Nuclear magnetic resonance-based metabolomics with machine learning for predicting progression from prediabetes to diabetes

Jiang Li[a,#], Yuefeng Yu[a,#], Ying Sun[a], Yanqi Fu[a], Wenqi Shen[a], Lingli Cai[a], Xiao Tan[b,c], Yan Cai[d], Ningjian Wang [a], Yingli Lu[a,*], Bin Wang[a,*]

[a]Institute and Department of Endocrinology and Metabolism, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[b]Department of Medical Sciences, Uppsala University, Uppsala, Sweden

[c]Department of Big Data in Health Science, School of Public Health, Zhejiang University School of Medicine, Hangzhou, China

[d]Department of Endocrinology, the Fifth Affiliated Hospital of Kunming Medical University, Yunnan Honghe Prefecture Central Hospital (Ge Jiu People's Hospital), Yunnan, China

#Jiang Li and Yuefeng Yu contributed equally to this manuscript.

*Corresponding Author

Bin Wang*, MD, PhD

Address: Institute and Department of Endocrinology and Metabolism, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200011 China. Telephone number: 0086-21-53315256.

E-mail: binwang1126@163.com

Yingli Lu*, MD, PhD

Address: Institute and Department of Endocrinology and Metabolism, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai,

200011 China.

E-mail: luyingli2008@126.com

Word count: 4,543

Number of tables and figures: 2 tables and 9 figures

Abstract

Background: Identification of individuals with prediabetes who are at high risk of developing diabetes allows for precise interventions. We aimed to determine the role of nuclear magnetic resonance (NMR)-based metabolomic signature in predicting the progression from prediabetes to diabetes.

Methods: This prospective study included 13,489 participants with prediabetes who had metabolomic data from the UK Biobank. Circulating metabolites were quantified via NMR spectroscopy. Cox proportional hazard (CPH) models were performed to estimate the associations between metabolites and diabetes risk. Supporting vector machine, random forest, and extreme gradient boosting were used to select the optimal metabolite panel for prediction. CPH and random survival forest (RSF) models were utilized to validate the predictive ability of the metabolites.

Results: During a median follow-up of 13.6 years, 2,525 participants developed diabetes. After adjusting for covariates, 94 of 168 metabolites were associated with risk of progression to diabetes. A panel of nine metabolites, selected by all three machine learning algorithms, was found to significantly improve diabetes risk prediction beyond conventional risk factors in the CPH model (area under the receiver operating characteristic curve [AUROC], 1-year: 0.823 for risk factors + metabolites vs 0.759 for risk factors, 5-year: 0.830 vs 0.798, 10-year: 0.801 vs 0.776, all P <0.05). Similar results were observed from the RSF model. Categorization of participants according to the predicted value thresholds revealed distinct cumulative risk of diabetes.

Conclusions: Our study lends support for use of the metabolite markers to help determine individuals with prediabetes who are at high risk of progressing to diabetes

and inform targeted and efficient interventions.

Funding: Shanghai Municipal Health Commission (2022XD017). Innovative Research Team of High-level Local Universities in Shanghai (SHSMU-ZDCX20212501). Shanghai Municipal Human Resources and Social Security Bureau (2020074). Clinical Research Plan of Shanghai Hospital Development Center (SHDC2020CR4006). Science and Technology Commission of Shanghai Municipality (22015810500).

Keywords: Prediabetes, diabetes, metabolomics, risk prediction, machine learning

Introduction

Prediabetes, an intermediate stage of glucose dysregulation that blood glucose levels are elevated but lower than in diabetes, has become a burgeoning global health emergency[1]. Prediabetes affected approximately 720 million individuals worldwide in 2021, with a project to 1 billion people by 2045[2]. Approximately 5% to 10% of people with prediabetes progress to having diabetes each year and the lifetime conversion rate to diabetes could be as high as 70%[3,4]. Therefore, preventing or delaying diabetes development among people with prediabetes will have substantial clinical and public health benefits.

Although lifestyle modification and medical therapy have been proven to be effective in preventing or delaying the diabetes onset among people with prediabetes[5-7], the substantial cost of modification programs and medications as well as drug-related side effects limit the widespread delivery of such interventions in this large high-risk population[8,9]. Notably, the progression from prediabetes to diabetes is highly heterogeneous, and a fraction of individuals with prediabetes may regress to normoglycemia without treatment.[10] Therefore, identifying targeted population who are at high risk of developing diabetes is the key step to tailor precise and efficient interventions. Glycemic indicators alone for risk stratification are deficient, with fasting glucose and glycosylated hemoglobin A1c (HbA1c) being convenient but less sensitive, while post-load glucose tolerance being sensitive but unfeasible in practice on a large scale[11,12]. In addition, several risk assessment models based on conventional clinical variables have been developed, but most of which had comparatively low performance and failed to take follow-up time into account[13-15].

Plasma metabolomics using high-throughput techniques could provide a comprehensive profiling of small-molecule metabolites in a specific physiological

period, which might yield valuable information for risk prediction. Previous studies have implied that incorporating circulating metabolites into basic models with conventional risk factors could improve prediction of diabetes risk [16-18]. However, we are aware of only one study that has assessed the relationship between metabolomic profiling and the progression to diabetes among individuals with prediabetes and investigated the predictive values of metabolites[19]. Nevertheless, it was limited by a nested case-control study design with a relatively short follow-up (median 5 years) and small sample size (n=~300). Whether addition of metabolic biomarkers improves the ability in predicting the progression from prediabetes to diabetes in prospective settings remains largely unknown.

To address these knowledge gaps, in the current study, we aimed to examine the longitudinal associations of circulating metabolic biomarkers, quantified using high-throughput nuclear magnetic resonance (NMR), with the risk of incident diabetes among individuals with prediabetes from the UK Biobank. Moreover, we evaluated whether metabolic signature adds anything to prediction models for diabetes development and risk stratification.

Methods

Study design and participants

The UK Biobank is a large population-based prospective cohort study enrolling more than 500,000 community-dwelling adults from 22 assessment centers across the UK between 2006 and 2010[20,21]. Participants completed touchscreen questionnaires and physical measurements and provided blood samples at baseline. The study was approved by the Northwest Multicenter Research Ethics Committee (REC reference for UK Biobank 11/NW/0382), and all participants provided informed consent.

For the identification of metabolomic biomarkers associated with the progression

from prediabetes to diabetes, the current study focused on participants with prediabetes at baseline with available circulating metabolite data. The diagnosis of prediabetes was defined by an HbA1c level of 5.7% to 6.4% (39 to 47 mmol/mol) in participants without diabetes, according to the American Diabetes Association (ADA) criteria[22]. After excluding individuals who developed diabetes or died within 1 month from the baseline, 13,489 participants with prediabetes were included in the final analyses.

Metabolite quantification

The metabolomics analysis of approximately 118,000 non-fasting ethylenediaminetetraacetic acid (EDTA) plasma samples at baseline was performed using the high-throughput NMR platform in Nightingale Health's laboratories of Finland. Details of the metabolic profiling platform and experimentation have been described elsewhere[23-25]. In brief, the EDTA samples were collected and stored at -80°C. Before preparation, frozen samples were slowly thawed at +4°C overnight and were centrifuged (3,400 g) for 3 minutes. Each sample was analyzed with a spectrometer and the metabolic biomarkers were quantified using Nightingale Health's proprietary software. The quality control procedures were implemented during the whole process and only samples and biomarkers that underwent the quality control process were stored in the UK Biobank dataset.

The metabolic biomarker profiling by Nightingale Health's NMR platform provides consistent results over time and across spectrometers. Furthermore, the sample preparation is minimal in the Nightingale Health's metabolic biomarker platform, circumventing all extraction steps. These aspects result in highly repeatable biomarker measurements. Pre-specified quality metrics were agreed between UK

Biobank and Nightingale Health to ensure consistent results across the samples, and pilot measurements were conducted. Nightingale Health performed real-time monitoring of the measurement consistency within and between spectrometers throughout the UK Biobank samples. Two control samples provided by Nightingale Health were included in each 96-well plate for tracking the consistency across multiple spectrometers. Furthermore, two blind duplicate samples provided by the UK Biobank were included in each well plate, with the position information unlocked only after results delivery. Coefficient of variation (CV) targets across the metabolic biomarker profile were pre-specified for both Nightingale Health's internal control samples and UK Biobank's blind duplicates. The targets were met for each consecutively measured batch of ~25,000 samples. For the majority of the metabolic biomarkers, the CVs were below 5% (https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=3000). Further, the distributions of measured biomarkers from 5 sample batches indicated absence of batch effects (https://biobank.ctsu.ox.ac.uk/ukb/ukb/docs/nmrm_app1).

A total of 249 metabolic biomarkers (168 directly measured and 81 ratios of these), spanning lipids, lipoprotein subclass, fatty acids, amino acids, ketone bodies, and glycolysis metabolites were quantified for each sample. In the present study, we analyzed 168 metabolic biomarkers that were directly measured (Supplementary file 1). The values of all metabolites were transformed using natural logarithmic transformation (ln[x+1]) followed by Z-transformation.

Covariate collection

Information on covariates was collected through a self-completed touchscreen questionnaire or verbal interview, including age, sex, ethnicity (white people or

others), Townsend deprivation index, household income (high: ≥ £52,000, medium: £18,000-£51,999, and low: < £18,000), education (college/university degree or others), employment status (current working, retired, or other), smoking status (never, previous, or current smoking), moderate alcohol (alcohol intake > 0 g and ≤ 14 g/day for women and alcohol intake > 0 g and ≤ 28 g/day for men), physical activity, healthy diet score, healthy sleep score, family history of diabetes (yes or no), history of cardiovascular disease (CVD, yes or no), history of hypertension (yes or no), history of dyslipidemia (yes or no), history of chronic lung diseases (CLD, e.g. chronic bronchitis, emphysema, and chronic obstructive pulmonary disease, yes or no), and history of cancer (yes or no). The Townsend deprivation index is a composite measure of area-level socioeconomic deprivation, with a higher score indicating higher levels of socioeconomic deprivation. Physical activity was measured by the metabolic equivalent task (MET) (sum of days performing walking, moderate activity, and vigorous activity)[27]. A healthy diet score was calculated based on the intake of vegetables (≥ median), fruits (≥ median), fish (≥ median), red meat (< median), and processed red meat (< median)[28]. One point was given for each favorable diet factor and the total diet score ranges from 0 to 5. A healthy sleep score was evaluated based on insomnia (sometimes or never), sleep duration (7~8 h), chronotype (morning person), daytime sleepiness (sometimes or never), and snoring (no)[29]. Each favorable sleep factor was given a score of 1, with the total sleep score ranging from 0 to 5. The history of dyslipidemia incorporated information on both the medical history of dyslipidemia and the use of lipid-lowering medications.

Physical measurements including blood pressure, height, weight, waist circumference (WC), and hip circumference (HC) were measured using calibrated instruments with standard protocols by trained nurses. Blood pressure was measured

using the Omron automatic digital monitor and two measurements were obtained at a few minutes' intervals. We calculated the mean systolic (SBP) and diastolic blood pressure (DBP) from two measurements. Body mass index (BMI) was calculated as weight in kilograms divided by the square of height in meters (kg/m^2). The HbA1c level was measured by high-performance liquid chromatography with the VARIANT II Turbo analyzer (Bio-Rad Laboratories). Missing covariates were imputed by the median value for continuous variables and a missing indicator for categorical variables.

Ascertainment of diabetes

Incident diabetes was ascertained from hospital inpatient records, death registers, and primary care records, according to the International Classification of Diseases, 10th revision (ICD-10) codes. Detailed information about the linkage procedure is available from https://content.digital.nhs.uk/services. The follow-up time was calculated from the baseline to the occurrence of diabetes, death, or the censoring date (30 March 2023), whichever came first.

Statistical analyses

Baseline characteristics were presented as numbers (percentages) for categorical variables and means (standard deviations, SDs) for continuous variables, respectively. Continuous variables were assessed for statistical differences using *t*-test and categorical variables were evaluated using the χ^2 test. Overall schematic workflow of the study is shown in **Figure 1**.

Metabolite selection

We first used Cox proportional hazards (CPH) model to assess the associations between individual metabolites and risk of diabetes progression with adjustment for

sociodemographic covariates (age, sex, ethnicity, education, Townsend Deprivation Index, employment status and household income), family history of diabetes, health conditions (history of CVD, hypertension, dyslipidemia, CLD and cancer), physical measurements (BMI, WC, HC, SBP and DBP), lifestyle factors (smoking status, moderate alcohol, healthy diet score, healthy sleep score and physical activity), and HbA1c. The potential confounders were selected based on prior knowledge of the risk factors for diabetes. Metabolites that were significantly associated with incident diabetes (P <0.05/168) were retained.

Secondly, we performed priority-Lasso to deal with multicollinearity in high dimensional data and to retain variables with nonzero coefficients. Priority-Lasso is a Lasso-based intuitive analysis strategy, which uses prior knowledge regarding the outcome by defining the blocks of different types of predictor variables[30]. In this study, we defined the 24 covariates as block 1, while all metabolites significantly associated with diabetes risk in the CPH model were defined as block 2. The penalization parameter λ was determined as values with maximum partial-likelihood in a 10-fold cross-validation.

Thirdly, three machine learning models including supporting vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) were adopted to further evaluate the importance of the Lasso-selected metabolites, as they can model nonlinear and nonadditive relations more flexibly[31]. Models were built by 10-fold cross-validation through the “caret” package. Common signals detected across diverse approaches are more likely to represent the strongest and true patterns in the data. We chose the intersection set of the top 20 most important variables selected by the three machine learning models, after balancing the performance of the final diabetes risk prediction model and the clinical applicability associated with

measurement costs of metabolites.

Model development

Participants were randomly subclassified into a training set and a test set at a ratio of 8:2 and two common algorithms for survival data including CPH model and random survival forest (RSF)[32] were adopted for model development. RSF, as a machine learning method, is designed to be used specifically for survival outcome prediction and has shown promising results in various settings[33,34]. It builds many decision trees using split points based on the log-rank test to identify different survival statuses and produces the predicted probability for an individual derived from the average prediction across all trees[35]. The RSF model was fitted using the "randomForestSRC" package and the grid search method was used for hyperparameter tuning (number of trees, number of variables to possibly split at each node, and minimum size of terminal node). Specifically, the grid search method was used to tune hyperparameters among the RSF model, through minimizing out-of-sample or out-of-bag error[36]. Each tree in the RSF is constructed from a random sample of the data, typically a bootstrap sample or 63.2% of the sample size (as in the present study). Consequently, not all observations are used to construct each tree. The observations that are not used in the construction of a tree are referred to as out-of-bag observations. In an RSF model, each tree is built from a different sample of the original data, so each observation is "out-of-bag" for some of the trees. The prediction for an observation can then be obtained using only those trees for which the observation was not used for the construction. A classification for each observation is obtained in this way and the error rate can be estimated from these predictions. The resulting error rate is referred to as the out-of-bag error. Through calculating the out-of-bag error in each iteration, the best hyperparameters were finally determined. The hyperparameters to be tuned and

range of grid search in the present study were below: number of trees (50-1000, by 50), number of variables to possibly split at each node (3-6, by 1), and minimum size of terminal node (1-20, by 1)[37].

Model evaluation

The model performance was assessed in the test set. The time-dependent area under the receiver operating characteristic curve (AUROC) was used to evaluate the model's discrimination ability. Continuous net reclassification improvement (NRI), and absolute integrated discrimination improvement (IDI) were used to assess whether adding the selected metabolites could improve risk discrimination and reclassification for the risk of progression from prediabetes to diabetes over the basic model that was built on 10 conventional clinical variables (age, sex, Townsend Deprivation Index, family history of diabetes mellitus, BMI, WC, HC, SBP, DBP, and HbA1c)[38]. The calibration ability of the model was estimated using calibration curve. Furthermore, we used decision curve analysis (DCA) to assess the clinical usefulness of prediction model-based guidance for prediabetes management, which calculates a clinical "net benefit" for one or more prediction models in comparison to default strategies of treating all or no patients[39]. To facilitate risk stratification, we classified participants into two risk groups according to the predictive value using "surv_cutpoint" function in the "survminer" R package[40]. We also divided participants into three categories according to the tertiles of probability. In addition, we included 90,688 participants with normal glucose from the UK Biobank and divided them into the training and test sets using an 8:2 ratio to further investigate the additive value of the selected metabolites in diabetes prediction among participants with normoglycemia. All analyses were conducted in R software (version 4.2.2). A two-sided P value < 0.05 was considered statistically significant. To control for the

false discovery rate in the association between multiple metabolic biomarkers and incident diabetes, Bonferroni correction for *P* value ($P < 0.05/168$) was used.

Results

Baseline characteristics

Among the 13,489 participants with baseline prediabetes, the mean age was 59.6 (SD, 7.1) years, and 6,166 (45.7%) were males. During a median follow-up of 13.6 (12.3-14.6) years, 2,525 (18.7%) participants progressed to diabetes. Baseline characteristics of the study population stratified by incident diabetes are summarized in **Table 1**. Participants who developed diabetes were more likely to be male, non-White, less educated, more deprived, and smokers. They also tended to have a family history of diabetes, comorbidities such as CVD, hypertension, dyslipidemia and CLD, and higher levels of BMI, WC, and HC.

Identification of metabolic biomarkers for progression to diabetes

After adjusting for covariates and correcting for multiple testing, 94 of 168 metabolic biomarkers were significantly associated with the risk of incident diabetes (**Figure 2**, Supplementary file 2). Concentrations of very low-density lipoprotein (VLDL) particles, particularly larger VLDL particles and composition within larger VLDL, were strongly associated with progression to diabetes. Triglyceride in all lipoprotein subclasses also demonstrated strong positive associations with diabetes risk. In contrast, concentrations of larger high-density lipoprotein (HDL) particles and composition within these particles were inversely associated with incident diabetes. For lipoprotein particle diameter, larger HDL and LDL particle sizes were associated with a lower risk of progression to diabetes, while larger VLDL particle size was associated with a higher risk.

Monounsaturated fatty acids and saturated fatty acids were positively associated with the risk of diabetes, whereas docosahexaenoic acid and the degree of fatty acid unsaturation were negatively associated with diabetes. Among the amino acids, higher concentrations of alanine, tyrosine, and branched-chain amino acid (BCAA) such as leucine and valine were associated with an increased risk of diabetes, but glutamine and glycine were inversely associated with diabetes. Neither of the ketone bodies showed an association with the risk of diabetes.

Of the 94 metabolites that were significantly associated with diabetes, 17 metabolites were selected by priority-Lasso (Supplementary file 3). When further evaluating the importance of these metabolites after adjustment for covariates using three machine learning algorithms, the intersection of the top 20 important predictors identified a total of 9 metabolites, namely cholesteryl esters in large HDL, cholesteryl esters in medium VLDL, triglycerides in very large VLDL, average diameter for LDL particles, triglycerides in IDL, glycine, tyrosine, glucose, and docosahexaenoic acid (**Figure 3**, Supplementary file 4).

Model development and evaluation

Build upon the selected 9 metabolites and 10 clinical variables, there was no obvious difference in the AUROC obtained from CPH model (1-year: 0.823 [95% confidence interval, CI 0.702, 0.945]; 5-year: 0.830 [0.797, 0.864]; 10-year: 0.801 [0.778, 0.825]) and RSF model (1-year: 0.828 [0.723, 0.933]; 5-year: 0.820 [0.785, 0.855]; 10-year: 0.802 [0.778, 0.826]). Hence, we chose CPH model as the final model because of its simplicity and interpretability. The addition of selected metabolites consecutively outperformed the basic model with conventional clinical variables in diabetes risk prediction from 1 to 10 years (**Figure 4**). Specifically, the AUROC increased from 0.759 (95% CI 0.608, 0.911) to 0.823 (0.702, 0.945), 0.798 (0.762, 0.834) to 0.830

(0.797, 0.864), and 0.776 (0.750, 0.801) to 0.801 (0.778, 0.825) for 1-year, 5-year, and 10-year diabetes risk, respectively (**Table 2**, **Figure 5**). Results from continuous NRI and absolute IDI also demonstrated improvement in the risk prediction for progression to diabetes (**Table 2**), although the model calibration was not significantly improved (**Figure 6**). The decision curve analysis showed that the inclusion of the metabolites had a higher net benefit across the threshold probabilities of 0-0.35 for predicting 5-year diabetes risk and 0-0.55 for predicting 10-year diabetes risk (**Figure 7**).

We further categorized the participants from the test set into low-risk and high-risk groups according to the optimal threshold of the predicted value (1.02) reflecting the best risk difference. Compared with the low-risk group, participants in the high-risk group had a significantly higher cumulative risk of incident diabetes (log-rank $P < 0.0001$) (**Figure 8**). When participants were alternatively classified into low-risk, medium-risk, and high-risk groups according to the tertile cut-off point of the predicted value, the high-risk group showed the highest risk of developing diabetes, followed by the medium-risk and low-risk groups (log-rank $P < 0.0001$). Similar results were also observed when considering the competing risk from death (Fine-Gray $P < 0.0001$) (**Figure 8-figure supplement 1**). In addition, the predicted risk of diabetes within 1 year ($P = 0.001$), 5 years ($P < 0.001$), or 10 years ($P < 0.001$) was generally higher among participants who progressed to diabetes than those who did not (**Figure 9**).

Among participants with normoglycemia, we also observed a significant improvement in the prediction of diabetes after the addition of metabolic biomarkers to the basic model. The AUROC increased from 0.821 (95% CI 0.736, 0.907) to 0.868 (0.802, 0.934), 0.790 (0.738, 0.842) to 0.811 (0.762, 0.860), and 0.791 (0.765, 0.816) to 0.806 (0.781, 0.831) for 1-year, 5-year, and 10-year diabetes risk, respectively

(Supplementary file 5). The increases in NRI and IDI were similar to or slightly lower than those found among participants with prediabetes.

Discussion

By leveraging data from the large UK Biobank cohort, this prospective study provided a comprehensive analysis of the associations of circulating metabolites with the risk of progression to diabetes and predictive ability in participants with prediabetes. We found that lipoprotein particles, lipoprotein particle size and composition, fatty acids, and amino acids were associated with the risk of incident diabetes. More importantly, our findings suggested that adding the selected metabolites (i.e., cholesteryl esters in large HDL, cholesteryl esters in medium VLDL, triglycerides in very large VLDL, average diameter for LDL particles, triglycerides in IDL, glycine, tyrosine, glucose, and docosahexaenoic acid) could significantly improve the risk prediction of progression from prediabetes to diabetes beyond the conventional clinical variables.

In the present study, the association between diabetes risk and lipid and lipoprotein profile, including VLDL particles and composition with larger VLDL, HDL particles and composition within larger HDL, triglyceride, smaller HDL and LDL particle sizes, and larger VLDL particle sizes, were broadly consistent with previous studies in the general population[41-44]. BCAAs have been widely reported to be involved in the pathogenesis of diabetes, which might impair insulin signaling and lead to increased insulin secretion and pancreatic β-cell exhaustion[45]. Furthermore, genetic association studies have shown higher BCAAS resulting from insulin resistance, which may in turn cause diabetes[46,47]. Our study confirmed the vital role of these metabolites in the progression to diabetes among individuals with prediabetes.

Several risk assessment models for predicting the risk of progression from prediabetes to diabetes have been reported[13-15]. Yokota et al. developed a logistic

regression model to predict the risk for conversion from prediabetes to diabetes based on family history of diabetes, sex, SBP, fasting plasma glucose (FPG), HbA1c, and alanine aminotransferase (ALT)[13]. The model derived from a retrospective longitudinal study design achieved an AUROC of 0.80 (0.70–0.87) but did not take follow-up time into account. Similarly, Liang et al developed a predictive model using three glycemic indicators (FPG, 2-h postprandial blood glucose [2-hPG], and HbA1c) alone[15] and obtained a relatively low AUROC of 0.732 (95% CI 0.688-0.776). In a cohort study of 852,454 individuals with prediabetes, a machine-learning model predicting the progression to diabetes within 1-year was established using data from electronic medical records[14]. The model built on age, gender, BMI, medication usage, and laboratory results achieved a high AUROC of 0.865 (0.860-0.869). However, the model's performance over a longer follow-up period was unclear and conventional parameters such as lifestyle, family history of diabetes or comorbidities were not taken into account.

Changes in circulating small-molecule metabolites may occur long before the disease onset. Although rapid development in the technology of metabolomics provides a powerful tool for precise disease prediction, few studies have investigated the role of metabolomics-derived metabolic biomarkers in predicting progression from prediabetes to diabetes. To our best knowledge, only one case-control study among 153 individuals with prediabetes and 160 matched controls reported that adding 13 metabolites to conventional clinical variables including BMI, waist-hip ratio, WC, SBP, DBP, triglyceride, LDL, and triglyceride-glucose index improved the risk prediction of diabetes progression within 5 years, with the AUROC increasing from 0.72 to 0.98[19]. However, the predictive ability of metabolites in prospective settings with large sample size remains uncertain. In this longitudinal study among

13,489 participants with prediabetes, we comprehensively used multiple machine learning algorithms to identify a panel of 9 circulating metabolites that were associated with diabetes incidence during a median follow-up of 13.6 years. The CPH model integrating conventional clinical variables and the selected metabolic signature achieved a comparatively high AUROC of 0.823, 0.830, and 0.801 for 1-year, 5-year, and 10-year diabetes risk, respectively. Importantly, the addition of the metabolites resulted in a significant improvement in the discrimination ability and risk reclassification of diabetes beyond conventional risk factors. Furthermore, we categorized participants according to the optimal threshold points of the predicted value and found that the high-risk group had a significantly higher cumulative incidence of diabetes than the low-risk group. Most importantly, a model with good discrimination does not necessarily have high clinical value. Hence, DCA was used to compare the clinical utility of the model before and after adding the metabolites, and this showed a higher net benefit for the latter than the basic model, suggesting the addition of the metabolites increased the clinical value of prediction, i.e., the potential benefit of guiding management in individuals with prediabetes[39,48]. These results provided novel evidence supporting the value of metabolic biomarkers in risk prediction and stratification for the progression from prediabetes to diabetes. Considering the epidemic proportion of prediabetes worldwide, even a modest improvement in diabetes risk prediction among individuals with prediabetes will have substantial clinical and public health implications. Early detection of individuals with prediabetes who are at high risk of developing diabetes would not only advance targeted screening initiatives, health management and interventions but also facilitate a rational allocation of medical resources while avoiding disproportionate healthcare expenditure, which could finally translate into precise and efficient prevention of

diabetes. The value of the selected metabolic biomarkers in diabetes prediction was also confirmed in individuals with normal glucose.

Our study presents several strengths. Circulating metabolites were quantified via NMR-based metabolome profiling within the UK Biobank, which offers metabolite qualification with relatively lower costs and better reproducibility[26]. Additional strengths of our study included large sample size, prospective study design with long-term follow-up, and comprehensive control of covariates. Moreover, we used multiple machine learning algorithms to identify the consistently important metabolic biomarkers based on which we developed the predictive models. The final model exhibited relatively high performance for 1-year, 5-year, and 10-year diabetes risk prediction. However, several limitations of our study should be noted. First, since FPG and 2-hPG were not available in the UK Biobank, we defined prediabetes using HbA1c alone and to what extent our results could be extrapolated to other people with prediabetes determined by multiple glycemic indicators requires further investigation. Second, circulating metabolites were measured at baseline, thus their dynamic change over time could not be captured. However, our models showed stable performance in predicting short-term and long-term progression to diabetes (1 to 10 years), indicating the validity of single measurements of metabolic biomarkers for risk prediction. Third, the Nightingale metabolomics platform primarily focused on lipids and lipoprotein sub-fractions, and thus the predictive value of other metabolites in the progression from prediabetes to diabetes warranted further research using an untargeted metabolomics approach. Additionally, the use of non-fasting blood samples might increase inter-individual variation in metabolic biomarker concentrations, however, fasting duration has been reported to account for only a small proportion of variation in plasma metabolic biomarker concentrations[49]. Therefore, we believe the impact of

non-fasting samples on our findings would be minor. Fourth, although incident diabetes cases were ascertained through different data sources, including hospital inpatient records, death registers, and primary care records, some undiagnosed diabetes might have been missed. This misclassification would underestimate the effect of the observed associations between metabolites and diabetes risk. Fifth, we could not draw any conclusion about the causality between the identified metabolites and the risk for progression to diabetes due to the observational nature, which remained to be validated in further experimental studies. Sixth, in this study, the prediction models were established and tested using the UK Biobank dataset, external validation in an independent cohort is warranted to confirm the predictive values of the metabolic biomarkers. Finally, the participants from the UK Biobank were mostly White, which might limit the generalizability of the findings to other populations.

Conclusions

In this large prospective study among individuals with prediabetes, we detected a panel of circulating metabolites that were associated with an increased risk of progressing to diabetes. Use of these metabolites significantly improved the risk prediction of progression from prediabetes to diabetes. Our findings provide evidence that integrating metabolite markers with conventional risk factors is a promising approach to advance effective screening strategies and precise interventions for individuals with prediabetes who are at high risk of developing diabetes.

Availability of data and materials

The data analyzed during this study are available at https://www.ukbiobank.ac.uk/. This research has been conducted using the UK Biobank Resource under application number 77740.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank all participants and staff in the UK Biobank for their dedication and contribution to this study.

Authors' contributions

B.W. and Y.L. conceived and designed the study. J.L. and Y.Y. performed the statistical analysis and drafted the manuscript. Y.S., Y.F., W.S., and L.C. participated in data collection. B.W., Y.L., X.T., N.W. critically revised the manuscript. All authors read and approved the final manuscript. B.W. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1 Echouffo-Tcheugui, J. B. & Selvin, E. Prediabetes and What It Means: The Epidemiological Evidence. *Annu Rev Public Health* **42**, 59-77 (2021). https://doi.org/10.1146/annurev-publhealth-090419-102644

2 Sun, H. *et al.* IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* **183**, 109119 (2022). https://doi.org/10.1016/j.diabres.2021.109119

3 Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J. & Kivimäki, M. Prediabetes: a high-risk state for diabetes development. *Lancet* **379**, 2279-2290 (2012). https://doi.org/10.1016/s0140-6736(12)60283-9

4 Ligthart, S. *et al.* Lifetime risk of developing impaired glucose metabolism and eventual progression from prediabetes to type 2 diabetes: a prospective cohort study. *Lancet Diabetes Endocrinol* **4**, 44-51 (2016). https://doi.org/10.1016/s2213-8587(15)00362-9

5 Gong, Q. *et al.* Morbidity and mortality after lifestyle intervention for people with impaired glucose tolerance: 30-year results of the Da Qing Diabetes Prevention Outcome Study. *Lancet Diabetes Endocrinol* **7**, 452-461 (2019). https://doi.org/10.1016/s2213-8587(19)30093-2

6 DeFronzo, R. A. *et al.* Pioglitazone for diabetes prevention in impaired glucose tolerance. *N Engl J Med* **364**, 1104-1115 (2011). https://doi.org/10.1056/NEJMoa1010949

7 Herman, W. H. Prediabetes Diagnosis and Management. *JAMA* **329**, 1157-1159 (2023). https://doi.org/10.1001/jama.2023.4406

8 Roberts, S. *et al.* Preventing type 2 diabetes: systematic review of studies of cost-effectiveness of lifestyle programmes and metformin, with and without screening, for pre-diabetes. *BMJ Open* **7**, e017184 (2017). https://doi.org/10.1136/bmjopen-2017-017184

9 Piller, C. Dubious diagnosis. *Science* **363**, 1026-1031 (2019). https://doi.org/doi:10.1126/science.363.6431.1026

10 Shang, Y. *et al.* Natural history of prediabetes in older adults from a population-based longitudinal study. *J Intern Med* **286**, 326-340 (2019). https://doi.org/10.1111/joim.12920

11 Phillips, L. S., Ratner, R. E., Buse, J. B. & Kahn, S. E. We can change the natural history of type 2 diabetes. *Diabetes Care* **37**, 2668-2676 (2014). https://doi.org/10.2337/dc14-0817

12 Ferrannini, E. Definition of intervention points in prediabetes. *Lancet Diabetes Endocrinol* **2**, 667-675 (2014). https://doi.org/10.1016/s2213-8587(13)70175-x

13 Yokota, N. *et al.* Predictive models for conversion of prediabetes to diabetes. *Journal of Diabetes and its Complications* **31**, 1266-1271 (2017). https://doi.org/10.1016/j.jdiacomp.2017.01.005

14 Cahn, A. *et al.* Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. *Diabetes Metab Res Rev* **36**, e3252 (2020). https://doi.org/10.1002/dmrr.3252

15 Liang, K. *et al.* Nomogram Predicting the Risk of Progression from Prediabetes to Diabetes After a 3-Year Follow-Up in Chinese Adults. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* **Volume 14**, 2641-2649 (2021). https://doi.org/10.2147/dmso.S307456

16 Merino, J. *et al.* Metabolomics insights into early type 2 diabetes pathogenesis and detection in individuals with normal fasting glucose. *Diabetologia* **61**, 1315-1324 (2018). https://doi.org/10.1007/s00125-018-4599-x

17 Peddinti, G. *et al.* Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia* **60**, 1740-1750 (2017). https://doi.org/10.1007/s00125-017-4325-0

18 Rebholz, C. M. *et al.* Serum metabolomic profile of incident diabetes. *Diabetologia* **61**, 1046-1054 (2018). https://doi.org/10.1007/s00125-018-4573-7

19 Ren, M. *et al.* Potential Novel Serum Metabolic Markers Associated With Progression of Prediabetes to Overt Diabetes in a Chinese Population. *Front Endocrinol (Lausanne)* **12**, 745214 (2021). https://doi.org/10.3389/fendo.2021.745214

20 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015). https://doi.org/10.1371/journal.pmed.1001779

21 Allen, N. *et al.* UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology* **1**, 123-126 (2012).

22 ElSayed, N. A. *et al.* 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes-2023. *Diabetes Care* **46**, S19-s40 (2023). https://doi.org/10.2337/dc23-S002

23 Würtz, P. *et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *Am J Epidemiol* **186**, 1084-1096

(2017). https://doi.org/10.1093/aje/kwx016

24 Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* **8**, 192-206 (2015). https://doi.org/10.1161/circgenetics.114.000216

25 Zhang, X. *et al.* Plasma metabolomic profiles of dementia: a prospective study of 110,655 participants in the UK Biobank. *BMC Med* **20**, 252 (2022). https://doi.org/10.1186/s12916-022-02449-3

26 Geng, T.-T. *et al.* Nuclear Magnetic Resonance–Based Metabolomics and Risk of CKD. *American Journal of Kidney Diseases* (2023). https://doi.org/10.1053/j.ajkd.2023.05.014

27 Liang, Y. Y. *et al.* Association of Social Isolation and Loneliness With Incident Heart Failure in a Population-Based Cohort Study. *JACC Heart Fail* **11**, 334-344 (2023). https://doi.org/10.1016/j.jchf.2022.11.028

28 Wang, X. *et al.* Joint association of loneliness and traditional risk factor control and incident cardiovascular disease in diabetes patients. *Eur Heart J* (2023). https://doi.org/10.1093/eurheartj/ehad306

29 Song, Y. *et al.* Social isolation, loneliness, and incident type 2 diabetes mellitus: results from two large prospective cohorts in Europe and East Asia and Mendelian randomization. *EClinicalMedicine* **64**, 102236 (2023). https://doi.org/10.1016/j.eclinm.2023.102236

30 Klau, S., Jurinovic, V., Hornung, R., Herold, T. & Boulesteix, A. L. Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics* **19**, 322 (2018). https://doi.org/10.1186/s12859-018-2344-6

31 Morgenstern, J. D., Rosella, L. C., Costa, A. P., de Souza, R. J. & Anderson, L. N. Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology. *Adv Nutr* **12**, 621-631 (2021). https://doi.org/10.1093/advances/nmaa183

32 Qiu, X. *et al.* A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy. *Front Oncol* **10**, 551420 (2020). https://doi.org/10.3389/fonc.2020.551420

33 Rahman, S. A. *et al.* The AUGIS Survival Predictor: Prediction of Long-Term and Conditional Survival After Esophagectomy Using Random Survival Forests. *Ann Surg* **277**, 267-274 (2023). https://doi.org/10.1097/sla.0000000000004794

34 Kwak, S. *et al.* Markers of Myocardial Damage Predict Mortality in Patients With Aortic Stenosis. *J Am Coll Cardiol* **78**, 545-558 (2021). https://doi.org/10.1016/j.jacc.2021.05.047

35 Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics* **2**, 841-860, 820 (2008).

36 Janitza, S. & Hornung, R. On the overestimation of random forest's out-of-bag error. *PLoS One* **13**, e0201904 (2018). https://doi.org/10.1371/journal.pone.0201904

37 Tian, D. *et al.* Machine Learning-Based Prognostic Model for Patients After Lung Transplantation. *JAMA Netw Open* **6**, e2312022 (2023). https://doi.org/10.1001/jamanetworkopen.2023.12022

38 Wilson, P. W. *et al.* Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med* **167**, 1068-1074 (2007). https://doi.org/10.1001/archinte.167.10.1068

39 Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* **3**, 18 (2019). https://doi.org/10.1186/s41512-019-0064-7

40 Fan, X. *et al.* Noninvasive radiomics model reveals macrophage infiltration in glioma. *Cancer Lett* **573**, 216380 (2023). https://doi.org/10.1016/j.canlet.2023.216380

41 Bragg, F. *et al.* Predictive value of circulating NMR metabolic biomarkers for type 2 diabetes risk in the UK Biobank study. *BMC Med* **20**, 159 (2022). https://doi.org/10.1186/s12916-022-02354-9

42 Mackey, R. H. *et al.* Lipoprotein particles and incident type 2 diabetes in the multi-ethnic study of atherosclerosis. *Diabetes Care* **38**, 628-636 (2015). https://doi.org/10.2337/dc14-0645

43 Bragg, F. *et al.* The role of NMR-based circulating metabolic biomarkers in development and risk prediction of new onset type 2 diabetes. *Sci Rep* **12**, 15071 (2022). https://doi.org/10.1038/s41598-022-19159-8

44 Bragg, F. *et al.* Circulating Metabolites and the Development of Type 2 Diabetes in Chinese Adults. *Diabetes Care* **45**, 477-480 (2022). https://doi.org/10.2337/dc21-1415

45 Morze, J. *et al.* Metabolomics and Type 2 Diabetes Risk: An Updated Systematic Review and

Meta-analysis of Prospective Cohort Studies. *Diabetes Care* **45**, 1013-1024 (2022). https://doi.org/10.2337/dc21-1705

46 Lotta, L. A. *et al*. Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. *PLoS Med* **13**, e1002179 (2016). https://doi.org/10.1371/journal.pmed.1002179

47 Mahendran, Y. *et al*. Genetic evidence of a causal effect of insulin resistance on branched-chain amino acid levels. *Diabetologia* **60**, 873-878 (2017). https://doi.org/10.1007/s00125-017-4222-6

48 Li, J., Xi, F., Yu, W., Sun, C. & Wang, X. Real-Time Prediction of Sepsis in Critical Trauma Patients: Machine Learning-Based Modeling Study. *JMIR Form Res* **7**, e42452 (2023). https://doi.org/10.2196/42452

49 Li-Gao, R. *et al.* Assessment of reproducibility and biological variability of fasting and postprandial plasma metabolite concentrations using 1H NMR spectroscopy. *PLoS One* **14**, e0218549 (2019). https://doi.org/10.1371/journal.pone.0218549

Table 1. Baseline characteristics of participants with prediabetes stratified by incident diabetes status.

Characteristics	Overall (n = 13,489)	Diabetes (n = 2,525)	Non-diabetes (n = 10,964)	*P* value
Age, years	59.6 (7.1)	59.7 (7.1)	59.6 (7.0)	0.347
Male	6,166 (45.7)	1,407 (55.7)	4,759 (43.4)	<0.001
Education				<0.001
College or university	3,409 (25.3)	498 (19.7)	2,911 (26.6)	
Others	10,056 (74.5)	2,022 (80.1)	8,034 (73.3)	
Unknown	24 (0.2)	5 (0.2)	19 (0.2)	
Ethnicity				0.013
White	12,172 (90.2)	2,239 (88.7)	9,933 (90.6)	
Others	1,293 (9.6)	281 (11.1)	1,012 (9.2)	
Unknown	24 (0.2)	5 (0.2)	19 (0.2)	
Employment status				<0.001
Working	6,608 (49.0)	1,172 (46.4)	5,436 (49.6)	
Retired	5,931 (44.0)	1,114 (44.1)	4,817 (43.9)	
Other	787 (5.8)	212 (8.4)	575 (5.2)	
Unknown	163 (1.2)	27 (1.1)	136 (1.2)	
Household income				<0.001
Low	3,529 (26.2)	2,734 (24.9)	795 (31.5)	
Medium	5,659 (42.0)	4,666 (42.6)	993 (39.3)	
High	1,897 (14.1)	1,611 (14.7)	286 (11.3)	
Unknown	2,404 (17.8)	1,953 (17.8)	451 (17.9)	
Townsend Deprivation Index	-1.0 (3.3)	-0.7 (3.4)	-1.1 (3.2)	<0.001
Family history of DM	3,068 (22.7)	786 (31.1)	2,282 (20.8)	<0.001
History of CVD	1,392 (10.3)	413 (16.4)	979 (8.9)	<0.001
History of hypertension	4,217 (31.3)	985 (39.0)	3,232 (29.5)	<0.001
History of dyslipidemia	1,932 (14.3)	417 (16.5)	1,515 (13.8)	0.001
History of CLD	1,847 (13.7)	413 (16.4)	1,434 (13.1)	<0.001
History of cancer				0.056
Yes	1,315 (9.7)	215 (8.5)	1,100 (10.0)	
No	12,171 (90.2)	2,309 (91.4)	9,862 (89.9)	
Unknown	3 (0.0)	1 (0.0)	2 (0.0)	
BMI, kg/m^2	29.0 (5.2)	31.3 (5.3)	28.4 (5.0)	<0.001
WC, cm	94.6 (13.5)	101.3 (13.1)	93.1 (13.1)	<0.001
HC, cm	105.4 (10.0)	108.6 (10.8)	104.6 (9.7)	<0.001
Smoking status (%)				<0.001
Never	6,478 (48.0)	1,104 (43.7)	5,374 (49.0)	
Previous	4,843 (35.9)	1,003 (39.7)	3,840 (35.0)	
Current	2,074 (15.4)	397 (15.7)	1,677 (15.3)	
Unknown	94 (0.7)	21 (0.8)	73 (0.7)	
Moderate alcohol				0.081
Yes	3,888 (28.8)	689 (27.3)	3,199 (29.2)	
No	9,595 (71.1)	1,836 (72.7)	7,759 (70.8)	
Unknown	6 (0.0)	0 (0.0)	6 (0.1)	
Healthy diet score	3.3 (1.1)	3.2 (1.1)	3.3 (1.1)	<0.001
Healthy sleep score	3.5 (1.0)	3.3 (1.1)	3.6 (1.0)	<0.001
Physical activity, METs	10.4 (4.9)	9.7 (5.1)	10.6 (4.9)	<0.001
SBP, mmHg	141.3 (18.5)	143.5 (18.2)	140.8 (18.5)	<0.001
DBP, mmHg	83.3 (10.2)	84.6 (10.4)	83.0 (10.1)	<0.001
HbA1c, %	5.9 (0.2)	6.0 (0.2)	5.9 (0.2)	<0.001

Data were presented as means (standard deviations, SDs) for continuous variables and numbers (percentages) for categorical variables.

BMI, body mass index; DM, diabetes mellitus; CVD, cardiovascular disease; CLD, chronic lung disease; DBP, diastolic blood pressure; HbA1c, glycated hemoglobin A1c; HC, hip circumference; MET, metabolic equivalent of task; SBP, systolic blood pressure; WC, waist circumference.

Table 2. Performance of Cox proportional hazards regression models in prediction of the progression of prediabetes to diabetes.

Performance metric	Basic model[a]	Basic model + 9 metabolites[b]	*P* value
AUROC			
T=1-year	0.759 (0.608, 0.911)	0.823 (0.702, 0.945)	0.009
T=5-year	0.798 (0.762, 0.834)	0.830 (0.797, 0.864)	<0.001
T=10-year	0.776 (0.750, 0.801)	0.801 (0.778, 0.825)	<0.001
Continuous NRI			
T=1-year	Reference	0.461 (0.134, 0.660)	<0.001
T=5-year	Reference	0.400 (0.277, 0.483)	<0.001
T=10-year	Reference	0.329 (0.252, 0.405)	<0.001
Absolute IDI			
T=1-year	Reference	0.006 (-0.002, 0.020)	0.132
T=5-year	Reference	0.028 (0.017, 0.040)	<0.001
T=10-year	Reference	0.040 (0.027, 0.054)	<0.001

[a]Basic model: age, sex, Townsend Deprivation Index, family history of diabetes mellitus, body mass index, Waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, and glycated hemoglobin A1c.
[b]The selected 9 metabolic biomarkers: cholesteryl esters in large HDL, triglycerides in very large VLDL, Glycine, average diameter for LDL particles, tyrosine, cholesteryl esters in medium VLDL, glucose, triglycerides in IDL, docosahexaenoic acid.
AUROC, area under the receiver operating characteristic curve; HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; IDI, absolute integrated discrimination improvement; LDL, low-density lipoprotein; NRI, net reclassification improvement; VLDL, very-low-density lipoprotein.

Figure legends

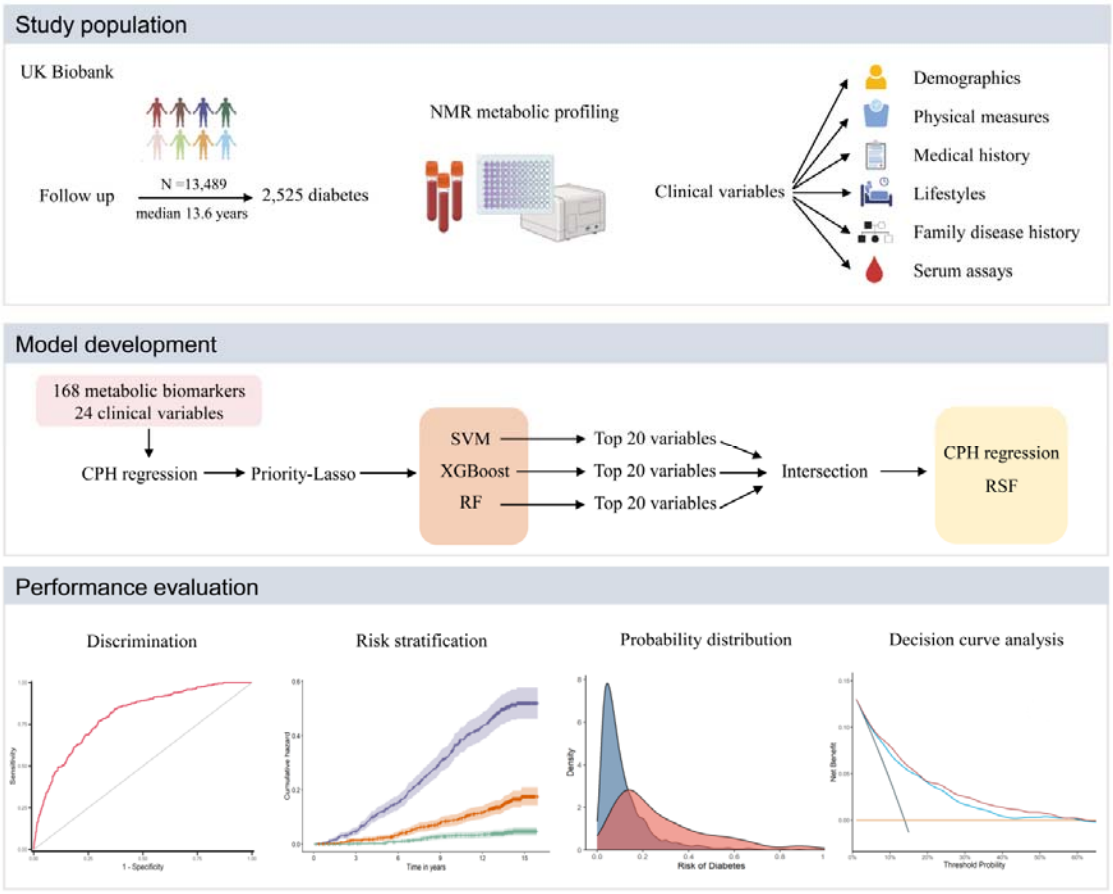

Figure 1. Overall schematic workflow of the study.

CPH, Cox proportional hazard; NMR, nuclear magnetic resonance; RF, random forest; RSF, Random survival forest; SVM, supporting vector machine; XGBoost, extreme gradient boosting.

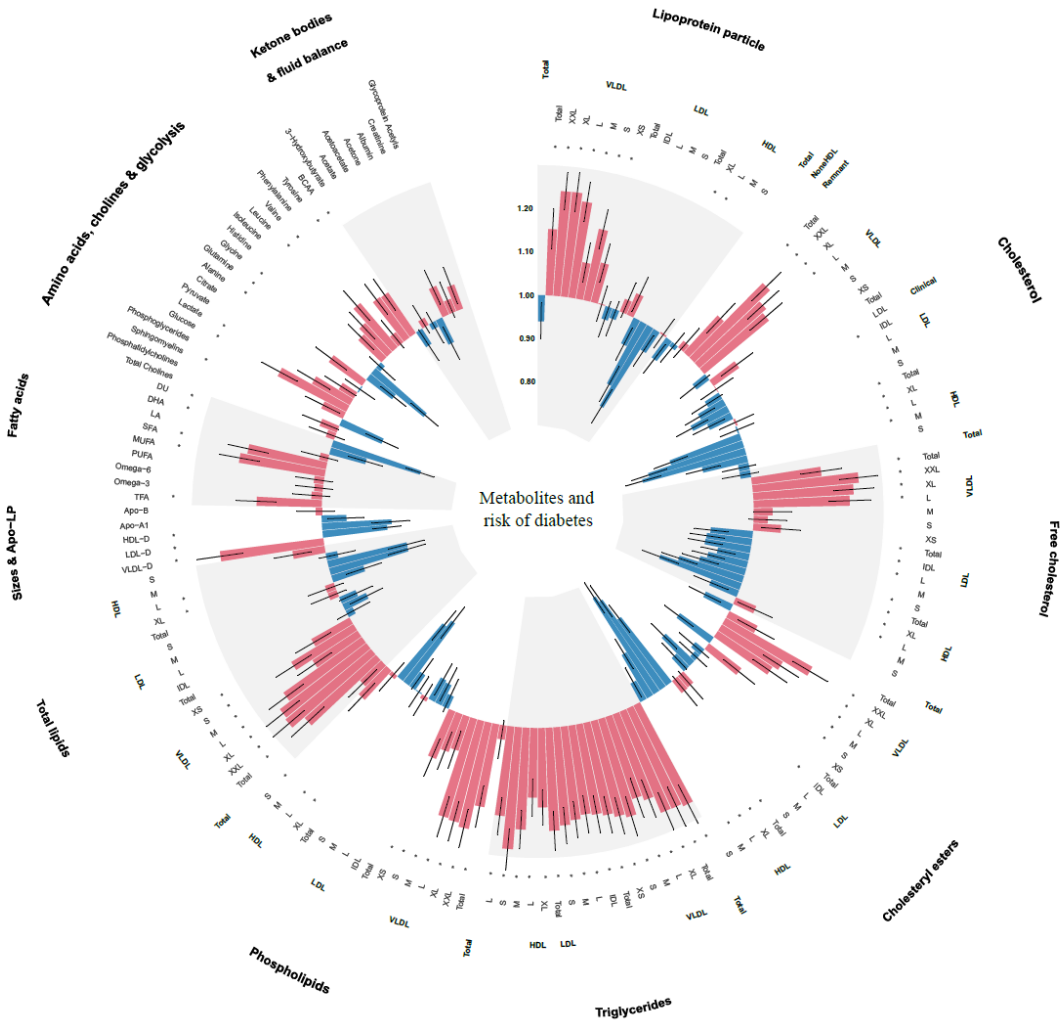

Figure 2. Associations of 168 metabolic biomarkers with risk of diabetes among 13,489 participants with prediabetes.

Hazard ratios (HR) were presented per 1 standard deviation (SD) higher of metabolic biomarker on the natural log scale and were adjusted for age, sex, ethnicity, education, Townsend Deprivation Index, employment status, household income, family history of diabetes, history of CVD, history of hypertension, history of dyslipidemia, history of CLD, history of cancer, body mass index, waist circumference, hip circumference, smoking status, moderate alcohol, healthy diet score, healthy sleep score, physical activity, systolic blood pressure, diastolic blood pressure and glycated hemoglobin A1c. *False discovery rate

controlled $P < 0.05/168$.

Apo-A1, apolipoprotein A1; Apo-B, apolipoprotein B; Apo-LP, apolipoprotein; BCAA, branched-chain amino acid; BMI, body mass index; CVD, cardiovascular disease; CLD, chronic lung disease; DHA, docosahexaenoic acid; FA, fatty acids; HDL, high-density lipoproteins; HDL-D, high-density lipoprotein particle diameter; IDL, intermediate-density lipoproteins; L, large; LA, linoleic acid; LDL, low-density lipoproteins; LDL-D, low-density lipoprotein particle diameter; LP, lipoprotein; M, medium; MUFA, monounsaturated fatty acids; PUFA, polyunsaturated fatty acids; S, small; SFA, saturated fatty acids; VLDL, very-low-density lipoproteins; VLDL-D, very-low-density lipoprotein particle diameter; XL, very large; XS, very small; XXL, extremely large.

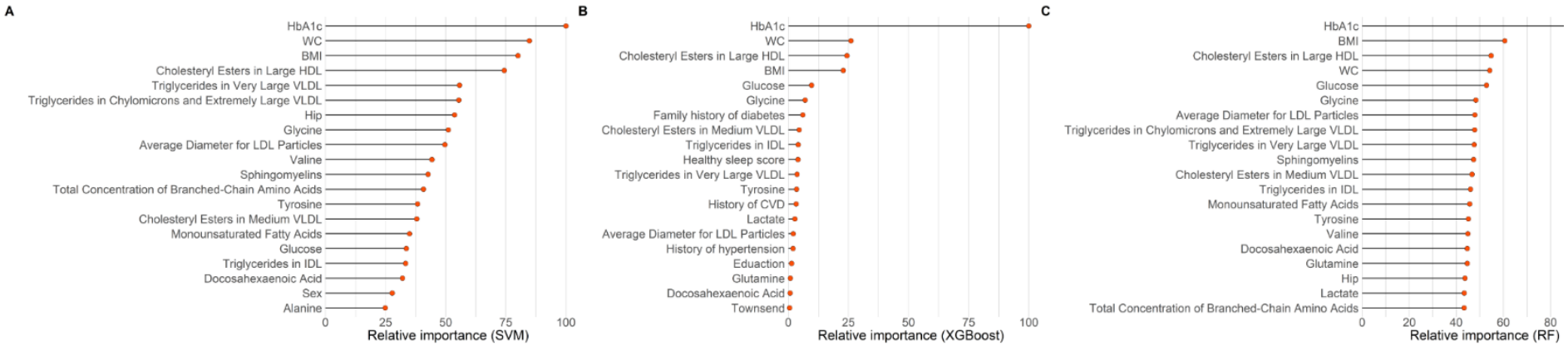

Figure 3. The top 20 important variables selected by three machine learning models: (A) supporting vector machine (SVM); (B) extreme gradient boosting (XGBoost); (C) random forest (RF).

The models were adjusted for age, sex, ethnicity, education, Townsend Deprivation Index, employment status, household income, family history of diabetes, history of CVD, history of hypertension, history of dyslipidemia, history of CLD, history of cancer, body mass index, waist circumference, hip circumference, smoking status, moderate alcohol, healthy diet score, healthy sleep score, physical activity, systolic blood pressure, diastolic blood pressure and glycated hemoglobin A1c. CVD, cardiovascular disease; CLD, chronic lung disease. HDL, high-density lipoproteins; IDL, intermediate-density lipoproteins; LDL, low-density lipoproteins; VLDL, very-low-density lipoproteins.

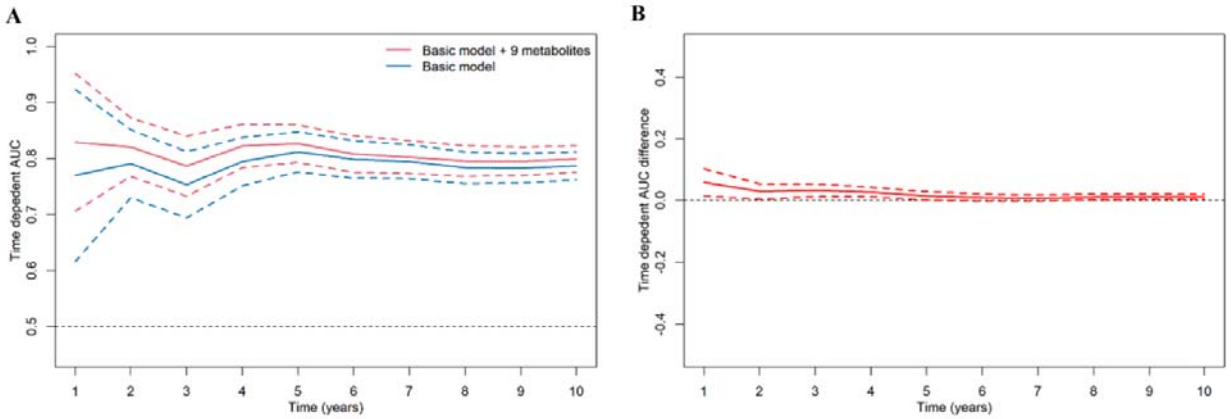

Figure 4. Consecutive area under time-dependent receiver-operating characteristic (AUROC) of basic model and basic model plus 9 metabolites (A), and the difference of these two time-dependent AUROCs over time (B).

The basic model used conventional clinical variables including age, sex, Townsend Deprivation Index, family history of diabetes mellitus, body mass index, Waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, and glycated hemoglobin A1c. The selected 9 metabolites included cholesteryl esters in large HDL, triglycerides in very large VLDL, Glycine, average diameter for LDL particles, tyrosine, cholesteryl esters in medium VLDL, glucose, triglycerides in IDL, and docosahexaenoic acid.

HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; VLDL, very-low-density lipoprotein.

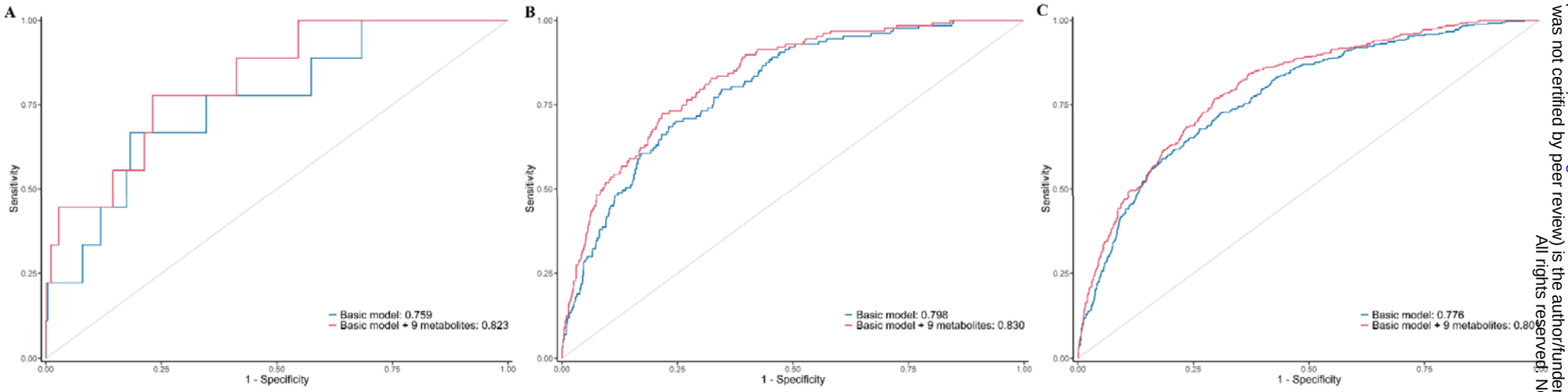

Figure 5. Time-dependent receiver-operating characteristic (ROC) curves of basic model and basic model plus 9 metabolites for predicting 1-year (A), 5-year (B), and 10-year (C) risk of developing diabetes in participants with prediabetes.

The basic model used conventional clinical variables including age, sex, Townsend Deprivation Index, family history of diabetes mellitus, body mass index, Waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, and glycated hemoglobin A1c. The selected 9 metabolites included cholesteryl esters in large HDL, triglycerides in very large VLDL, Glycine, average diameter for LDL particles, tyrosine, cholesteryl esters in medium VLDL, glucose, triglycerides in IDL, and docosahexaenoic acid.

HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; VLDL, very-low-density lipoprotein.

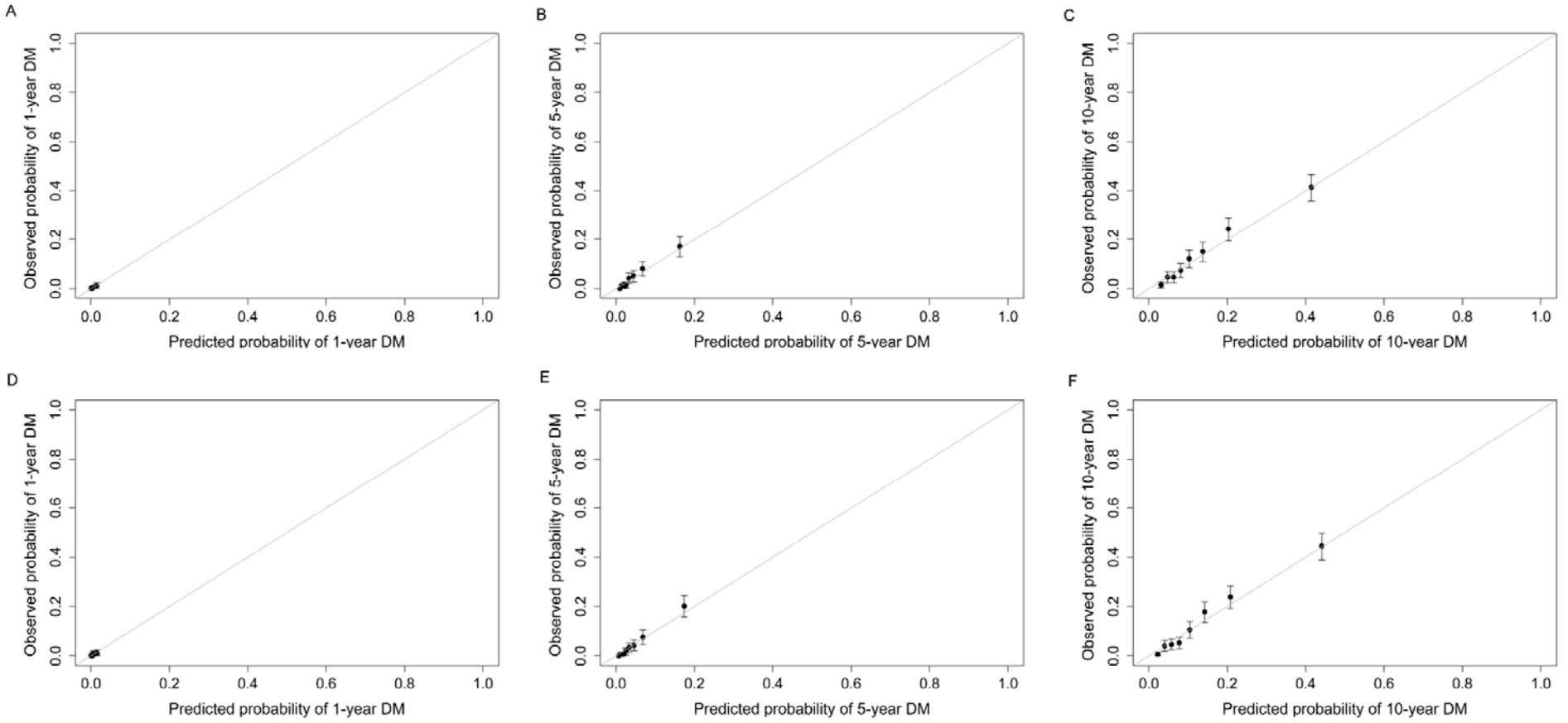

Figure 6. Calibration plots of basic model (A, B, C) and basic model plus 9 metabolites (D, E, F) for predicting 1- year, 5-year, and 10-year risk of developing diabetes in participants with prediabetes.

The basic model used conventional clinical variables including age, sex, Townsend Deprivation Index, family history of diabetes mellitus, body mass index, Waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, and glycated hemoglobin A1c. The selected 9 metabolites included cholesteryl esters in large HDL, triglycerides in very large VLDL, Glycine, average diameter for LDL particles, tyrosine, cholesteryl esters in medium VLDL, glucose, triglycerides in IDL, and docosahexaenoic acid.

HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; VLDL, very-low-density lipoprotein.

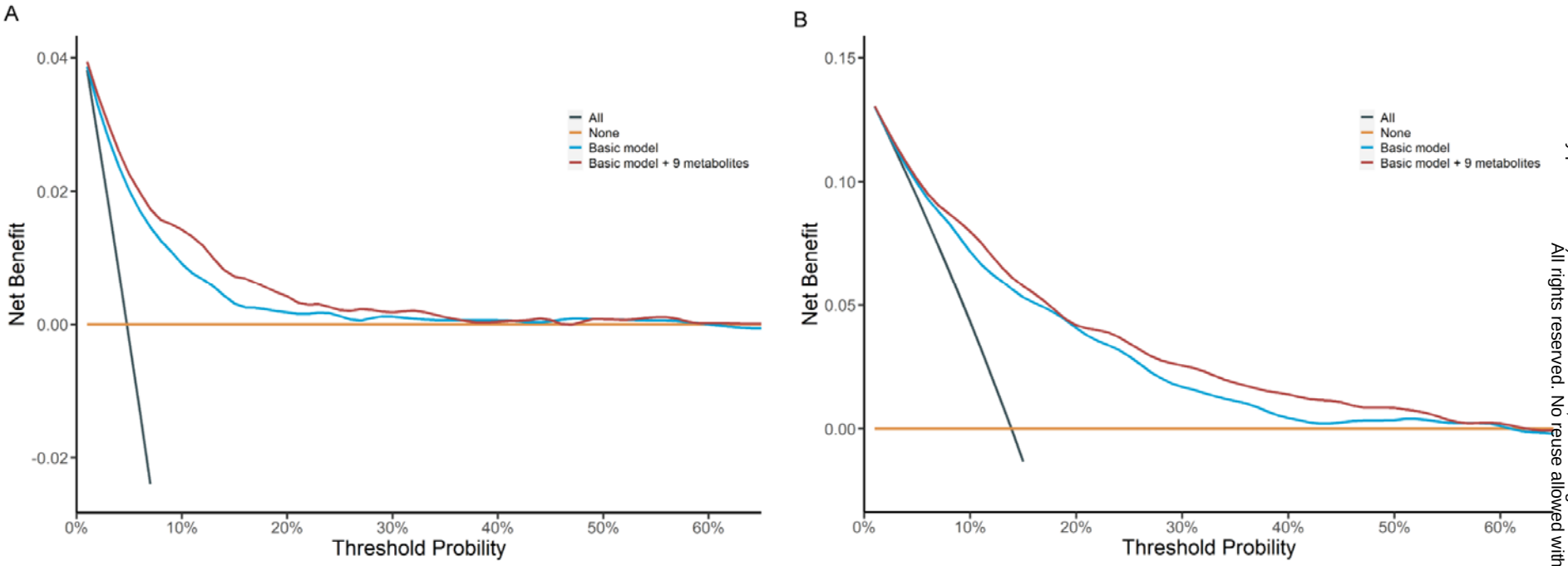

Figure 7. Decision curve analysis of basic model and basic model plus 9 metabolites for predicting 5-year (A) and 10-year (B) risk of developing diabetes in participants with prediabetes.

Decision curve analysis was not performed on 1-year prediction considering the relatively small number of prediabetic patients who develop diabetes within a year in the test set and small net benefit from intervention. The basic model used conventional clinical variables including age, sex, Townsend Deprivation Index, family history of diabetes mellitus, body mass index, Waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, and glycated hemoglobin A1c. The selected 9 metabolites included cholesteryl esters in large HDL,

triglycerides in very large VLDL, Glycine, average diameter for LDL particles, tyrosine, cholesteryl esters in medium VLDL, glucose, triglycerides in IDL, and docosahexaenoic acid.

HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; VLDL, very-low-density lipoprotein.

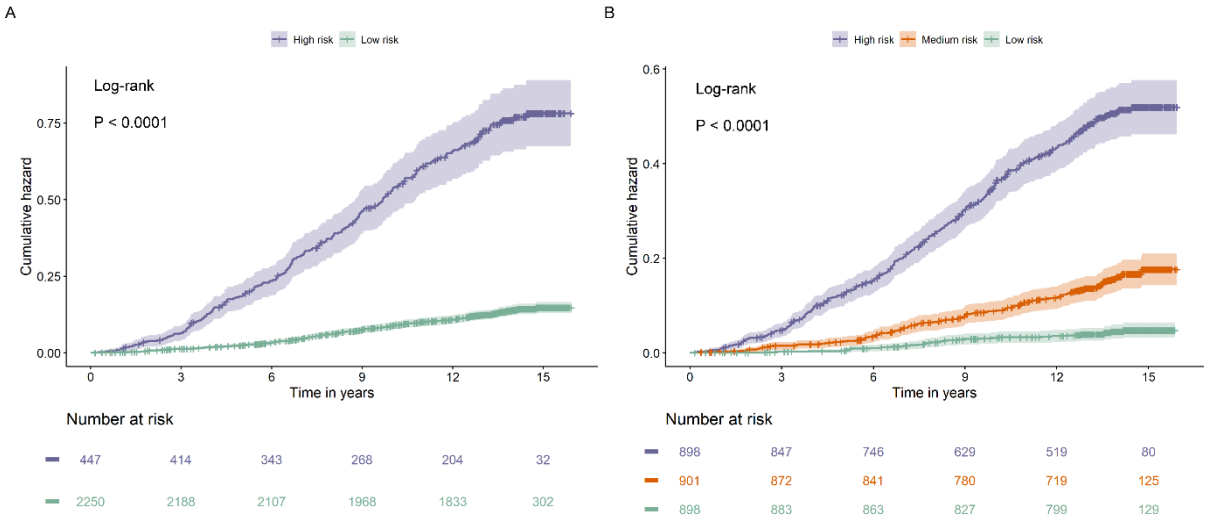

Figure 8. Cumulative hazard curves for participants with prediabetes with different risks stratified by the Cox model based on clinical variables and 9 metabolites.

The Cox model divided participants with prediabetes in the test set to two categories (A) and three categories (B) with significant differences in cumulative hazard of diabetes during the follow-up (both P <0.0001).

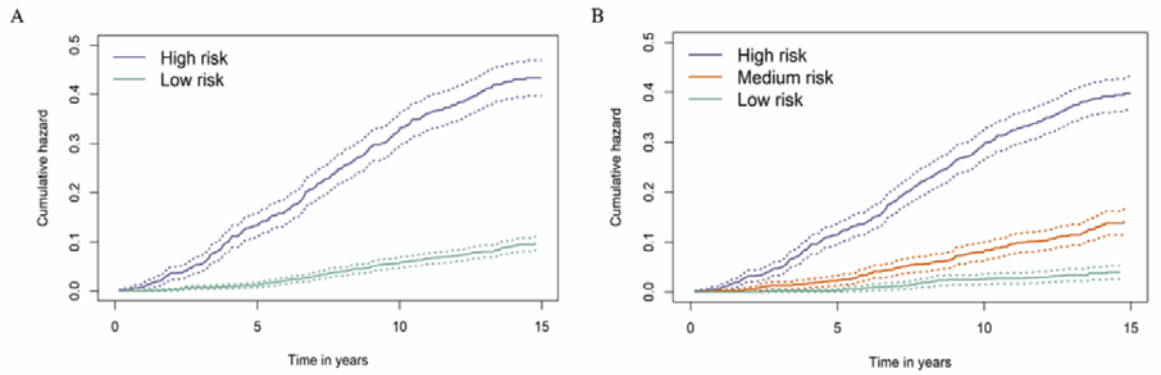

Figure 8-figure supplement 1. Cumulative hazard curves for participants with prediabetes with different risks stratified by the Cox model based on clinical variables and 9 metabolites when considering competing risk from death.

The Cox model divided participants with prediabetes in the test set to two categories (A) and three categories (B) with significant differences in cumulative hazard of diabetes during the follow-up (both Fine-Gray P <0.0001).

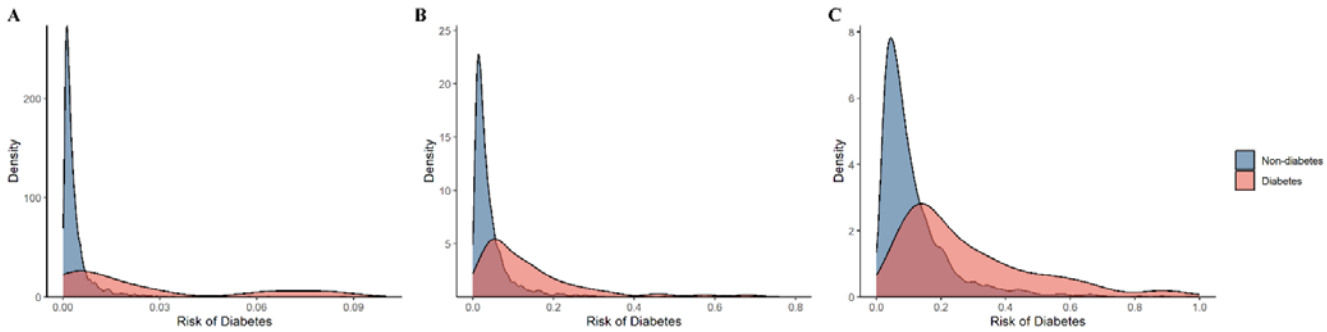

Figure 9. The distribution of the predictive probability of developing diabetes among participants with prediabetes by incident diabetes status within 1- year (A), 5-year (B), and 10-year (C).

Additional files

Supplementary file 1. List of 168 NMR-based metabolomic biomarkers in the UK Biobank.

HDL, high-density lipoproteins; IDL, intermediate-density lipoproteins; LDL, low-density lipoproteins; VLDL, very-low-density lipoproteins.

Supplementary file 2. Associations of 168 metabolic biomarkers with risk of diabetes among 13,489 participants with prediabetes.

Hazard ratios (HR) were presented per 1 standard deviation (SD) higher of metabolic biomarker on the natural log scale and were adjusted for age, sex, ethnicity, education, Townsend Deprivation Index, employment status, household income, family history of diabetes, history of CVD, history of hypertension, history of dyslipidemia, history of CLD, history of cancer, body mass index, waist circumference, hip circumference, smoking status, moderate alcohol, healthy diet score, healthy sleep score, physical activity, systolic blood pressure, diastolic blood pressure and glycated hemoglobin A1c. *P* value< 0.05/168 were highlighted in bold.

Apo-A1, apolipoprotein A1; Apo-B, apolipoprotein B; Apo-LP, apolipoprotein; BMI, body mass index; CVD, cardiovascular disease; CLD, chronic lung disease; DHA, docosahexaenoic acid; FA, fatty acids; HDL, high-density lipoproteins; HDL-D, high-density lipoprotein particle diameter; IDL, intermediate-density lipoproteins; L, large; LA, linoleic acid; LDL, low-density lipoproteins; LDL-D, low-density lipoprotein particle diameter; LP, lipoprotein; M, medium; MUFA, monounsaturated fatty acids; PUFA, polyunsaturated fatty acids; S, small; SFA, saturated fatty acids; VLDL, very-low-density lipoproteins; VLDL-D, very-low-density lipoprotein particle diameter; XL, very large; XS, very small; XXL, extremely large.

Supplementary file 3. Coefficients of the selected 17 metabolites by priority-Lasso.

HDL, high-density lipoproteins; IDL, intermediate-density lipoproteins; LDL, low-density lipoproteins; VLDL, very-low-density lipoproteins.

Supplementary file 4. Associations of the selected 9 metabolites with risk of diabetes among 13,489 participants with prediabetes after adjusting for conventional clinical variables.

Hazard ratios (HR) were presented per 1 standard deviation (SD) higher of metabolic biomarker on the natural log scale and were adjusted for age, sex, Townsend Deprivation Index, family history of diabetes, body mass index, waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, and glycated hemoglobin A1c.

HDL, high-density lipoproteins; IDL, intermediate-density lipoproteins; LDL, low-density lipoproteins; VLDL, very-low-density lipoproteins.

Supplementary file 5. Performance of Cox proportional hazards prediction models for the risk of diabetes among participants with normoglycemia.

[a]Basic model: age, sex, Townsend Deprivation Index, family history of diabetes mellitus, body mass index, Waist circumference, hip circumference, systolic blood pressure, diastolic blood pressure, and glycated hemoglobin A1c.

[b]The selected 9 metabolic biomarkers: cholesteryl esters in large HDL, triglycerides in very large VLDL, Glycine, average diameter for LDL particles, tyrosine, cholesteryl esters in medium VLDL, glucose, triglycerides in IDL, docosahexaenoic acid.

AUROC, area under the receiver operating characteristic curve; HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; IDI, absolute integrated discrimination improvement; LDL, low-density lipoprotein; NRI, net reclassification improvement; VLDL, very-low-density lipoprotein.