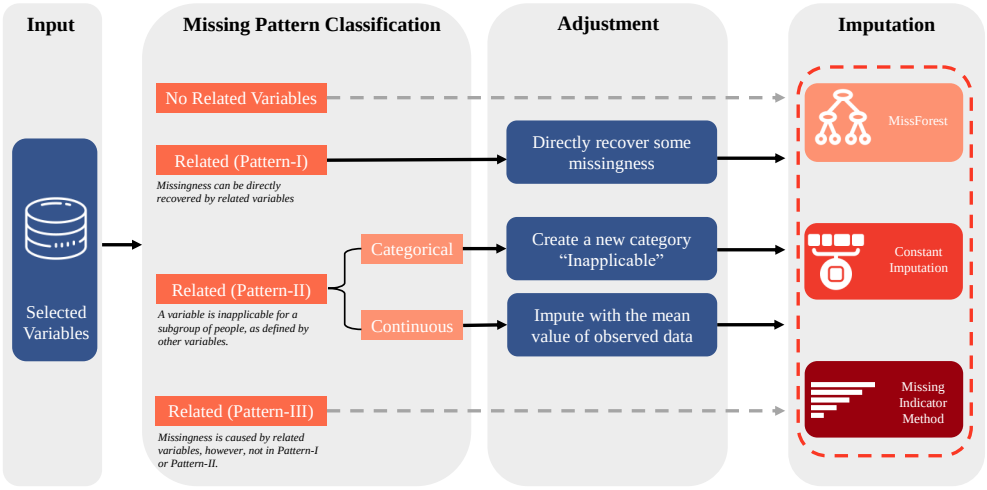
Supplementary Table 3. Data fields and categories with minor information.

Category	Subcategory	Number of data fields
Basic Information	appearance	2
	baseline	1
Life	activity	1
	alcohol	2
	appearance	2
	diet	2
	household	2
	job	1
Measures	hearing	6
	physical	16
Natural and social environment	household	1
	other	1
Imaging [1]	brain MRI	884
	heart MRI	27
	Carotid Ultrasound	12



Supplementary Fig. 3. Disease prediction and risk assessment based on three different priority levels of variables. The left panel shows the disease prediction results, presenting a comparison of AUC values on the testing set using a fully connected neural network (FCNN). The right panel shows the risk assessment results, presenting a comparison of C-index values on the testing set using the DeepSurv model. Box plots indicate the median (central line), interquartile range (box), and whiskers extending to the minimum and maximum values, excluding outliers—defined as data points lying beyond 1.5 times the interquartile range from the first and third quartiles.

2.3 Missing data processing



Supplementary Fig. 4. Pipeline for missing data processing. For input variables selected during preprocessing, we first classify their missing patterns into different types and then apply adjustments accordingly. Finally, we impute the remaining missing values and compare four classical imputation methods. Icons provided by Icons8 (https://icons8.com).

Supplementary Table 4. Summary of different sources of the response variable Y. The table includes the number of patients for different source-specific diseases, along with respective records and time ranges, both before and after time alignment considerations.

Data source	Time alignment	Number of patients	Number of records
Self-report (cancer)	No	47,641	51,564
	Yes		6,312
Self-report (non-cancer)	No	338,168	749,519
	Yes		54,918
Hospital inpatient	No	442,305	5,261,448
	Yes		4,032,539
Death registration	No	40,082	40,108
	Yes		40,096
Primary care	No	224,767	5,207,055
	Yes		2,009,735
First occurrence	No	222,903	2,385,577
	Yes		1,156,954

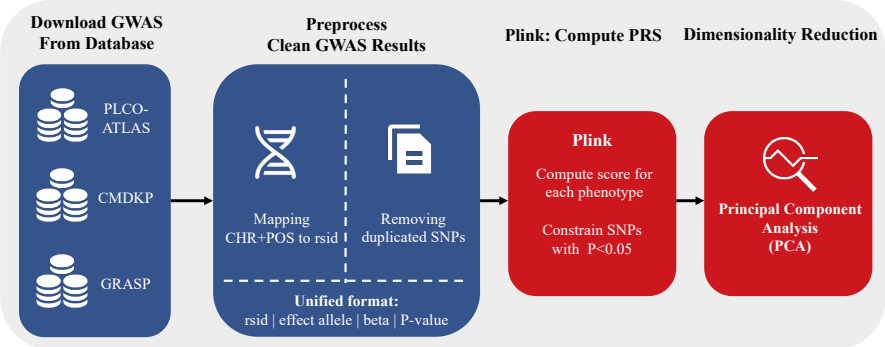
2.4 Statistics of response variables

Supplementary Table 4 presents statistics for different source response variables. The original data sources include self-report (Data Fields 20001, 20002), inpatient (Data Field 41270), death register (Data Field 40001), and primary care data (Data Field 42040), covering five subtypes. An additional subtype, first occurrences (Category 1712), provides valuable supplementary records for certain diseases. We analyzed response variable data for all records and after time alignment, indicating records where ICD codes occurred post-enrollment. Hospital inpatient data constitutes the largest data source, with most individuals having at least one inpatient record. Death register data is recorded post-enrollment.

2.5 Genetic data processing

To extract genetic variables, we used the 40 principal components of the genetic data from the UK Biobank, referred to as Data-Field 22009. We also considered incorporating Polygenic Risk Scores (PRS) [1]. To maintain the independence of the training, validation, and test sets, and to prevent information leakage, we used PRS weights derived from GWAS results of 111 phenotypes from PLCO-ATLAS [2], CMDKP [3], and GRASP [4], rather than UKB-specific GWAS weights. These phenotypes include various cancers (kidney, liver, lung, etc.), chronic diseases (type 1/2 diabetes, cardiovascular diseases, etc.), and biochemical indicators (BASO, EOS, HCT, etc.).

After preprocessing the GWAS results, we extracted SNPs with a p-value < 0.05 for each trait. Using PLINK, we computed the weighted sum of each SNP for each individual, with beta serving as the weight, to generate a PRS for each trait. We then merged the PRS for all traits and conducted a parallel analysis to determine the appropriate number of principal components, finding 27 to be optimal. The first 27 principal components were then calculated as the final result.



Supplementary Fig. 5. Processing pipeline for PRS from genetic data in the UK Biobank. Icons provided by Icons8 (https://icons8.com).

2.6 Imaging data processing

The UK Biobank includes data for over 500,000 individuals, but less than 50,000 have medical imaging data. Direct imputation of missing values for the entire data would require filling in over 90% of the data. Additionally, individuals with imaging data may also have partial missing data. The imaging data were not uniformly collected at enrollment but in multiple batches starting from 2014. This temporal misalignment with the enrollment date prevents the direct inclusion of imaging data alongside the other five types of covariates, as responses (Phecodes) must occur after covariate collection. Therefore, specialized handling and modeling of imaging data are required.

First, we applied the missing data imputation strategy described in Section **??** to the subset of individuals with imaging data, filling in missing values before extracting and summarizing the principal components of the imaging data. To prevent temporal leakage from the imaging data, we assigned zero values (no information imputation) to individuals without imaging data and adjusted the time for individuals with imaging data to their imaging collection time for comparison with disease occurrence time. For those without imaging data, we used the enrollment time to match disease occurrence time, ensuring disease occurrence post-covariate collection. To maintain the independence of training, validation, and test sets, all related PCA models were first trained on the training set to estimate weights. These weights were then used to calculate the principal components for the validation and test sets.

Supplementary Note 3: Details of prediction and survival Models

For disease prediction models, seven different models were implemented. Logistic Regression, POPDx, CATI, and FCNN are built using PyTorch, Random Forest with scikit-learn, XGBoost with xgboost, and LightGBM with lightgbm. Risk assessment models are also implemented using PyTorch.

For neural network-based models, we used the Adam optimizer. Each model's architecture was optimized using grid search to find the best performance.

For disease prediction models, we choose the model with best performance over a range of parameters. POPDx has four fully connected layers and an embedding matrix, followed by a multi-label output layer. The neuron counts in each of the four layers are 50, 25, 500, 100, 25, 100 for the six categories respectively. CATI's architecture consists of four layers. The neuron counts in each of the four layers are 25, 100, 25, 200, 500, 500 for the six categories respectively. FCNN has three layers with neuron counts of 200 or 100 (for + Measurment model only). ReLU activation functions are used between layers. Cross-entropy loss, suitable for multi-task prediction, is used. The learning rate is 0.0001, and training is conducted for 1000 epochs with early stopping if the validation loss does not decrease for 10 consecutive epochs.

For risk assessment models, POPDxSurv, CATISurv, and DeepSurv, follow a similar strategy. These models incorporate an embedding matrix and consist of five layers with 400 neurons each for CATISurv and POPDxSurv and 300 neurons for DeepSurv. The activation function is SELU. The loss function is the negative log-likelihood loss, suitable for multi-task prediction. The learning rate is 0.005, and training is conducted for 1000 epochs with early stopping if the validation loss does not decrease for 15 consecutive epochs.

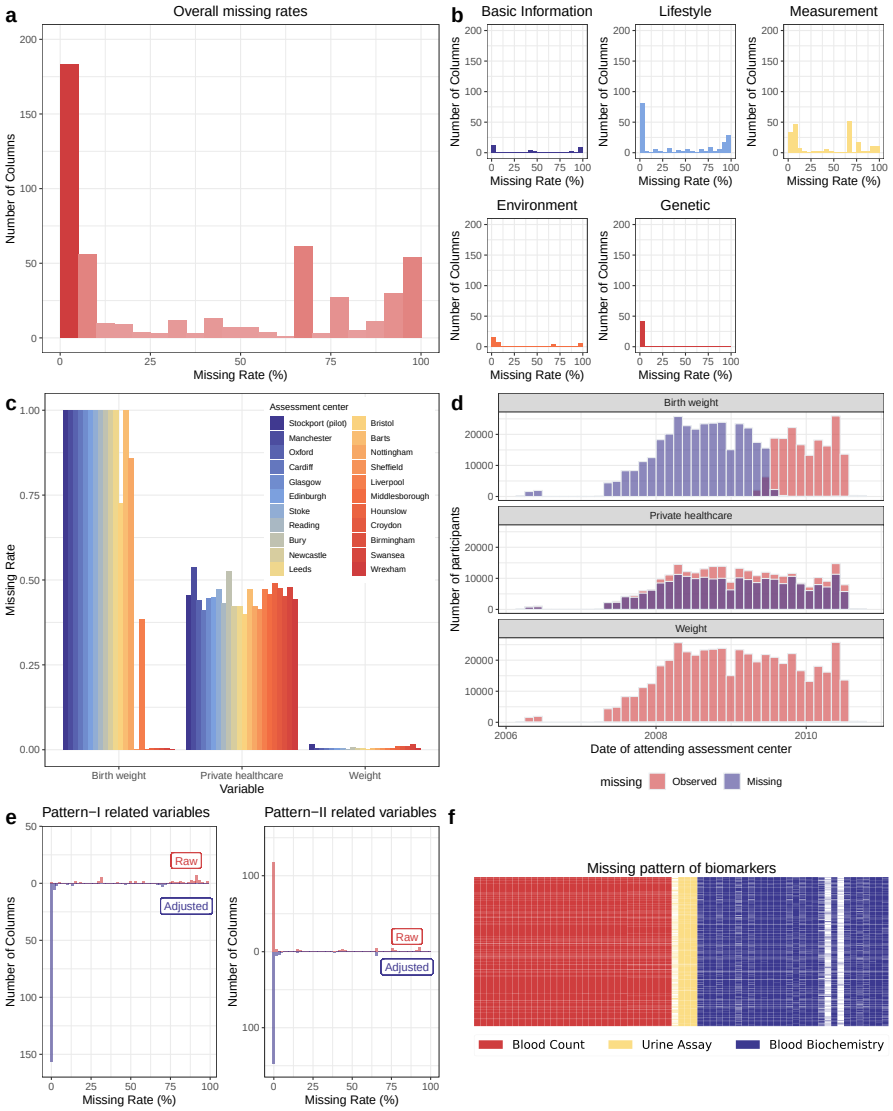
Supplementary Note 4: Detailed analysis results

4.1 Analysis of missing data

4.1.1 An overview of missing data in UK Biobank

To properly impute missing values in the UK Biobank data and facilitate the construction of the UKB-MDRMF framework, we conducted a systematic analysis of the missing data patterns. The patterns are summarized in Supplementary Fig. 6. Most selected variables have low rates of missing values, as shown in Supplementary Fig. 6(a). However, some variables have missing rates higher than 95%. Overall, the distribution is low in the middle and high on both sides. Supplementary Fig. 6(b) further categorizes missing values into six major groups: basic, lifestyle, measurement, environment, genetic, and imaging data. Imaging data, due to its different populations, is detailed in Supplementary Note 4.1.2. Genetic variables, derived from SNPs, generally have a missing rate lower than 10%.

High missing rates for some variables are partly due to differences in data collection across various centers and periods. The UK Biobank, a cohort study, collected data over an extended period from multiple centers, leading to discrepancies in data availability. Supplementary Fig. 6(c) illustrates three representative scenarios. For example, the "Private healthcare" variable shows significant missingness in certain centers, as indicated in Supplementary Fig. 6(d), since it was included post-2009. Conversely, the "Birth Weight" variable has a high missing rate unrelated to the center or timing of data collection, as about 40% of participants could not recall their birth weight. The "Weight" variable, however, has a low missing rate and is not significantly affected by data collection phases. More detailed analyses of missingness by time and center are presented in Supplementary Note 4.1.2.

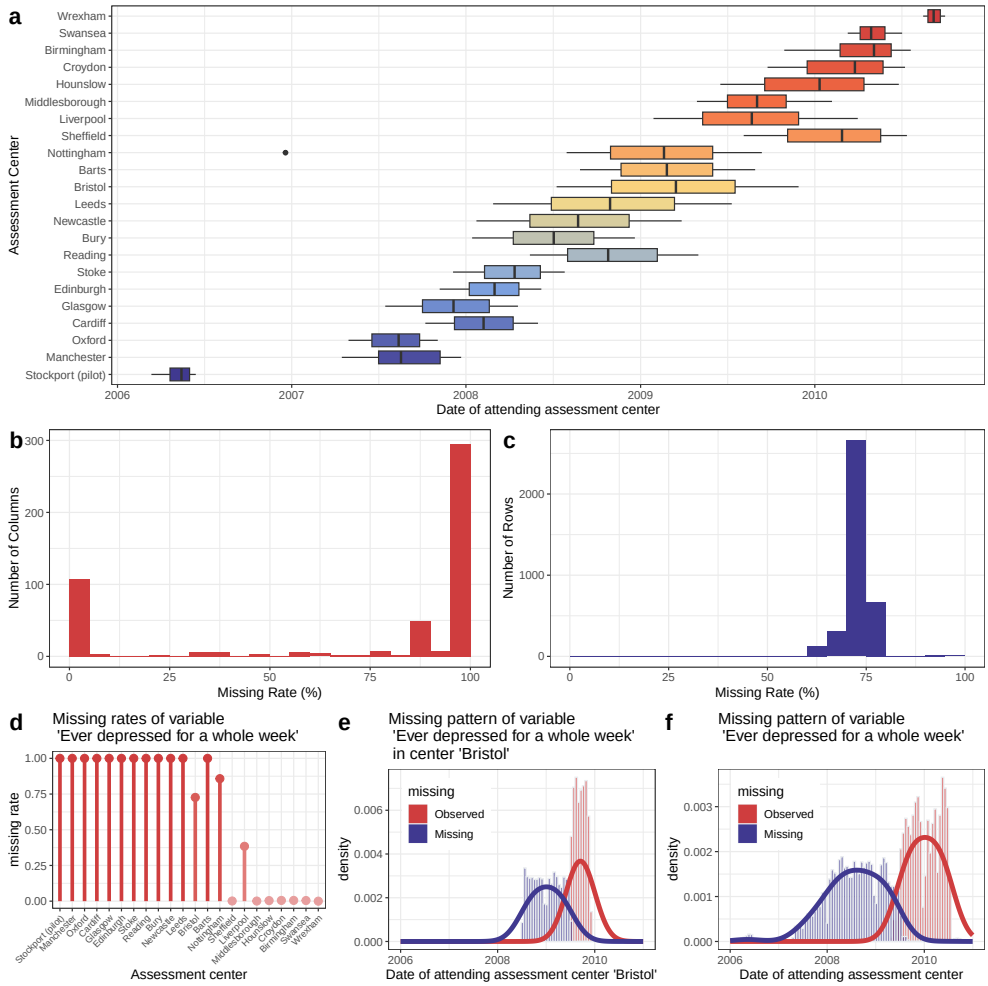


Supplementary Fig. 6. Missing patterns in the selected data, excluding imaging, which is reported elsewhere as only a sub-population has imaging measurements. (a). Overall missing rates of our selected variables. (b). Overall missing rates of selected variables in each category. (c). Typical examples of missingness across different assessment centers. The colored bars denote missing rates across different centers. (d). Typical examples of missingness across different time points. (e). Missing rates of Pattern-I and Pattern-II related variables before and after our manual adjustment. It illustrates that our adjustment leads to a significant decrease in missing rates of Pattern-I and Pattern-II related variables. (f). Missing patterns of different biomarkers. Each column represents a variable and each row represents a participant. A colored block means observed and a white block means missing. The white horizontal line denotes blockwise missingness, meaning that several variables are missing simultaneously for certain participants.

4.1.2 Missingness across different assessment centers

Further analysis of missingness related to assessment centers is presented in Supplementary Note 4.1.1. Supplementary Fig. 7(a) shows the distribution of attendance dates across different centers, indicating a stage-by-stage collection strategy. Notably, 3,967 participants attended the pilot assessment center. Before the initial data collection procedure, UK Biobank took a pilot phase, where thousands of people were recruited to the pilot assessment center. Data collected in this stage are used to improve the data collection procedure in the following phases. Only a subset of data fields are encompassed into the pilot phase, some of which are later revised or even deleted in formal collection. Thus, data quality for these participants is not guaranteed, see Supplementary Figs. 7(b) and 7(c) for an illustration. Both high missing rates for variables and participants are prevalent.

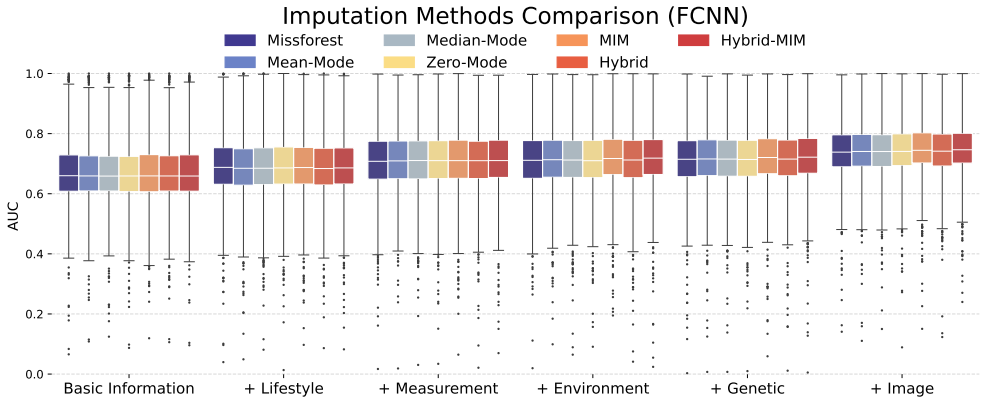
The Supplementary Figs. 7(d)-7(f) examine the variability in missingness across assessment centers and time intervals for the questionnaire item: "Looking back over your life, have you ever had a time when you were feeling depressed or down for at least a whole week?" (Data-Field 4598). Supplementary Fig. 7(d) shows the missing rates across distinct centers, while Supplementary Figs. 7(e)-7(f) detail the missing data patterns over different temporal segments. This field was bypassed in certain centers, such as "Stockport (pilot)" and "Manchester", and within other centers, it was limited to a subset of participants. For instance, in the "Bristol" center, its use began in July 2009, as evidenced in Supplementary Fig. 7(g), showing its progressive integration around that time.



Supplementary Fig. 7. Missingness across assessment centers and time points. (a). Distribution of attendance dates across different centers. (b). Missing rates for variables in the pilot center. (c). Missing rates for participants in the pilot center. (d). An example of missing patterns across different assessment centers. The x-axis presents the coding for different assessment centers, while the y-axis presents the proportion of individuals that have no records on the variable "Ever depressed for a whole week". (e). An example of missing patterns across different time periods. The y-axis represents the proportion of participants with the variable "Ever depressed for a whole week" observed or missing in the assessment center "Bristol" according to different colors, respectively. (f). Analogous to (e) for all assessment centers.

4.1.3 Comparison of different imputation methods

A comparison of different imputation methods is presented in Supplementary Fig. 8. Our method, labeled "Hybrid", uniquely combines additional adjustments with the MissForest method. We also incorporate the "Hybrid-MIM" method, which additionally adds missing indicators alongside the Hybrid method, to evaluate the predictive benefits of adding missing indicators. The FCNN method, showing the best overall performance, serves as a benchmark for comparing various imputation approaches. The six categories are sequentially incorporated into the model for performance evaluation, using AUC as the metric. While MIM and Hybrid-MIM generally exhibit slightly better performance in terms of median AUC, the performances of the various imputation methods are relatively similar. However, adding missing indicators for prediction complicates interpretability by introducing binary variables highly correlated with existing data variables, such as age and sex. Detailed illustrations are provided in Supplementary Note 4.1.5. In contrast, our Hybrid method employs careful adjustments, resulting in more meaningful imputed values and enhanced interpretability.



Supplementary Fig. 8. Performance for classification using FCNN under different imputation methods. The x-axis is the number of input categories (from basic to imaging data), while the y-axis is the AUC value for classification. Box plots depict the median (central line), interquartile range (box), and whiskers extending to the minimum and maximum values, excluding outliers—defined as points beyond 1.5× the interquartile range from the first and third quartiles.

4.1.4 Details of pattern classification and processing of missing data

This section provides details on the pattern classification and processing of missing data. Subcategories with Pattern-I and Pattern-II missingness are summarized in the left and right panels of Supplementary Table 5, respectively. For illustration, examples for these two patterns are provided as follows.

For Pattern-I missingness, activity-related variables are an example. Participants who indicated they were unable to work according to Data-Field 864 "Number of

Supplementary Table 5. Number of variables with Pattern-I and Pattern-II related missingness in different categories.

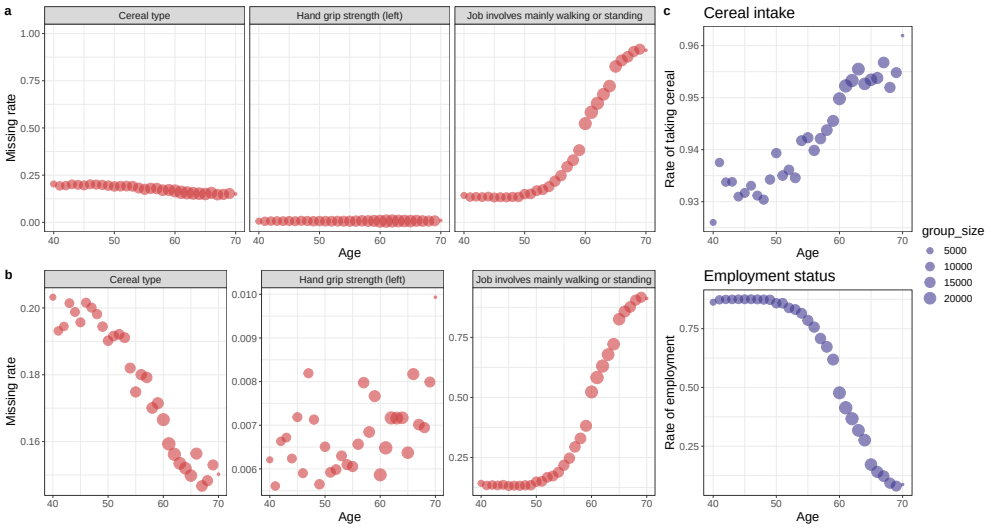
Pattern-I		Pattern-II	
Category	Number of Fields	Category	Number of Fields
Feeling	18	Smoking	16
Activity	12	Hearing	7
Alcohol	6	Activity	6
Mental health	6	Household	6
Electronic	2	Job	5
		Alcohol	4
		Diet	4
		Other	2
		Appearance	1
		Education	1
		Electronic	1
		Eyesight	1
		Mental health	1

days/week walked 10+ minutes" were not expected to answer questions about the duration of walks, usual walking pace, etc. Similarly, participants who indicated they hardly engaged in moderate physical activities were not expected to answer questions about the details of moderate activities. The same rationale applies to vigorous activity, strenuous sports, etc. For Pattern-II missingness, consider smoking-related variables as an example. Data-Field 6157 "Why stopped smoking" is meaningless for participants who gave the answer "smoke on most or all days" in Data-Field 22506 "Tobacco smoking". Such missingness caused by the inapplicability of certain subgroups, is referred to as Pattern-II missingness.

4.1.5 MAR missingness

MAR (Missing at Random) missingness is prevalent in UK Biobank data. For instance, the missingness of many variables is related to age. Supplementary Fig. 9 illustrates typical variables. The phenomena introduced earlier can often be related to other variables. For example, the missing rates of the "Cereal type" variable increase with age because people who do not consume cereal were not supposed to fill in this questionnaire item. As shown in Supplementary Fig. 9(c), cereal intake is more common among older people, leading to a lower missing rate of "Cereal type". The same applies to "Employment status". Since employment rates are lower among older people, the missing rate of job-related variables is higher. In contrast, the variable "Hand grip strength (left)" has a much lower missing rate and is not largely related to age.

4.2 P-values for 21 types of diseases with progressively added data categories



Supplementary Fig. 9. Relationship between missingness and age. a, Missing rates of three variables within different age groups. b, Magnified version of (a). c, Relationship between cereal intake/employment status and age. This largely explains the association in (b).

Supplementary Table 6. P-values for 21 types of diseases as different data types are incrementally added in the disease prediction model. Each added data type is compared with the model without that specific data type. The two-sided Wilcoxon test is used for hypothesis testing.

Category	+ Lifestyle	+ Measurement	+ Environment	+ Genetic	+ Image
Intestine	2.33×10^{-10}	1.33×10^{-6}	5.37×10^{-1}	1.83×10^{-1}	6.85×10^{-5}
Upper digestive	1.88×10^{-5}	8.37×10^{-5}	8.78×10^{-1}	5.94×10^{-3}	1.49×10^{-8}
Digestive gland	2.00×10^{-8}	4.66×10^{-9}	4.88×10^{-1}	6.50×10^{-2}	9.31×10^{-10}
Genital	2.56×10^{-5}	5.99×10^{-1}	1.73×10^{-3}	4.27×10^{-1}	1.12×10^{-12}
Pregnancy	6.98×10^{-1}	8.11×10^{-2}	5.80×10^{-3}	1.68×10^{-1}	1.37×10^{-2}
Oral	2.10×10^{-4}	2.44×10^{-1}	3.62×10^{-2}	5.91×10^{-1}	4.90×10^{-8}
Skin	2.63×10^{-5}	5.89×10^{-5}	1.80×10^{-1}	1.20×10^{-10}	8.74×10^{-10}
Eye	5.50×10^{-6}	1.03×10^{-7}	3.45×10^{-1}	2.05×10^{-3}	3.45×10^{-7}
Ear	3.32×10^{-4}	2.05×10^{-7}	2.29×10^{-2}	7.45×10^{-9}	4.44×10^{-3}
Tissues	1.09×10^{-8}	2.65×10^{-9}	1.38×10^{-3}	5.15×10^{-3}	4.35×10^{-10}
Urinary	1.66×10^{-7}	7.70×10^{-8}	1.51×10^{-1}	5.20×10^{-2}	2.77×10^{-7}
CNS	1.59×10^{-6}	3.25×10^{-1}	2.36×10^{-1}	1.05×10^{-1}	1.65×10^{-5}
Mental	8.44×10^{-9}	8.06×10^{-8}	7.88×10^{-1}	1.50×10^{-4}	5.82×10^{-9}
PNS	1.25×10^{-2}	3.36×10^{-3}	4.54×10^{-1}	8.90×10^{-1}	2.01×10^{-3}
Endocrine	7.38×10^{-12}	1.06×10^{-12}	7.92×10^{-1}	2.64×10^{-3}	4.57×10^{-11}
Blood	9.28×10^{-5}	2.99×10^{-8}	3.87×10^{-1}	6.87×10^{-1}	2.05×10^{-5}
Vessel	4.26×10^{-7}	5.01×10^{-6}	4.58×10^{-1}	2.38×10^{-1}	3.31×10^{-9}
Heart	3.44×10^{-9}	7.47×10^{-11}	2.27×10^{-1}	9.23×10^{-3}	2.07×10^{-11}
Upper respiratory	1.14×10^{-5}	1.91×10^{-5}	9.55×10^{-2}	8.23×10^{-3}	3.81×10^{-6}
Lung	8.04×10^{-7}	1.60×10^{-9}	4.79×10^{-1}	8.54×10^{-2}	6.91×10^{-8}
Externalities	9.19×10^{-2}	3.93×10^{-4}	6.23×10^{-1}	9.02×10^{-6}	4.95×10^{-7}

Supplementary Table 7. P-values for 21 types of diseases as different data types are incrementally added in the risk assessment model. Each added data type is compared with the model without that specific data type. The two-sided Wilcoxon test is used for hypothesis testing.

Category	+ Lifestyle	+ Measurement	+ Environment	+ Genetic	+ Image
Intestine	2.02×10^{-6}	4.80×10^{-1}	3.32×10^{-3}	1.45×10^{-4}	3.26×10^{-9}
Upper digestive	3.81×10^{-4}	8.15×10^{-2}	1.78×10^{-1}	8.28×10^{-4}	2.83×10^{-7}
Digestive gland	3.51×10^{-6}	1.11×10^{-6}	1.66×10^{-1}	6.92×10^{-1}	2.50×10^{-7}
Genital	4.90×10^{-5}	5.47×10^{-3}	7.16×10^{-4}	1.23×10^{-4}	3.86×10^{-14}
Pregnancy	5.77×10^{-1}	3.60×10^{-1}	1.69×10^{-2}	9.06×10^{-1}	1.21×10^{-4}
Oral	2.12×10^{-1}	7.22×10^{-1}	1.35×10^{-1}	3.89×10^{-3}	7.61×10^{-7}
Skin	1.52×10^{-4}	1.01×10^{-1}	8.85×10^{-4}	6.07×10^{-6}	1.62×10^{-9}
Eye	1.90×10^{-4}	2.62×10^{-4}	5.46×10^{-1}	2.38×10^{-7}	1.23×10^{-9}
Ear	4.05×10^{-1}	5.34×10^{-5}	7.73×10^{-5}	2.69×10^{-4}	3.73×10^{-9}
Tissues	9.07×10^{-6}	5.34×10^{-5}	8.74×10^{-2}	2.07×10^{-9}	1.49×10^{-13}
Urinary	5.57×10^{-6}	5.56×10^{-2}	7.87×10^{-3}	8.70×10^{-8}	1.65×10^{-6}
CNS	4.55×10^{-4}	9.49×10^{-1}	7.57×10^{-1}	2.89×10^{-4}	1.56×10^{-9}
Mental	1.69×10^{-8}	9.30×10^{-1}	5.68×10^{-2}	6.71×10^{-3}	6.54×10^{-8}
PNS	4.89×10^{-1}	1.03×10^{-2}	2.52×10^{-1}	8.04×10^{-1}	6.10×10^{-5}
Endocrine	4.46×10^{-8}	1.19×10^{-8}	1.76×10^{-3}	1.67×10^{-6}	9.03×10^{-14}
Blood	1.17×10^{-2}	7.04×10^{-7}	6.61×10^{-1}	1.35×10^{-2}	2.70×10^{-4}
Vessel	1.44×10^{-6}	4.85×10^{-3}	1.52×10^{-1}	3.33×10^{-3}	7.23×10^{-4}
Heart	1.76×10^{-5}	3.19×10^{-7}	6.23×10^{-3}	8.12×10^{-6}	1.67×10^{-11}
Upper respiratory	2.40×10^{-3}	2.25×10^{-1}	4.01×10^{-2}	3.61×10^{-2}	1.14×10^{-5}
Lung	3.20×10^{-9}	1.45×10^{-3}	3.63×10^{-2}	5.92×10^{-1}	1.30×10^{-6}
Externalities	1.24×10^{-3}	8.94×10^{-1}	2.20×10^{-3}	7.04×10^{-3}	3.46×10^{-7}

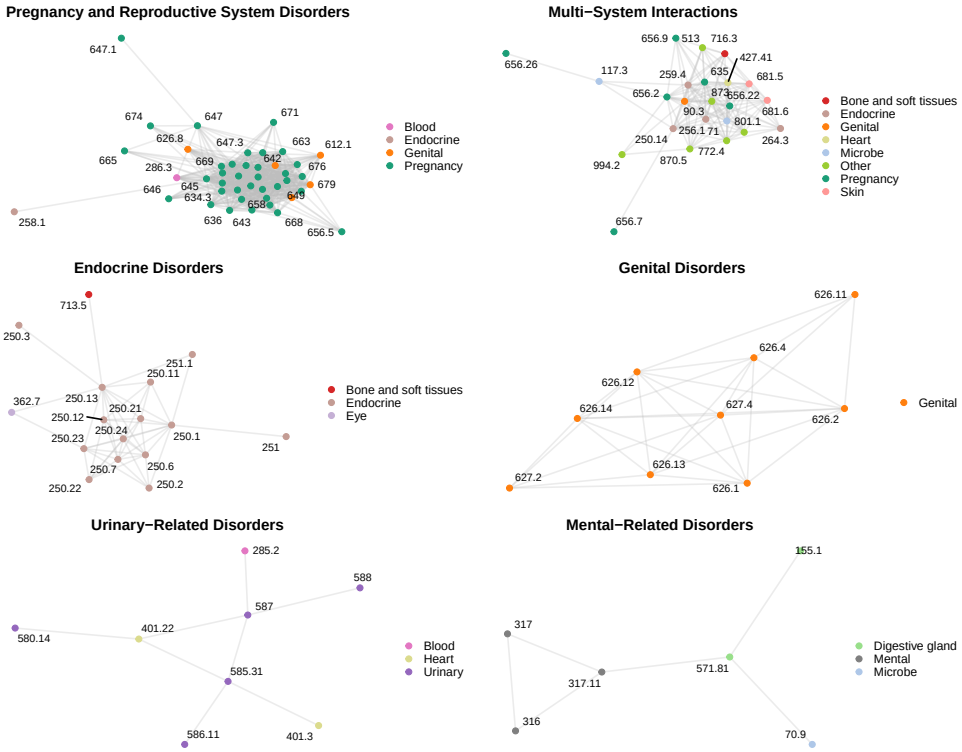
4.3 Additional results of UKB-MDRMF

4.3.1 Multimorbidity Network

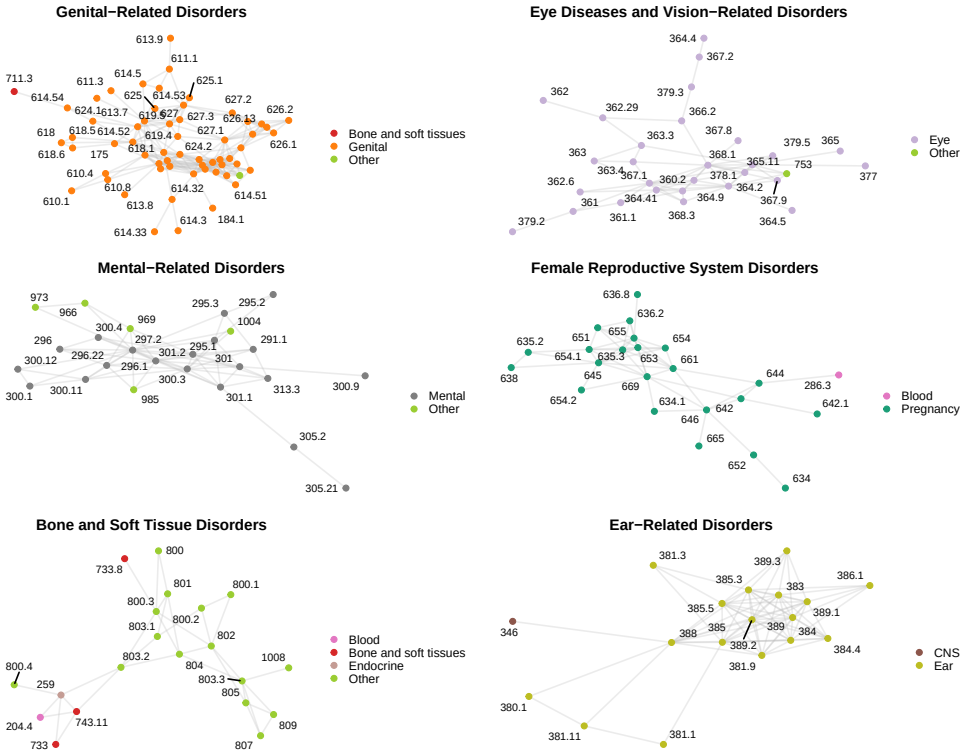
The multimorbidity network developed in UKB-MDRMF leverages embeddings extracted from the penultimate layer of a neural network model, utilizing FCNN for disease prediction and DeepSurv for risk assessment. Each of the 1,560 Phecodes is represented as a high-dimensional vector with 200 dimensions for FCNN and 300 dimensions for DeepSurv), and pairwise correlations between these embeddings are calculated to construct a correlation matrix. To ensure relevance, a threshold of 0.39 is applied to retain only significant correlations; for survival models like DeepSurv, where correlations tend to be higher, this threshold is increased to 0.6. From this refined correlation matrix, an undirected graph is constructed where nodes represent Phecodes and edges signify strong correlations. The Louvain method is then employed to detect communities within the graph [5], effectively identifying clusters of diseases with notable comorbid relationships. For clarity in presentation, the six largest communities are visualized, revealing distinct patterns of comorbidity and shared disease mechanisms.

Supplementary Fig. 10 displays the six major disease communities as identified by the Louvain algorithm using FCNN-generated embeddings. Each node corresponds to a specific Phecode, with edges reflecting significant correlations among them, suggesting potential comorbid relationships. These clusters encompass a variety of disease categories, including pregnancy and reproductive disorders; multi-system interactions involving endocrine, microbe, skin, and reproductive health; endocrine disorders; genital health issues; and mental health and behavioral conditions. Similarly, the multimorbidity network derived from the DeepSurv model presents comparable clustering patterns with additional granularity, distinguishing genital health issues into male- and female-specific clusters as shown in Supplementary Fig. 11. These findings underscore the model's ability to discern nuanced disease relationships and clustering patterns across a broad spectrum of conditions.

This analytical approach not only pinpoints clusters of diseases with robust comorbid relationships but also unveils links between diseases that share mechanisms but are not traditionally categorized together. For instance, within urinary-related diseases, the network identifies Hypertensive chronic kidney disease (401.22), typically classified under cardiovascular conditions, and Anemia of chronic disease (285.2), commonly grouped under blood disorders. The identification of these six primary clusters corroborates findings from other large-scale biomedical comorbidity studies [5, 6], reinforcing the robustness of our methodology. By revealing connections between diseases with overlapping pathways, our analysis offers invaluable insights into disease mechanisms, fostering further research into shared risk factors and facilitating targeted interventions.



Supplementary Fig. 10. Network visualization of six disease communities identified by Phecode clusters based on FCNN. Each node represents a specific Phecode, with edges indicating potential comorbid relationships between diseases. The node colors correspond to different disease categories, such as endocrine, urinary, reproductive health, and mental health disorders, derived from their Phecode classification. This visualization highlights the comorbidity patterns captured through the FCNN model's penultimate layer embeddings.



Supplementary Fig. 11. Network visualization of six disease communities identified by Phecode clusters based on DeepSurv. Each node represents a specific Phecode, with edges indicating potential comorbid relationships between diseases. The node colors correspond to distinct disease categories such as reproductive health, mental health, bone, soft tissue, vision-related, hearing, and genital disorders, derived from their Phecode.

4.3.2 Multi-center validation using All of US data

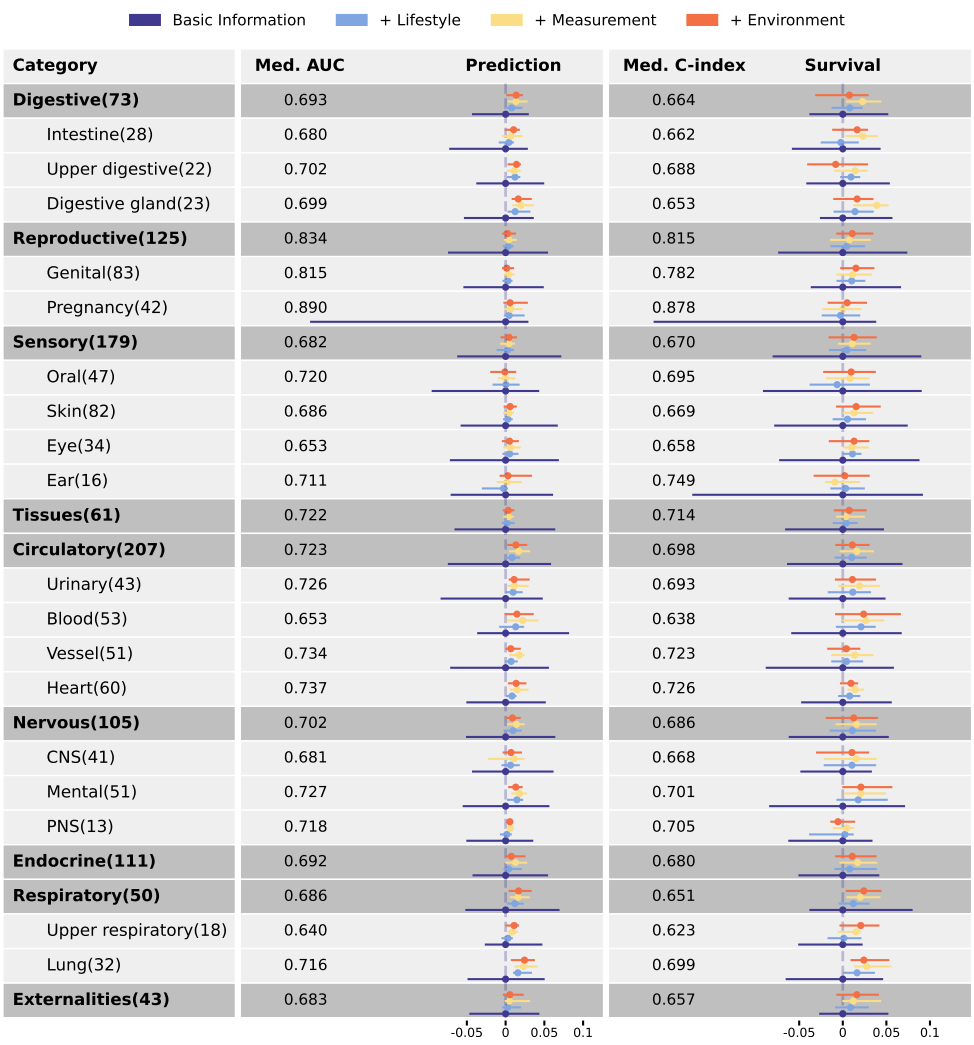
The "All of Us" Research Program is a major, nationwide initiative in the United States designed to collect and analyze health data from a diverse population to advance biomedical research [7]. This program encompasses a variety of data sources and modalities including survey data, EHR data, physical measurements, and genetics data. To validate our findings from the UK Biobank, we utilized survey-derived covariates and physical measurements, alongside outcomes derived from EHR data.

Mirroring our approach with the UK Biobank, we classified the covariates into four categories for the All of Us dataset: Basic information (14 variables including age, biological sex at birth, and annual income), Lifestyle (30 variables such as smoking habits and alcohol consumption), Environmental (5 variables including home ownership and housing stability), and Measurement data (3 variables including heart rate and blood pressure). The comprehensive data dictionary is available in Supplementary Data 3.

Our validation cohort included 218,743 subjects who had available EHR records and at least one variable from each covariate category. Survey responses were recoded into categorical or continuous formats as appropriate, with categorical data transformed into dummy variables—missing data were treated as a separate category—and continuous data subjected to mean imputation for missing values. EHR data were processed similarly to the UK Biobank, with ICD-10 codes mapped to Phecodes and diagnosis dates used to calculate survival times.

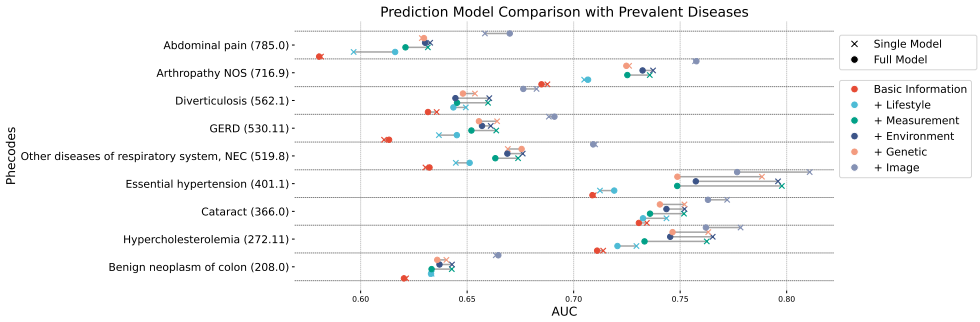
Supplementary Fig. 12 illustrates the performance of the FCNN models for disease prediction and DeepSurv models for survival analysis using the All of Us dataset. The median AUC (Med. AUC) for disease prediction demonstrated incremental improvement with the addition of lifestyle, clinical measurements, and environmental data, starting from basic demographic and health information. Notably, reproductive system diseases exhibited the highest predictive accuracy, with a median AUC of 0.834, followed by circulatory and nervous system diseases. Similarly, the median C-Index (Med. C-Index) for survival analysis showed consistent improvement, with reproductive diseases achieving the highest scores of 0.902.

These results align with those from the UK Biobank, emphasizing the significance of integrating diverse data types to enhance both prediction and survival modeling. The magnitude of improvement varied across disease categories; reproductive system diseases consistently performed well based primarily on basic information, while sensory-related diseases proved more challenging to predict. This performance trend is consistent with the modeling results from the UK Biobank. Furthermore, these findings underscore not only the variability in model performance across diseases but also the consistent predictive patterns observed in large-scale biomedical datasets from different centers. The supplementary analysis confirms the robustness and generalizability of the FCNN and DeepSurv models, reinforcing their applicability across multi-center datasets and diverse patient populations.

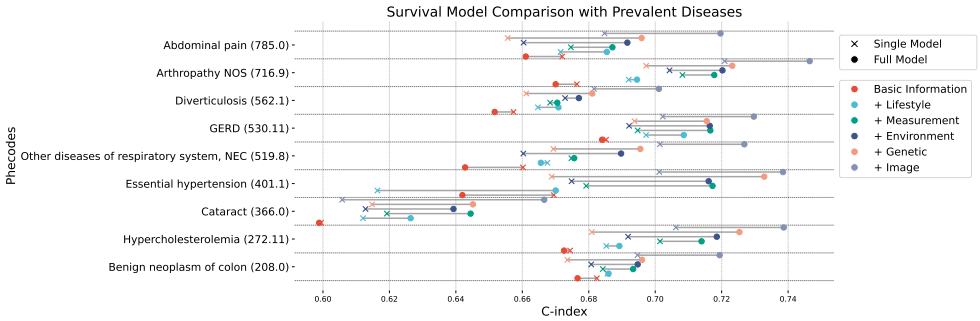


Supplementary Fig. 12. Model performance forest plot for different disease types using FCNN and DeepSurv models in the All of Us data. The figure displays the accuracy of disease prediction and survival modeling across various disease types, evaluated as additional data types are included incrementally. Med. AUC represents the median AUC of the FCNN model's performance for disease prediction, initialized with basic demographic and health information. Med. C-Index indicates the median C-Index of the DeepSurv model for survival analysis, based on the same initial data category. Each point denotes the median metric value, while the horizontal lines extend to the 25th and 75th percentiles, reflecting the variability in performance across diseases within each category.

4.3.3 Comparison of single vs. full model performance in disease prediction and risk assessment



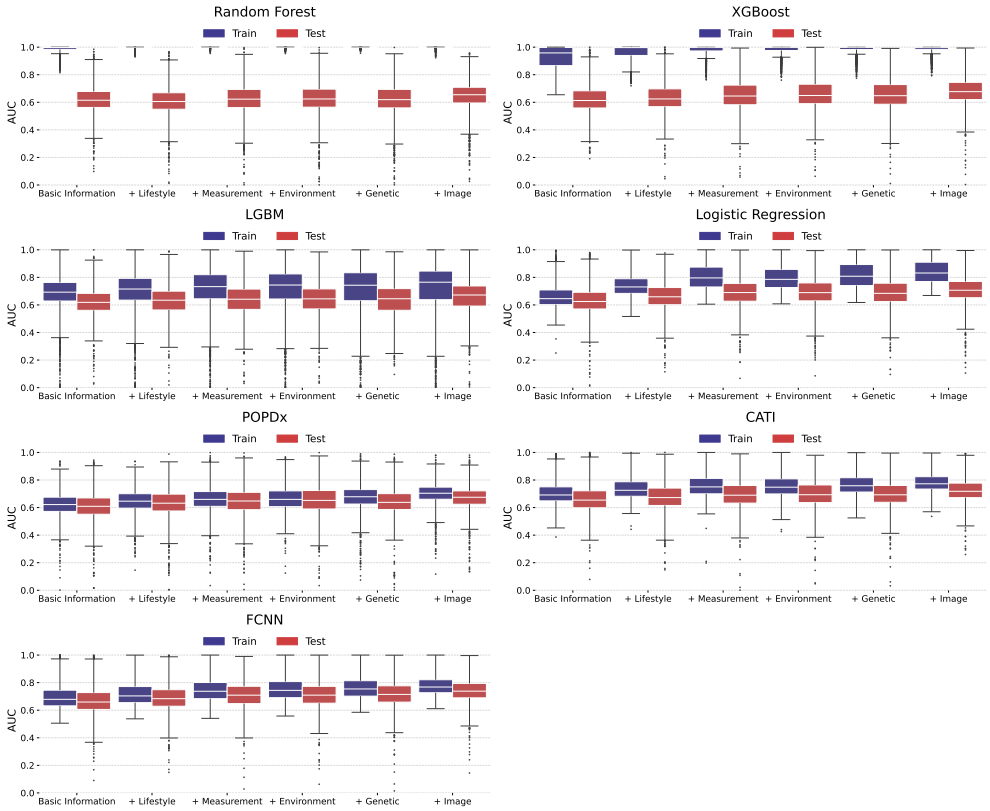
Supplementary Fig. 13. Comparison of AUC performance for disease prediction in Single vs. Full Model (UKB-MDRMF) on the test set. We evaluated nine high-prevalence diseases by comparing a model trained solely on individual diseases (Single Model) with the full model trained jointly on 1,560 Phecodes (UKB-MDRMF). Dots represent the results of the Full Model, while crosses indicate the performance of the Single Model. Different colors correspond to the sequential addition of different data categories.



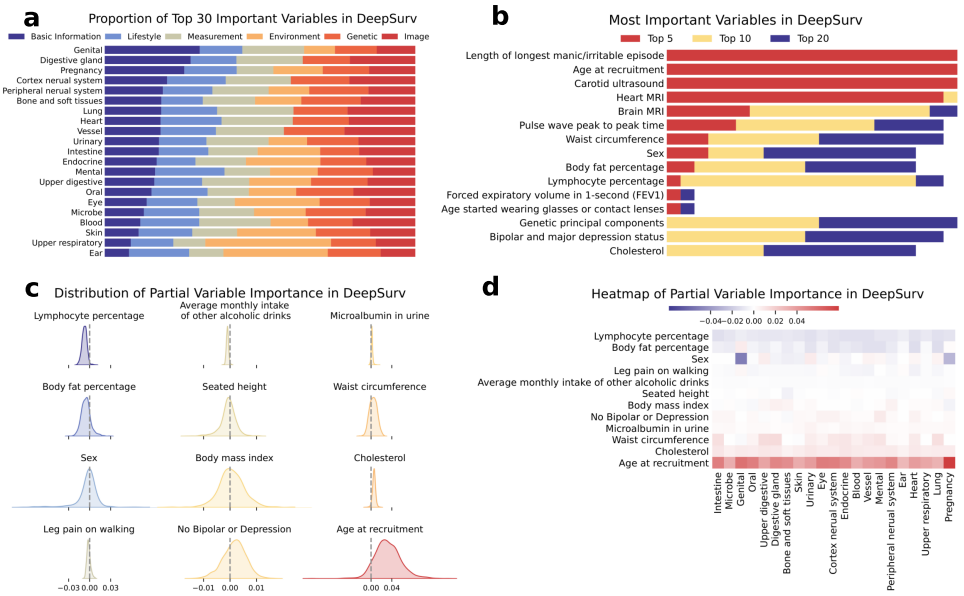
Supplementary Fig. 14. Comparison of AUC performance for individual risk assessment in Single vs. Full Model (UKB-MDRMF) on the test set. We evaluated nine high-prevalence diseases by comparing a model trained solely on individual diseases (Single Model) with the full model trained jointly on over 1,560 Phecodes (UKB-MDRMF). Dots represent the results of the Full Model, while crosses indicate the performance of the Single Model. Different colors correspond to the sequential addition of different data categories.

While our framework (UKB-MDRMF) simultaneously models over 1,500 Phecodes, the results for each disease can also be independently extracted and analyzed, allowing for a more detailed assessment of specific clinical outcomes. As shown in Supplementary Figs. 13 and 14, we compared the predictive performance of the full model, UKB-MDRMF (joint modeling of all Phecodes), with single-disease models trained independently for nine high-prevalence diseases, considering both disease prediction and survival analysis tasks. The results indicate that for disease prediction (Supplementary Fig. 13), joint modeling generally performs comparably to or better than single-disease modeling, except for Essential Hypertension (Phecode: 401.1) and Hypercholesterolemia (Phecode: 272.11), where the single-disease models showed slightly better performance. In contrast, for risk assessment (Supplementary Fig. 14), nearly all solid dots are positioned to the right of the "x" marks, demonstrating that the UKB-MDRMF framework significantly improves C-index performance through joint modeling. This suggests that incorporating shared information across diseases and risk factors enhances overall predictive accuracy, particularly in survival analysis.

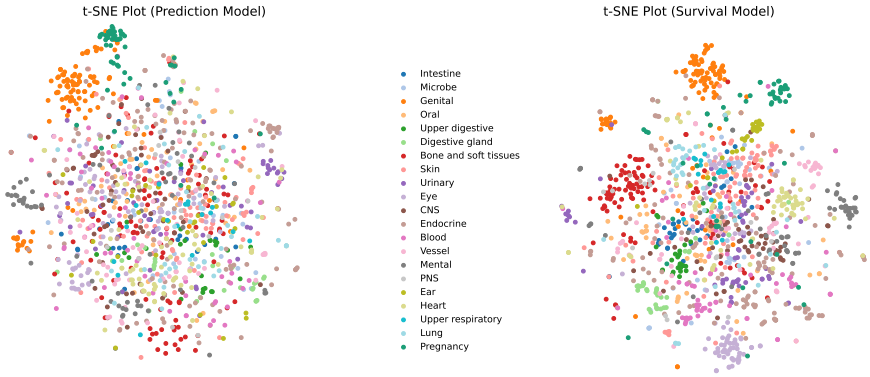
4.3.4 Additional figures



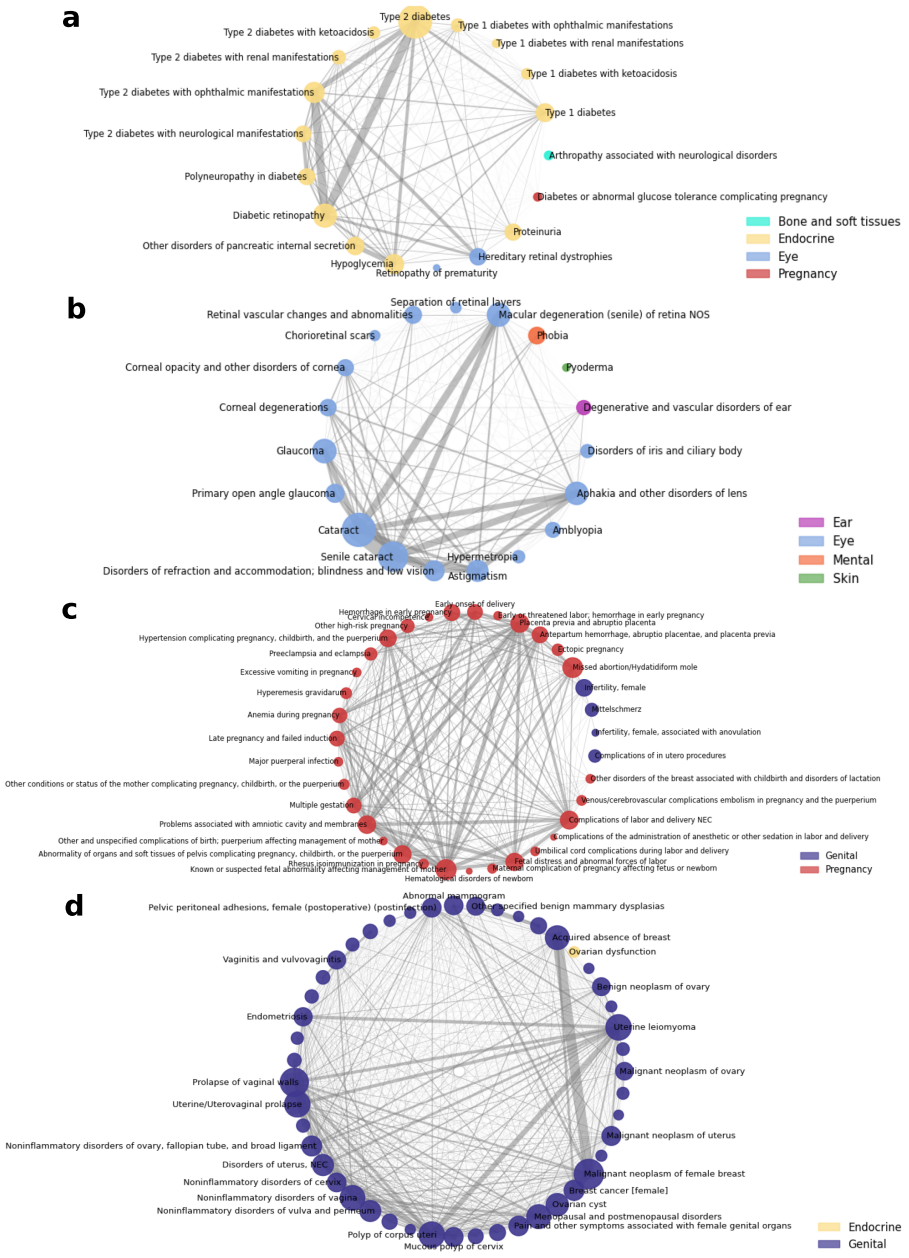
Supplementary Fig. 15. AUC on training and testing sets for different prediction models. Traditional machine learning models such as Random Forest, XGBoost, LGBM, and Logistic Regression commonly exhibit overfitting, whereas methods based on deep learning such as POPDx, CATI, and FCNN show relatively weaker overfitting, with higher consistency between performance on training and testing sets. Box plots depict the median (central line), interquartile range (box), and whiskers extending to the minimum and maximum values, excluding outliers—defined as points beyond 1.5× the interquartile range from the first and third quartiles.



Supplementary Fig. 16. Assessing the importance of various disease risk factors using SHAP value from DeepSurv. (a) Normalized proportion of the top 30 significant risk factors among six categories of independent variables for each of the 21 disease types. (b) Frequencies for the top 5, top 10, and top 20 important variables in each of the 21 types of diseases. (c) Distribution of variable importance for all Phecodes, with colors ranging from blue (negative effect) to red (positive effect). (d) Average importance values for each variable category and disease type, with blue indicating negative effects and red indicating positive effects.



Supplementary Fig. 17. Two-dimensional projection of the prediction and survival models' multimorbidity mechanisms using t-SNE for all Phecodes, where each point represents a predicted Phecode, and each color represents a major disease category.



Supplementary Fig. 18. Multimorbidity patterns of the remaining four selected clusters of internal Phecodes not shown in the main text. (a) Endocrine-related disease cluster. (b) Eye-related disease cluster. (c) Pregnancy-related disease cluster. (d) Genital-related disease cluster. The size of each data point indicates the number of affected individuals, with larger points representing higher frequencies of occurrence. The thickness of the lines represents the frequency of comorbidity, with thicker lines indicating higher frequencies.

4.3.5 Details of data processing for comparative analysis applications

In practical research applications, the challenge lies in converting disease codes. Most studies use the ICD-10 coding system, while our study uses a collection of ICD codes known as Phecodes. In our comparisons with other studies, we ensured code consistency. The correspondence between the ICD codes used in other works and the Phecodes selected for our study was presented in Supplementary Data 4. In the comparative analysis, we also altered the framework to binary input/output to train a single disease model and transferred the parameters of the all-disease model for initialization. Bold numbers in the table indicate the model that performs better between ours and theirs.

Supplementary Table 8. More results for comparative analysis (C-index)

Study	Disease	Their C-index	Our C-index
You J, et al. [8]	Diseases of Infections	0.62	**0.91**
	Bacterial infections	0.68	**0.91**
	Viral infections	0.54	**0.88**
	Cancer	0.66	**0.84**
	Lung cancer	**0.84**	0.77
	Breast cancer	0.74	**0.81**
	Prostate cancer	**0.85**	0.78
	Leukaemia	0.75	**0.81**
	Blood and immune disorders	0.73	**0.88**
	Anemia	0.76	**0.88**
	Endocrine disorders	0.72	**0.91**
	Diabetes	0.87	**0.93**
	Obesity	0.85	**0.90**
	Mental and behavioural disorders	0.70	**0.92**
	Dementia	**0.85**	0.78
	Mood disorders	0.66	**0.88**
	Nervous system disorders	0.65	**0.91**
	Neurotic disorders	0.60	**0.92**
	Parkinson disease	0.80	**0.88**
	Sleep disorders	0.69	**0.88**
	Eye disorders	0.67	**0.92**
	Ear disorders	0.58	**0.92**
	Circulatory system disorders	0.70	**0.89**
	Hypertension	0.79	**0.86**
	Ischemic heart diseases	0.76	**0.84**
	Arrhythmias	0.73	**0.84**
	Heart failure	**0.79**	0.77
	Stroke	0.73	**0.84**

Continued on next page

Supplementary Table 8. (continued)

Study	Disease	Their C-index	Our C-index
	Peripheral artery diseases	0.71	**0.88**
	Respiratory system disorders	0.64	**0.90**
	Chronic obstructive pulmonary disease	0.82	**0.90**
	Asthma	0.64	**0.82**
	Digestive system disorders	0.50	**0.93**
	Inflammatory bowel disease	**0.90**	0.86
	Liver diseases	0.72	**0.89**
	Skin disorders	0.59	**0.91**
	Musculoskeletal system disorders	0.60	**0.92**
	Osteoarthritis	0.69	**0.89**
	Genitourinary system disorders	0.64	**0.92**
	Renal failure	0.82	**0.83**
	Coronary heart disease	0.83	**0.84**
	Atrial fibrillation	**0.75**	0.69
Mars N, et al. [9]	Type 2 diabetes	0.84	**0.94**
	Breast cancer	0.75	**0.82**
	Prostate cancer	**0.86**	0.79
Markovitz A, et al. [10]	Pregnant	0.79	**0.88**
	Cardiovascular disease	0.72	**0.87**
Sun L, et al. [11]	Coronary heart disease	0.74	**0.83**
	Stroke	0.71	**0.86**

Supplementary References

[1] S.W. Choi, T.S.H. Mak, P.F. O'Reilly, Tutorial: a guide to performing polygenic risk score analyses. Nature protocols **15**(9), 2759–2772 (2020)

[2] M.J. Machiela, W.Y. Huang, W. Wong, S.I. Berndt, J. Sampson, J. De Almeida, M. Abubakar, J. Hislop, K.L. Chen, C. Dagnall, et al., Gwas explorer: an open-source tool to explore, visualize, and access gwas summary statistics in the plco atlas. Scientific Data **10**(1), 25 (2023)

[3] J.D. Eicher, C. Landowski, B. Stackhouse, A. Sloan, W. Chen, N. Jensen, J.P. Lien, R. Leslie, A.D. Johnson, Grasp v2. 0: an update on the genome-wide repository of associations between snps and phenotypes. Nucleic acids research **43**(D1), D799–D804 (2015)

[4] M. Costanzo, K. Bruskiewicz, L. Caulkins, M. Duby, C. Gilbert, Q. Hoang, D.K. Jang, A. Kluge, R. Koesterer, J. Massung, et al., An open-access platform for translating diabetes and cardiometabolic disease genetics into accessible knowledge. Journal of the Endocrine Society **5**(Supplement_1), A406–A406 (2021)

[5] G. Dong, J. Feng, F. Sun, J. Chen, X.M. Zhao, A global overview of genetically interpretable multimorbidities among common diseases in the uk biobank. Genome medicine **13**, 1–20 (2021)

[6] T. Beaney, J. Clarke, D. Salman, T. Woodcock, A. Majeed, P. Aylin, M. Barahona, Identifying multi-resolution clusters of diseases in ten million patients with multimorbidity in primary care in england. Communications Medicine **4**(1), 102 (2024)

[7] A.H. Ramirez, L. Sulieman, D.J. Schlueter, A. Halvorson, J. Qian, F. Ratsimbazafy, R. Loperena, K. Mayo, M. Basford, N. Deflaux, et al., The all of us research program: data quality, utility, and diversity. Patterns **3**(8) (2022)

[8] J. You, Y. Guo, Y. Zhang, J.J. Kang, L.B. Wang, J.F. Feng, W. Cheng, J.T. Yu, Plasma proteomic profiles predict individual future health risk. Nature Communications **14**(1), 7817 (2023)

[9] N. Mars, J.T. Koskela, P. Ripatti, T.T. Kiiskinen, A.S. Havulinna, J.V. Lindbohm, A. Ahola-Olli, M. Kurki, J. Karjalainen, P. Palta, et al., Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nature medicine **26**(4), 549–557 (2020)

[10] A.R. Markovitz, J.J. Stuart, J. Horn, P.L. Williams, E.B. Rimm, S.A. Missmer, L.J. Tanz, E.B. Haug, A. Fraser, S. Timpka, et al., Does pregnancy complication history improve cardiovascular disease risk prediction? findings from the hunt study in norway. European Heart Journal **40**(14), 1113–1120 (2019)

[11] L. Sun, L. Pennells, S. Kaptoge, C.P. Nelson, S.C. Ritchie, G. Abraham, M. Arnold, S. Bell, T. Bolton, S. Burgess, et al., Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. PLoS medicine **18**(1), e1003498 (2021)