Machine learning approaches revealed metabolic signatures of incident chronic kidney disease in persons with pre-and type 2 diabetes

Jialing Huang[1,2,3], Cornelia Huth[2,3], Marcela Covic[1,2,3], Martina Troll[1,2], Jonathan Adam[1,2], Sven Zukunft[4,5], Cornelia Prehn[4], Li Wang[1,2,6], Jana Nano[2,3], Markus F. Scheerer[7,8], Susanne Neschen[7,9], Gabi Kastenmüller[10], Karsten Suhre[11], Michael Laxy[12], Freimut Schliess[13], Christian Gieger[1,2,3], Jerzy Adamski[4,14,15], Martin Hrabe de Angelis[3, 7,15], Annette Peters[2,3], Rui Wang-Sattler[1,2,3,*]

[1] Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.
[2] Institute of Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
[3] German Center for Diabetes Research (DZD), Ingolstädter Landstraße 1, 85764 München-Neuherberg, Germany.
[4] Research Unit of Molecular Endocrinology and Metabolism, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.
[5] Current address: Institute for Vascular Signalling, Centre for Molecular Medicine, Goethe University, Frankfurt am Main, Germany
[6] Liaocheng People's Hospital - Shandong University Postdoctoral Work Station, Department of Scientific Research, P.R.China.
[7] Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.
[8] Current address: Bayer AG, Medical Affairs & Pharmacovigilance, Müllerstr. 178, 13353 Berlin, Germany.
[9] Current address: Sanofi Aventis Deutschland GmbH, Germany.
[10] Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany.
[11] Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar (WCMC-Q), PO Box 24144, Education City - Qatar Foundation, Doha, Qatar.
[12] Institute of Health Economics and Health Care Management, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.
[13] Profil GmbH 41460 Neuss, Germany
[14] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597, Singapore
[15] Chair of Experimental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, 85353 Freising, Germany

Diabetes Publish Ahead of Print, published online October 6, 2020

*Corresponding author. Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, 85764 Neuherberg, Germany. Tel.: + 49 89 3187 3978; Fax: + 49 89 3187 2428; E-mail: rui.wang-sattler@helmholtz-muenchen.de

ABSTRACT

Early and precise identification of individuals with pre-diabetes and type 2 diabetes (T2D) at risk of progressing to chronic kidney disease (CKD) is essential to prevent complications of diabetes. Here, we identify and evaluate prospective metabolite biomarkers and the best set of predictors of CKD in the longitudinal, population-based Cooperative Health Research in the Region of Augsburg (KORA) cohort by targeted metabolomics and machine learning approaches. Out of 125 targeted metabolites, sphingomyelin (SM) C18:1 and phosphatidylcholine diacyl (PC aa) C38:0 were identified as candidate metabolite biomarkers of incident CKD specifically in hyperglycemic individuals followed during 6.5 years. Sets of predictors for incident CKD developed from 125 metabolites and 14 clinical variables showed highly stable performances in all three machine learning approaches and outperformed the currently established clinical algorithm for CKD. The two metabolites in combination with five clinical variables were identified as the best set of predictors and their predictive performance yielded a mean area value under the receiver operating characteristic curve of 0.857. The inclusion of metabolite variables in the clinical prediction of future CKD may thus improve the risk prediction in persons with pre- and T2D. The metabolite link with hyperglycemia-related early kidney dysfunction warrants further investigation.

KEYWORDS

Chronic kidney disease, complications of diabetes, metabolite biomarkers, metabolomics, machine learning algorithms, improved prediction of incident chronic kidney disease

INTRODUCTION

Chronic kidney disease (CKD) affects about 9.1% of the general population worldwide (1). From 1990 to 2017, the global all-age mortality rate due to CKD increased by 41.5%, resulting in 1.2 million deaths only in 2017 (1).

Among the established risk factors for CKD, diabetes mellitus accounts for 30%-50% of all CKD cases (2) and its microvascular complication, diabetic nephropathy, is the leading cause of end-stage kidney disease (3). Moreover, undiagnosed diabetes and pre-diabetes have been related with high prevalence of CKD in US, European and Asian populations (4-7). Early screening of hyperglycemic individuals under risk of developing CKD is therefore crucial for effective prevention and management of incident CKD in the framework of an integrated personalized diabetes management (8).

Increased urinary albumin-to-creatinine ratio (UACR) and reduced estimated glomerular filtration rate (eGFR) are two clinical biomarkers of kidney-related structural damage and functional decline used to diagnose CKD (9). UACR, eGFR, age and sex were reported to be highly predictive for progression of CKD (10). Albuminuria and eGFR were also found to be the most important variables to predict onset and progression of early CKD in individuals with type 2 diabetes (T2D). However, their predictive ability was modest with an externally validated c-statistic of 0.68 even when combined with age and sex (11). Since the traditional risk factors for CKD are insufficient for reliable prediction of CKD in individuals with T2D, there is an urgent

need for more sensitive and specific biomarkers for CKD prognosis in (pre-) diabetes management.

A comprehensive individual profiling by means of metabolomics is a promising approach to discover previously unconsidered associations between metabolic signatures and clinical outcomes such as obesity, pre-diabetes and T2D (12-19). Several studies investigated the metabolite profiles of CKD, both in the general and T2D population (20-22). However, to the best of our knowledge, none of them had explored the metabolites associated with future development of CKD in persons with (pre-) T2D.

In this study, we applied priority-Lasso and multivariate logistic regression (MLR) to identify metabolites associated with incident CKD in the population-based adult cohort KORA (Cooperative Health Research in the Region of Augsburg) (23; 24). Using three machine learning approaches (support vector machine (SVM), random forest (RF), adaptive boosting (AdaBoost)), we furthermore assessed the predictive power of predictor sets constructed with metabolites and clinical phenotypes and compared their performance with the typically used clinical algorithm for CKD. We finally presented the best set of predictors for incident CKD in individuals with (pre-) T2D.

RESEARCH DESIGN AND METHODS

Study design and participants

We investigated the two follow-ups of the longitudinal cohort KORA survey 4 conducted in the area of Augsburg, Southern Germany. The first follow-up (F4) involved 3,080 individuals (aged 32–81 years) examined between 2006 and 2008. The second follow-up (FF4) examined 2,269 participants from 2013 to 2014 (23). Because the metabolomics data and the clinical variables of CKD (eGFR and UACR) were measured in the F4 study, we used F4 as baseline.

Individuals with hyperglycemia and normal glucose tolerance (NGT) were classified according to baseline fasting and two hour post load glucose (2-h glucose) values using the World Health Organization diagnostic criteria (25). Hyperglycemic group comprised participants with pre-diabetes and newly diagnosed T2D (i.e., fasting glucose ≥ 110 mg/dl and/or 2-h-glucose glucose ≥ 140 mg/dl), as well as known T2D that was diagnosed by physician validated self-reporting and/or current use of anti-diabetes agents (13; 23).

We examined 2,142 individuals who participated in both KORA F4 and FF4. Exclusion criteria were: 1) non-fasting samples (n = 5 at F4); 2) missing eGFR and UACR (n = 16 at F4, n = 64 at FF4) or covariate values (n = 19 at F4); 3) diagnosis for type 1 diabetes (n = 6 at F4), unclear type of diabetes mellitus (n = 21 at F4) or CKD (n = 173 at F4). The remaining dataset comprised 385 hyperglycemic participants and 1,453 individuals with NGT (Fig. 1, Table 1). The hyperglycemic participants were

used to identify candidate metabolite biomarkers for incident CKD and to develop and evaluate sets of metabolite and clinical predictors. The NGT participants were used for sensitivity analyses of candidate biomarkers.

All study participants gave written informed consent. The KORA study was approved by the ethics committee of the Bavarian Medical Association, Munich, Germany.

Outcome definition

The eGFR was calculated from serum creatinine (mg/dl) and cystatin-C (mg/dl) (IDMS and IFCC standardized values) using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation (26). Non-CKD was defined as an eGFR $\geq$ 60 ml/min/1.73 m^2 and an UACR < 30 mg/g at both F4 and FF4 (9). Incident cases of CKD consisted of participants that were non-CKD at baseline (F4) but had reduced kidney function (eGFR < 60 ml/min/1.73 m^2) and/or kidney damage (UACR $\geq$ 30 mg/g) at follow-up (FF4).

Metabolite quantification and normalization

The serum samples from participants in the KORA F4 study were measured with the Absolute*IDQ*™ p150 Kit (BIOCRATES Life Sciences AG, Innsbruck, Austria) (24; 27). In total, 3,061 serum samples of the F4 study were quantified for 163 metabolites in 38 randomly distributed kit plates (Table S1). Each plate also contained three quality

control (QC) samples (gender mixed human plasma provided by the manufacturer) and one zero sample (PBS).

Identical QC procedures were used (13). Each metabolite met two criteria: 1) average value of the coefficient of variance in the three QCs < 25%; 2) 50% of all measured sample concentrations above 3 times median of the 38 zero samples. In total, 125 metabolites passed the criteria and were used in the subsequent analysis (Table S1). To minimize the plate effect, metabolite concentrations were adjusted for the plate normalization factors (NFs). For each metabolite, the plate NFs were calculated by dividing the mean of QC sample values in each plate with the mean of all QC sample values in 38 plates. As shown in Fig. S1, plate normalization efficiently corrected the inter-plate variations in metabolite concentration.

To ensure comparability between different metabolites, their concentrations were natural-log transformed and scaled to a mean value of zero and standard deviation (SD) of one.

Three-step feature selection

Since feature reduction is an important aspect of predictive modeling we defined a three-step feature selection procedure.

In order to decrease the false positive rate of the final discovery, we firstly used MLR adjusted for the two sets of covariates based on medical knowledge (11). Basic model was adjusted for age, sex, BMI, systolic blood pressure (BP), smoking status,

triglyceride, total cholesterol, HDL cholesterol and fasting glucose. Full model was additionally adjusted for the use of lipid-lowering, antihypertensive and anti-diabetic medication, and for baseline eGFR and UACR (Fig. 1). Metabolites that were significantly associated with incident CKD in the full model ($P < 0.05$) were retained.

Secondly, we applied the machine learning method priority-Lasso to deal with multi-collinearity of included variables and to retain metabolite and clinical variables with non-zero coefficients. Priority-Lasso is a Lasso-based intuitive procedure that utilizes prior knowledge of the study outcome by defining the blocks of different types of predictor variables (28). We defined 14 clinical variables in the full model as block 1 whereas the metabolites retained after the first-step screen were defined as block 2. The penalization parameters λ in each block were determined as values with maximum area under the receiver operating characteristic curve (AUC) estimated in a 10-fold cross-validation.

Thirdly, we used logistic regression with backward stepwise selection according to the Akaike information criterion (*AIC*) to select for the most strongly associated variables with incident CKD and reduce model complexity (Fig. 1).

After the three-step feature selection, the selected metabolites from the 385 hyperglycemic individuals were regarded as candidate biomarkers.

Sensitivity analyses of candidate biomarkers

We conducted four sensitivity analyses to reduce the possibility of chance

findings (Fig. 1): 1) a nearest-neighbor propensity score matching in nested case-control study design was used to balance cases and controls on conventional risk factors of CKD. MLR analysis was used to generate propensity scores using incident CKD as outcome and covariates from the full model. The caliper was defined as 0.1. After 1-to-1 propensity score matching, we investigated the association of candidate biomarkers with incident CKD by conditional logistic regression; 2) we investigated whether the predictive effect of candidate biomarkers for incident CKD was dependent of the hyperglycemic status. We examined the association of the candidate biomarkers with incident CKD in 1453 normoglycemic participants by MLR; 3) we explored the interaction effects of candidate biomarkers with glucose levels for incident CKD in 1838 individuals and performed a stratified analysis by MLR. We next examined the multiplicative interaction effects between candidate biomarkers and glucose groups by adding related multiplicative terms in the MLR models. The significance of interaction terms was tested by ANOVA LRT-test; 4) we examined the association of candidate biomarkers with UACR-based (UACR $\geq$ 30 mg/g) and eGFR-based (eGFR $<$ 60 ml/min/1.73 m^2) incident CKD separately in hyperglycemic participants.

Development and evaluation of predictor sets

We performed the three-step feature selection with 100 random repeats of 10-fold cross-validation to develop the sets of metabolite and clinical predictors for incident CKD in hyperglycemia (Fig. 1). Their predictive performances were evaluated using

AUC. The AUC values of developed predictors were compared with the established prediction model consisting of age, sex, eGFR, and UACR (10; 11). These four clinical variables were used as reference predictors.

In each 10-fold cross-validation, the data from 385 hyperglycemic individuals were randomly partitioned into 10 non-overlapping subsets. Each of these 10 subsets was regarded in turn as testing data, whereas the remaining nine subsets were used as training data (Fig. 1). In each iteration, a set of metabolite and clinical variables for incident CKD was identified with the three-step feature selection procedure using one of the training datasets. The identified predictor set and the reference predictors were used to develop respective prediction models with SVM. In this way, two prediction models were built using one training dataset. The AUC values of respective two models were computed for the testing data only (Fig. 1). The average AUC value over 10 iterations of one 10-fold cross-validation was calculated and finally presented. In order to assess the robustness of the predictive results, the predictive models were furthermore built using another two machine learning approaches (i.e., RF and AdaBoost) and the corresponding AUC values were reported.

SVM models were fitted with the R e1071 package (29). The kernel parameter was defined as radial (i.e., Gaussian radial basis function). The corresponding parameters gamma and cost (i.e. cost of constraints violation) of radial basis kernel were defined as the values with the best performance estimation using 10-fold cross-validation with a grid search over supplied value ranges; RF models were fitted with

the R randomForest package, which implements Breiman's classic algorithm (30). The two RF parameters, nTree (i.e., the number of trees to grow for each forest) and mTry (i.e., the number of input variables randomly chosen at each split), were set to 600 and the default setting (floor of square root of the number of features), respectively. The R ada package was used to fit the AdaBoost models (31). The three AdaBoost parameters loss (i.e., loss function), type (i.e., type of boosting algorithm to perform) and iter (i.e., number of boosting iterations to perform) were set to ada (corresponding to the default boosting under exponential loss), discrete (discrete boosting) and 200, respectively.

In total, we performed 100 repeats of 10-fold cross-validations including 1000 times of three-step feature selection. The most frequently selected set of metabolites and clinical variables among these 1000 selection rounds was subsequently defined as the best set of predictors for incident CKD in hyperglycemia.

All statistical analyses were performed in R (version 3.5.0) and two-sided *P*-value < 0.05 was considered as statistically significant.

Data and resource availability

The KORA F4/FF4 data sets are not publicly available because of data protection agreements but can be provided upon request through the KORA-PASST (Project application self-service tool, www.helmholtz-muenchen.de/kora-gen).

RESULTS

Baseline characteristics of study participants

Among 1,838 eligible, non-CKD participants of the KORA F4 study, 200 individuals developed CKD during a mean follow-up of 6.5 years (Fig. 1, Table 1). Incident CKD was diagnosed more frequently in hyperglycemic participants (22.1%) than in individuals with NGT (7.9%) (Table 1). Compared with non-CKD individuals, the incident cases of CKD in hyperglycemic and NGT groups were significantly older and had significantly higher baseline values of HbA1c, fasting and 2-h glucose and UACR, whereas their baseline eGFR values were significantly lower. They also self-reported a significantly higher intake of anti-hypertensive and lipid-lowering medication (Table 1).

Identification of metabolite biomarkers for incident CKD in hyperglycemia

Of 125 analyzed metabolites in 385 hyperglycemic participants, the baseline values of 13 metabolites were nominally associated ($P < 0.05$) with incident CKD, both in basic and full MLR models (Fig. 2A, Table S2). Among the 13 metabolites, nine corresponded to sphingomyelins (SMs) and SM C18:1 remained significant after stringent Bonferroni correction (Figs. 2A, S2). Of the 13 metabolites, four metabolites were selected by priority-Lasso and two (SM C18:1 and phosphatidylcholine diacyl (PC aa) C38:0) remained significant after stepwise *AIC* selection (Fig. 1). The relative concentrations of the two metabolites were significantly higher in 85 incident CKD

cases when compared to 300 non-CKD individuals (Fig. 2B). For example, a SD increase in the ln-transformed SM C18:1 concentration at baseline was associated with a 122% increased odds of CKD at follow-up (full model P = 3.315E-04; Table S2).

The results of the three-step feature selection thus identified two metabolites, SM C18:1 and PC aa C38:0, as candidate biomarkers of incident CKD in hyperglycemic individuals.

Sensitivity analyses consolidate the candidate CKD biomarkers

Propensity score matching in 385 hyperglycemic individuals resulted in 62 one-to-one matched incident CKD and non-CKD pairs. All covariates from the full model showed similar characteristics between the cases and matched controls (Table S3) and the two candidate biomarkers showed significant risk associations with incident CKD (Table S4).

Both metabolites were not significantly associated with incident CKD in 1453 normoglycemic individuals, i.e. when 115 incident CKD cases were compared with 1338 non-CKD individuals that were both NGT at baseline (Tables 1 and S5, Fig. 2B). This result indicates that the two candidate biomarkers of incident CKD are specific for hyperglycemia.

Their specificity for hyperglycemia was further confirmed by metabolite-glucose interaction analysis. The risk estimates of SM C18:1 and PC aa C38:0 association with incident CKD were significant only in the hyperglycemic subgroup as

well as in the top tertile of fasting and 2-h glucose, respectively (Table S5). Moreover, SM C18:1 demonstrated significant multiplicative interaction effects with glycemic status and 2-h glucose (Fig. 3, Table S5).

The fourth sensitivity analysis aimed to address the UACR- and eGFR-based outcomes separately. Among 385 hyperglycemic participants, 32 and 65 developed incident CKD according to UACR- and eGFR criteria, respectively. Both metabolites showed consistently significant risk effects for the UACR-based incident CKD in hyperglycemic participants, both in basic and full MLR (Table S6). Moreover, SM C18:1 was a significant predictor for eGFR-based incident CKD in the basic MLR (Table S6).

Superior discrimination ability and the best set of predictors of incident CKD in hyperglycemia

During 100 times of 10-fold cross-validation, the median AUC values of our developed sets of predictors (i.e. metabolites and clinical variables) were stable in all three machine learning algorithms with corresponding values above 0.813 (Fig. 4 and Table S7). When compared to the reference predictors (age, sex, eGFR, UACR), the median AUC value of our developed sets of predictors increased by 2.5% and reached 0.825 (95% *CI* = 0.801-0.849, SVM algorithm, Table S7), thereby outperforming the reference predictors in 97 out of 100 times of 10-fold cross-validation (Table S7). The improvement remained consistent after applying other two machine learning

approaches, RF (2.9% absolute increase in median AUC value) and AdaBoost (1.6%) (Table S7). These results suggest that our developed sets of predictors outperform the established clinical predictors for incident CKD.

We further identified the best set of predictors for incident CKD, which consisted of two metabolites (SM C18:1, PC aa C38:0) and five clinical variables (age, total cholesterol, fasting glucose, eGFR, UACR). This set was the most frequently selected set with 113 times over 1000 selection rounds (Table S8). Moreover, these seven variables were the most important ones, and metabolites SM C18:1 and PC aa C38:0 were selected 857 and 593 times over these 1000 rounds (Table S9). The mean AUC value of the best set of predictors for incident CKD was 0.857, which was 4.8% higher than the corresponding AUC value of the full model containing 14 clinical variables including two known CKD biomarkers eGFR and UACR (Table S10).

DISCUSSION

This longitudinal study revealed significant accumulation of sphingo- and glycerophospholipids (SM C18:1 and PC aa C38:0) in (pre-) T2D individuals up to 6.5 years before their clinical onset of CKD. These candidate metabolite biomarkers of incident CKD were specific for hyperglycemic state, i.e. individuals with increased fasting and/or 2-h glucose levels. Highly stable performances of the sets of predictors for incident CKD developed from 125 metabolites and 14 clinical variables were furthermore independently confirmed with three machine learning algorithms. The best

set of predictors consisted of the two metabolites (SM C18:1, PC aa C38:0) and five clinical variables (age, total cholesterol, fasting glucose, eGFR, UACR) and showed the best predictive power for early discrimination of hyperglycemic individuals at high risk of progressing to CKD.

Despite the relatively low coverage of our targeted metabolomics approach, i.e. lack of ceramides and other sphingolipids, our results support evidence on sphingomyelin accumulation in glomerular diseases of genetic and non-genetic origin (32). Out of 125 analyzed metabolites comprising amino acids, acylcarnitines, hexoses, glycerophospho- and sphingolipids (Table S1), SMs represented the majority of metabolites associated with incident CKD in hyperglycemic participants ($P < 0.05$, Fig. 2A). Increased SM levels in relation with CKD were also reported in individuals with Type 1 Diabetes (T1D) (33) and T2D (34), except for the non-targeted lipidomic study of T1D (35). Isomer annotation of the top significant metabolite SM C18:1 in our study revealed that it may consist of several sphingoid backbones (d16:1, d18:0, d18:1, d18:2, d19:1) bound to mainly saturated or monounsaturated fatty acyls with 16-18 carbons (36). A similar preference for saturated fatty acyl chains was found for PC aa C38:0 and PC aa C42:0, two diacyl PCs with positive association trends with incident CKD (Fig. 2A).

Circulatory levels of several other metabolites associated with CKD in our study (SM C16:0, SM C16:1, SM C24:1 and PC aa C38:0) have previously been shown to positively associate with coronary artery disease mortality (37). SM C16:0 and SM

C16:1 were also found to be positively associated with myocardial infarction (38). Moreover, higher plasma SMs were found in patients with coronary artery disease and causally related with progression of atherosclerosis lesions in animal models (39; 40). The PC aa C32:2 that showed an inverse association with incident CKD in our study was previously found to be protective for coronary artery disease mortality (37).These observations suggest that metabolic alterations associated with incident CKD may also reflect underlying cardiovascular disease, for which CKD is an independent risk factor (41).

Circulatory accumulation in SMs and saturated PCs in individuals with pre-diabetes and T2D may also reflect early stages of diabetic nephropathy such as mesangial matrix expansion, podocyte injury and glomerular enlargement (42). The sphingomyelin SM (d18:1/16:0) was reported to accumulate in the enlarged glomeruli of diabetic and obese mice and was detected in the glomeruli and vasculature of human kidney (43). SM (d18:1/16:0) is one of the possible isomers for SM C16:0 that was positively associated with incident CKD in our study (Fig. 2A) and highly correlated with our top hit SM C18:1 (Pearson's correlation coefficient = 0.66, $P < 2.2e\text{-}16$, Fig. S2). Renal accumulation in SM (d18:1/16:0) was related with reduced enzyme activity of AMP-activated protein kinase (AMPK) in the diabetic kidney glomeruli, mitochondrial dysfunction and CKD progression (43).

The altered levels of certain SM and PC species in hyperglycemic individuals under increased risk for CKD could be caused by fluctuations in their fatty acid profile,

which influences the first rate-limiting step in *de novo* SM synthesis, due to nutritional oversupply, dyslipidemia (44) or gut microbiome (45). The severity of CKD correlates with increased levels of saturated and mono-unsaturated fatty acids (46) and enzymes involved in *de novo* synthesis and the ceramide-sphingomyelin homeostasis such as sphingomyelin synthase 2 (SMS2) show fatty acyl-chain specificity and may determine the regional expression of SM species in the kidney (47). Reduced SM levels in the plasma membranes and lipoproteins improves whole-body insulin sensitivity (48) and SMS2 inhibition was suggested as a potential therapeutic target for controlling inflammatory responses and atherosclerosis (49; 50). Whether SMS2 inhibition could prevent the development of CKD in hyperglycemic individuals requires further investigation.

The current predictive models for CKD mainly rely on clinical variables (10; 11; 51; 52). Our study demonstrates that two candidate metabolite biomarkers, in combination with five clinical variables, yield the best performing set of predictors for incident CKD in hyperglycemic individuals. Furthermore, we show the power of appropriate combination of state-of-the-art machine learning and classical statistical approaches to reveal novel biomarkers and improve the performance of classical clinical predictors of CKD. The three-step feature selection, which we define in this study, was able to capture as few predictors as possible but achieve better predictive performance, which fulfills the ideal setting of clinical practice. Many epidemiological studies have used inappropriate ways to evaluate the performance of the identified

variables, in which, for example, certain variables were selected from the whole data set and then the predictive performance was only evaluated on those selected variables using resampling approaches on the same data set (53). Consequently, this could have potentially strongly overestimated the predictive performance, because the testing data set has been included as part of the whole data set to perform variable selection, and it cannot be regarded as testing data set anymore (53). In our study, we employed cross-validation in a combination with three-step feature selection and applied stringent internal validation procedures to evaluate the performance of the identified sets of predictors. In each round, the variable selection was only conducted in the training data and the performance evaluation was only performed in the testing data. In this way, we were able to attain accurate and unbiased internal AUC estimates. Given these advantages as described above, the consistent improvement of our developed sets of predictors on top of four established reference predictors in all three machine learning algorithms can be regarded as a significant progress.

Our study has several additional advantages. We used a well-characterized, population-based human cohort that allows to adjust for the influence of demographic parameters, medication and other clinical variables. Our stringent QC of metabolite profiles and adjustment for plate effects reduced the noise among all 3,061 measured samples. We performed sensitivity analyses to confirm the candidate metabolite biomarkers and investigate their interaction with glycemia.

A limitation of our study is a missing replication (of ten international human cohorts, none included at least 50 incident CKD cases in hyperglycemia and metabolites we measured). Discriminatory power of the candidate biomarkers and the best set of predictors cannot be generalized due to lack of external validation. Thus, we are aware that larger prospective studies are needed to validate our discoveries.

In summary, we identified two candidate metabolite biomarkers and the best set of predictors for incident CKD that are specific for individuals with pre-diabetes and T2D. This study demonstrates the value of metabolomics and appropriate combination of predictors in the improvement of accurate detection of hyperglycemic individuals with enhanced risk for CKD. With rising worldwide prevalence and burden of (pre)-diabetes-related CKD, combining metabolite and clinical predictors is a promising approach for effective predictions of future CKD in the framework of an integrated personalized diabetes management.

ACKNOWLEDGMENTS

We express our appreciation to all KORA study participants for donating their blood and time. We thank the field staff in Augsburg conducting the KORA studies. We are grateful to the staff (J. Scarpa, K. Faschinger, N. Lindemann) from the Institute of Epidemiology and the Genome Analysis Center Metabolomics Platform at the Helmholtz Zentrum München, who helped in the sample logistics, data and straw collection, and metabolomic measurements. Additionally, we thank the staff (e.g. A. Ludolph, S. Jelic and B. Langer) from the Institute of Genetic Epidemiology at the Helmholtz Zentrum München, and the platform KORA-PASST, for their help with KORA data logistics.

We thank Dr. Anne-Laure Boulesteix for tips on statistical methods.

AUTHOR CONTRIBUTIONS

J.H. conceived the study, analyzed the data and wrote the manuscript. C.H. researched cohort data and edited manuscript. M.C. contributed to pathway analysis and wrote the manuscript. M.T. researched data and edited manuscript. Jo.A. edited manuscript. S.Z. researched data. C.P. researched metabolomic data. L.W. edited manuscript. J.N. edited manuscript. M.F.S. researched data and edited manuscript. S.N. researched data. G.K. researched metabolomic data. K.S. researched metabolomic data. M.L. reviewed manuscript. F.S. edited manuscript. C.G. researched cohort data. Je.A. researched metabolomic data. M.R.dA. researched data. A.P. researched cohort data. R.W-S.

designed the study, researched metabolomic data and wrote the manuscript. R.W-S. is the guarantor of this work and, as such, had full access to all study data and takes responsibility for data integrity and accuracy of data analysis.

CONFLICT OF INTEREST STATEMENT

M.F.S was employed at Helmholtz Center Munich during his PhD thesis and is currently employed in the CardioRenal Medical Department of Bayer AG, however, the company was not involved in work related to data and manuscript generation.

FUNDING

The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. Part of this project was supported by EU FP7 grants HEALTH-2013-2.4.2-1/602936 (Project CarTarDis) and the 19076 & 20679 iPDM-GO "Integrated Personalized Diabetes Management Goes Europe" innovation project supported by the European Institute of Innovation and Technology (EIT) Health. EIT Health is supported by the EIT, a body of the European Union. K.S. is supported by Biomedical Research Program

funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation.

PRIOR PRESENTATION INFORMATION

Parts of this study were presented in poster form at the 15th Annual Conference of the metabolomics society 2019 in The Hague, the Netherlands, and at the 7th DZD Diabetes Research School 2019 in Barcelona, Spain.

REFERENCES

1. Bikbov B, Purcell CA, Levey AS, et al.: Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2020;395:709-733

2. Webster AC, Nagler EV, Morton RL, Masson P: Chronic Kidney Disease. Lancet 2017;389:1238-1252

3. Alicic RZ, Neumiller JJ, Johnson EJ, Dieter B, Tuttle KR: Sodium-Glucose Cotransporter 2 Inhibition and Diabetic Kidney Disease. Diabetes 2019;68:248-257

4. Plantinga LC, Crews DC, Coresh J, Miller ER, 3rd, Saran R, Yee J, Hedgeman E, Pavkov M, Eberhardt MS, Williams DE, Powe NR: Prevalence of chronic kidney disease in US adults with undiagnosed diabetes or prediabetes. Clin J Am Soc Nephrol 2010;5:673-682

5. Melsom T, Schei J, Stefansson VT, Solbu MD, Jenssen TG, Mathisen UD, Wilsgaard T, Eriksen BO: Prediabetes and Risk of Glomerular Hyperfiltration and Albuminuria in the General Nondiabetic Population: A Prospective Cohort Study. Am J Kidney Dis 2016;67:841-850

6. Markus MRP, Ittermann T, Baumeister SE, Huth C, Thorand B, Herder C, Roden M, Siewert-Markus U, Rathmann W, Koenig W, Dorr M, Volzke H, Schipf S, Meisinger C: Prediabetes is associated with microalbuminuria, reduced kidney function and chronic kidney disease in the general population: The KORA (Cooperative Health Research in the Augsburg Region) F4-Study. Nutr Metab Cardiovasc Dis 2018;28:234-242

7. Li W, Wang A, Jiang J, Liu G, Wang M, Li D, Wen J, Mu Y, Du X, Gaisano H, Dou J, He Y, Kim GS, Oh HH, Kim SH, Kim BO, Byun YS: Risk of chronic kidney disease defined by

decreased estimated glomerular filtration rate in individuals with different prediabetic phenotypes: results from a prospective cohort study in China. BMJ Open Diabetes Res Care 2020;8:130

8. Ceriello A, Barkai L, Christiansen JS, Czupryniak L, Gomis R, Harno K, Kulzer B, Ludvigsson J, Nemethyova Z, Owens D, Schnell O, Tankova T, Taskinen MR, Verges B, Weitgasser R, Wens J: Diabetes as a case study of chronic disease management with a personalized approach: the role of a structured feedback loop. Diabetes Res Clin Pract 2012;98:5-10

9. Levin A, Stevens PE, Bilous RW, Coresh J, De Francisco ALM, De Jong PE, Griffith KE, Hemmelgarn BR, Iseki K, Lamb EJ, Levey AS, Riella MC, Shlipak MG, Wang H, White CT, Winearls CG: Kidney disease: Improving global outcomes (KDIGO) CKD work group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney International Supplements 2013;3:1--150

10. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, Levin A, Levey AS: A predictive model for progression of chronic kidney disease to kidney failure. Jama 2011;305:1553-1559

11. Dunkler D, Gao P, Lee SF, Heinze G, Clase CM, Tobe S, Teo KK, Gerstein H, Mann JF, Oberbauer R: Risk Prediction for Early CKD in Type 2 Diabetes. Clin J Am Soc Nephrol 2015;10:1371-1379

12. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, Joost HG, Fritsche A, Haring HU, Hrabe de Angelis M, Peters A, Roden M, Prehn C, Wang-Sattler R, Illig T, Schulze MB, Adamski J, Boeing H, Pischon T: Identification of serum metabolites associated with risk of type 2

diabetes using a targeted metabolomic approach. Diabetes 2013;62:639-648

13. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, Heim K, Campillos M, Holzapfel C, Thorand B, Grallert H, Xu T, Bader E, Huth C, Mittelstrass K, Doring A, Meisinger C, Gieger C, Prehn C, Roemisch-Margl W, Carstensen M, Xie L, Yamanaka-Okumura H, Xing G, Ceglarek U, Thiery J, Giani G, Lickert H, Lin X, Li Y, Boeing H, Joost HG, de Angelis MH, Rathmann W, Suhre K, Prokisch H, Peters A, Meitinger T, Roden M, Wichmann HE, Pischon T, Adamski J, Illig T: Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol 2012;8:615

14. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, Lewis GD, Fox CS, Jacques PF, Fernandez C, O'Donnell CJ, Carr SA, Mootha VK, Florez JC, Souza A, Melander O, Clish CB, Gerszten RE: Metabolite profiles and the risk of developing diabetes. Nat Med 2011;17:448-453

15. Chen GC, Chai JC, Yu B, Michelotti GA, Grove ML, Fretts AM, Daviglus ML, Garcia-Bedoya OL, Thyagarajan B, Schneiderman N, Cai J, Kaplan RC, Boerwinkle E, Qi Q: Serum sphingolipids and incident diabetes in a US population with high diabetes burden: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Am J Clin Nutr 2020;112:57-65

16. Carayol M, Leitzmann MF, Ferrari P, Zamora-Ros R, Achaintre D, Stepien M, Schmidt JA, Travis RC, Overvad K, Tjonneland A, Hansen L, Kaaks R, Kuhn T, Boeing H, Bachlechner U, Trichopoulou A, Bamia C, Palli D, Agnoli C, Tumino R, Vineis P, Panico S, Quiros JR, Sanchez-Cantalejo E, Huerta JM, Ardanaz E, Arriola L, Agudo A, Nilsson J, Melander O, Bueno-de-Mesquita B, Peeters PH, Wareham N, Khaw KT, Jenab M, Key TJ, Scalbert A, Rinaldi S: Blood

Metabolic Signatures of Body Mass Index: A Targeted Metabolomics Study in the EPIC Cohort. J Proteome Res 2017;16:3137-3146

17. Leal-Witt MJ, Ramon-Krauel M, Samino S, Llobet M, Cuadras D, Jimenez-Chillaron JC, Yanes O, Lerin C: Untargeted metabolomics identifies a plasma sphingolipid-related signature associated with lifestyle intervention in prepubertal children with obesity. Int J Obes (Lond) 2018;42:72-78

18. Razquin C, Toledo E, Clish CB, Ruiz-Canela M, Dennis C, Corella D, Papandreou C, Ros E, Estruch R, Guasch-Ferre M, Gomez-Gracia E, Fito M, Yu E, Lapetra J, Wang D, Romaguera D, Liang L, Alonso-Gomez A, Deik A, Bullo M, Serra-Majem L, Salas-Salvado J, Hu FB, Martinez-Gonzalez MA: Plasma Lipidomic Profiling and Risk of Type 2 Diabetes in the PREDIMED Trial. Diabetes Care 2018;41:2617-2624

19. Alderete TL, Jin R, Walker DI, Valvi D, Chen Z, Jones DP, Peng C, Gilliland FD, Berhane K, Conti DV, Goran MI, Chatzi L: Perfluoroalkyl substances, metabolomic profiling, and alterations in glucose homeostasis among overweight and obese Hispanic children: A proof-of-concept analysis. Environ Int 2019;126:445-453

20. Hocher B, Adamski J: Metabolomics for clinical use and research in chronic kidney disease. Nat Rev Nephrol 2017;13:269-284

21. Goek ON, Prehn C, Sekula P, Romisch-Margl W, Doring A, Gieger C, Heier M, Koenig W, Wang-Sattler R, Illig T, Suhre K, Adamski J, Kottgen A, Meisinger C: Metabolites associate with kidney function decline and incident chronic kidney disease in the general population. Nephrol Dial Transplant 2013;28:2131-2138

22. Solini A, Manca ML, Penno G, Pugliese G, Cobb JE, Ferrannini E: Prediction of Declining Renal Function and Albuminuria in Patients With Type 2 Diabetes by Metabolomics. J Clin Endocrinol Metab 2016;101:696-704

23. Herder C, Kannenberg JM, Huth C, Carstensen-Kirberg M, Rathmann W, Koenig W, Heier M, Puttgen S, Thorand B, Peters A, Roden M, Meisinger C, Ziegler D: Proinflammatory Cytokines Predict the Incidence and Progression of Distal Sensorimotor Polyneuropathy: KORA F4/FF4 Study. Diabetes Care 2017;40:569-576

24. Chak CM, Lacruz ME, Adam J, Brandmaier S, Covic M, Huang J, Meisinger C, Tiller D, Prehn C, Adamski J, Berger U, Gieger C, Peters A, Kluttig A, Wang-Sattler R: Ageing Investigation Using Two-Time-Point Metabolomics Data from KORA and CARLA Studies. Metabolites 2019;9

25. World Health O, International Diabetes F: Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia : report of a WHO/IDF consultation. Geneva, World Health Organization, 2006

26. Inker LA, Schmid CH, Tighiouart H, Eckfeldt JH, Feldman HI, Greene T, Kusek JW, Manzi J, Van Lente F, Zhang YL, Coresh J, Levey AS: Estimating glomerular filtration rate from serum creatinine and cystatin C. N Engl J Med 2012;367:20-29

27. Römisch-Margl W, Prehn C, Bogumil R, Röhring C, Suhre K, Adamski J: Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. Metabolomics 2012;8:133-142

28. Klau S, Jurinovic V, Hornung R, Herold T, Boulesteix AL: Priority-Lasso: a simple

hierarchical approach to the prediction of clinical outcome using multi-omics data. BMC Bioinformatics 2018;19:322

29. Chang C-C, Lin C-J: LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1-27

30. Liaw A, Wiener M: Classification and Regression by randomForest. R News 2002;2:18--22

31. Culp M, Johnson K, Michailides G: ada: An R Package for Stochastic Boosting. Journal of Statistical Software 2006;017

32. Merscher S, Fornoni A: Podocyte pathology and nephropathy - sphingolipids in glomerular diseases. Front Endocrinol (Lausanne) 2014;5:127

33. Makinen VP, Tynkkynen T, Soininen P, Forsblom C, Peltola T, Kangas AJ, Groop PH, Ala-Korpela M: Sphingomyelin is associated with kidney disease in type 1 diabetes (The FinnDiane Study). Metabolomics 2012;8:369-375

34. Liu JJ, Ghosh S, Kovalik JP, Ching J, Choi HW, Tavintharan S, Ong CN, Sum CF, Summers SA, Tai ES, Lim SC: Profiling of Plasma Metabolites Suggests Altered Mitochondrial Fuel Usage and Remodeling of Sphingolipid Metabolism in Individuals With Type 2 Diabetes and Kidney Disease. Kidney Int Rep 2017;2:470-480

35. Tofte N, Suvitaival T, Ahonen L, Winther SA, Theilade S, Frimodt-Moller M, Ahluwalia TS, Rossing P: Lipidomic analysis reveals sphingomyelin and phosphatidylcholine species associated with renal impairment and all-cause mortality in type 1 diabetes. Sci Rep 2019;9:16398

36. Annotation of potential isobaric and isomeric lipid species measured with the AbsoluteIDQ

p180 Kit (and p150 Kit) [article online], Available from https://www.biocrates.com/images/p180_List_of_Isobars_and_Isomers_v1_2019.pdf. 2019

37. Sigruener A, Kleber ME, Heimerl S, Liebisch G, Schmitz G, Maerz W: Glycerophospholipid and sphingolipid species and mortality: the Ludwigshafen Risk and Cardiovascular Health (LURIC) study. PLoS One 2014;9:e85724

38. Floegel A, Kuhn T, Sookthai D, Johnson T, Prehn C, Rolle-Kampczyk U, Otto W, Weikert C, Illig T, von Bergen M, Adamski J, Boeing H, Kaaks R, Pischon T: Serum metabolites and risk of myocardial infarction and ischemic stroke: a targeted metabolomic approach in two German prospective cohorts. Eur J Epidemiol 2018;33:55-66

39. Jiang XC, Paultre F, Pearson TA, Reed RG, Francis CK, Lin M, Berglund L, Tall AR: Plasma sphingomyelin level as a risk factor for coronary artery disease. Arterioscler Thromb Vasc Biol 2000;20:2614-2618

40. Li Z, Basterr MJ, Hailemariam TK, Hojjati MR, Lu S, Liu J, Liu R, Zhou H, Jiang XC: The effect of dietary sphingolipids on plasma sphingomyelin metabolism and atherosclerosis. Biochim Biophys Acta 2005;1735:130-134

41. Cai Q, Mukku VK, Ahmad M: Coronary artery disease in patients with chronic kidney disease: a clinical update. Curr Cardiol Rev 2013;9:331-339

42. Alicic RZ, Rooney MT, Tuttle KR: Diabetic Kidney Disease: Challenges, Progress, and Possibilities. Clin J Am Soc Nephrol 2017;12:2032-2045

43. Miyamoto S, Hsu C-C, Hamm G, Darshi M, Diamond-Stanic M, Declèves A-E, Slater L, Pennathur S, Stauber J, Dorrestein PC, Sharma K: Mass Spectrometry Imaging Reveals

Elevated Glomerular ATP/AMP in Diabetes/obesity and Identifies Sphingomyelin as a Possible Mediator. EBioMedicine 2016;7:121-134

44. Torretta E, Barbacini P, Al-Daghri NM, Gelfi C: Sphingolipids in Obesity and Correlated Co-Morbidities: The Contribution of Gender, Age and Environment. Int J Mol Sci 2019;20

45. Johnson EL, Heaver SL, Waters JL, Kim BI, Bretin A, Goodman AL, Gewirtz AT, Worgall TS, Ley RE: Sphingolipids produced by gut bacteria enter host metabolic pathways impacting ceramide levels. Nat Commun 2020;11:2471

46. Czumaj A, Śledziński T, Carrero JJ, Stepnowski P, Sikorska-Wisniewska M, Chmielewski M, Mika A: Alterations of Fatty Acid Profile May Contribute to Dyslipidemia in Chronic Kidney Disease by Influencing Hepatocyte Metabolism. Int J Mol Sci 2019;20

47. Sugimoto M, Wakabayashi M, Shimizu Y, Yoshioka T, Higashino K, Numata Y, Okuda T, Zhao S, Sakai S, Igarashi Y, Kuge Y: Imaging Mass Spectrometry Reveals Acyl-Chain- and Region-Specific Sphingolipid Metabolism in the Kidneys of Sphingomyelin Synthase 2-Deficient Mice. PLoS One 2016;11:e0152191

48. Li Z, Zhang H, Liu J, Liang CP, Li Y, Li Y, Teitelman G, Beyer T, Bui HH, Peake DA, Zhang Y, Sanders PE, Kuo MS, Park TS, Cao G, Jiang XC: Reducing plasma membrane sphingomyelin increases insulin sensitivity. Mol Cell Biol 2011;31:4205-4218

49. Fan Y, Shi F, Liu J, Dong J, Bui HH, Peake DA, Kuo MS, Cao G, Jiang XC: Selective reduction in the sphingomyelin content of atherogenic lipoproteins inhibits their retention in murine aortas and the subsequent development of atherosclerosis. Arterioscler Thromb Vasc Biol 2010;30:2114-2120

50. Adachi R, Ogawa K, Matsumoto SI, Satou T, Tanaka Y, Sakamoto J, Nakahata T, Okamoto R, Kamaura M, Kawamoto T: Discovery and characterization of selective human sphingomyelin synthase 2 inhibitors. Eur J Med Chem 2017;136:283-293

51. Ravizza S, Huschto T, Adamov A, Bohm L, Busser A, Flother FF, Hinzmann R, Konig H, McAhren SM, Robertson DH, Schleyer T, Schneidinger B, Petrich W: Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med 2019;25:57-59

52. Echouffo-Tcheugui JB, Kengne AP: Risk models to predict chronic kidney disease and its progression: a systematic review. PLoS Med 2012;9:e1001344

53. Boulesteix AL, Wright MN, Hoffmann S, Konig IR: Statistical learning approaches in the genetic epidemiology of complex diseases. Hum Genet 2020;139:73-84

TABLES

Table 1. Characteristics of the KORA study population

KORA participants were classified according to their hyperglycemic status at baseline (F4) and incident chronic kidney disease status at follow-up (FF4). Unless indicated, variables show baseline measurements. Mean ± standard deviation is provided for quantitative variables if not indicated otherwise. *P*-values were calculated by univariate logistic regression. *P*-values shown in bold represent statistical significance at 0.05 level.

Abbreviations: CKD, chronic kidney disease; HbA_{1c}, glycated hemoglobin; 2-h glucose, two hour post load glucose; NGT, normal glucose tolerance; BP, blood pressure; eGFR, estimated glomerular filtration rate; UACR, urinary albumin-to-creatinine ratio.

Clinical variables	Hyperglycemic participants			NGT participants		
	Incident CKD N = 85	Non-CKD N = 300	*P*-value	Incident CKD N = 115	Non-CKD N = 1338	*P*-value
Age, years	67.78 ± 8.78	59.44 ± 9.39	**1.29E-10**	60.97 ± 12	50.05 ± 10.82	**4.81E-20**
Sex, male, %	55.29	58.00	0.656	46.09	46.64	0.910
BMI, kg/m²	30.11 ± 4.58	29.74 ± 4.80	0.522	27.39 ± 4.51	26.29 ± 4.09	**0.007**
HbA_{1c} (%)	6.06 ± 0.86	5.82 ± 0.57	**0.004**	5.49 ± 0.29	5.33 ± 0.30	**3.71E-08**
HbA_{1c} (mmol/mol)	42.81 ± 9.32	40.14 ± 6.24	**0.004**	36.56 ± 3.24	34.76 ± 3.39	**1.03E-07**
Fasting glucose, mg/dl	116.02 ± 28.6	110.23 ± 18.82	**0.031**	93.61 ± 7.42	91.4 ± 7.56	**0.003**
2-h glucose, mg/dl	173.59 ± 43.17[b]	159.82 ± 39.87[b]	**0.019**	102.7 ± 20.68	96.37 ± 20.53	**0.002**
Systolic BP, mmHg	132.01 ± 18.72	128.78 ± 17.16	0.135	124.73 ± 18.42	117.69 ± 15.87	**9.59E-06**
Diastolic BP, mmHg	75.14 ± 9.53	78.25 ± 9.47	**0.009**	76.36 ± 10.51	74.81 ± 9.3	0.089
Triglyceride, mg/dl[a]	130.0 [93 - 186]	133.5 [94.8 - 195.3]	0.859	107 [75 - 143]	91 [63 - 130]	0.220
Total cholesterol, mg/dl	212.87 ± 38.32	225.2 ± 39.7	**0.012**	219.39 ± 40.24	213.45 ± 37.75	0.108
HDL cholesterol, mg/dl	51.87 ± 11.64	51.66 ± 13.66	0.897	57.06 ± 15.27	58.00 ± 14.70	0.514
LDL cholesterol, mg/dl	130.64 ± 35.47	144.77 ± 34.47	**0.001**	138.45 ± 35.56	134.03 ± 33.84	0.180
Baseline eGFR, mL/min/1.73 m²	78.42 ± 13.6	90.48 ± 12.48	**2.18E-11**	83.13 ± 15.85	98.38 ± 12.79	**1.39E-25**
Follow-up eGFR, mL/min/1.73 m²	57.5 ± 18.3	81.67 ± 13.12		66.68 ± 19.32	89.5 ± 13.48	
Baseline UACR, mg/g[a]	10.22 [4.8 - 15.0]	5.45 [3.8 - 9.1]	**2.54E-07**	7.16 [4.7 - 13.8]	4.64 [3.2 - 7.2]	**3.81E-13**
Follow-up UACR, mg/g[a]	14.47 [6.02 - 41.02]	5.54 [3.34 - 9.47]		18.51 [5.4 - 54.1]	4.22 [2.9 - 6.6]	
Smoking, %			0.321			0.699
Non-smoker	47.06	41.33	Ref.	41.74	42.15	Ref.
Former smoker	47.06	48.00	0.558	41.74	38.57	0.676
Current smoker	5.88	10.67	0.159	16.52	19.28	0.607
Medication usage, %						
Lipid-lowering	30.59	11.33	**3.20E-05**	15.65	6.28	**2.78E-04**
Antihypertensive	71.76	42.67	**4.49E-06**	50.43	16.07	**8.88E-17**
Anti-diabetic	16.47	11.33	0.208	0	0	-

[a] values are presented as median [25th–75th percentile];

[b] In the hyperglycemic participants, 2-h glucose levels were only available in 61 individuals with incident CKD and 254 individuals without CKD.

FIGURE LEGENDS

Figure 1. Study design

Abbreviations: CKD, chronic kidney disease; BP, blood pressure; eGFR, estimated glomerular filtration rate; UACR, urinary albumin-to-creatinine ratio; *AIC*, Akaike information criterion; SVM, support vector machine; RF, random forest; AdaBoost, adaptive boosting.

Figure 2. Serum metabolite associations with incident chronic kidney disease

A, Volcano plot of the association results for 125 metabolites with incident CKD in hyperglycemic individuals. Odds ratios and *P*-values are from logistic regression analysis adjusted for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, fasting glucose, use of lipid lowering drugs, antihypertensive and anti-diabetic medication, and baseline values of estimated glomerular filtration rate and urinary albumin-to-creatinine ratio. The upper and the lower interrupted lines represent Bonferroni-corrected and uncorrected ($P = 0.05$) significance levels, respectively. B, Mean residuals (with standard errors) of SM C18:1 and PC aa C38:0 for non-CKD and incident CKD in hyperglycemic and NGT individuals, respectively. Metabolite residuals were calculated with linear regression models adjusted for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, and fasting glucose. **Abbreviations:** CKD, chronic kidney disease; SM, sphingomyelin; PC aa, phosphatidylcholine diacyl; NGT, normal glucose tolerance.

Figure 3. Stratified associations of candidate biomarkers with incident chronic kidney disease according to glucose status

Associations of SM C18:1 and PC aa C38:0 with incident chronic kidney disease stratified by hyperglycemic status (A), and each tertile of fasting glucose (B) and 2-h glucose (C) values. Regression coefficients in NGT, first and second tertile of fasting and 2-h glucose were adjusted for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, fasting glucose, use of lipid lowering drug and antihypertensive medication, and baseline values of estimated glomerular filtration rate and urinary albumin-to-creatinine ratio. Regression coefficients in hyperglycemic group and the top tertile of fasting and 2-h glucose were additionally adjusted for anti-diabetic medication. **Abbreviations:** NGT, normal glucose tolerance; 2-h glucose, two hour post load glucose; SM, sphingomyelin; PC aa, phosphatidylcholine diacyl.

Figure 4. Prediction performance of incident chronic kidney disease in hyperglycemic individuals in three machine learning approaches

The boxplots show the AUC values of two models applying three machine learning approaches over 100 times of 10-fold cross-validation. Reference predictors: baseline age, sex, estimated glomerular filtration rate and urinary albumin-to-creatinine ratio.

Developed sets of predictors: combination of metabolites and clinical variables which were identified by the three-step feature selection in each round. For the resampling rounds, in each iteration of each 10-fold cross-validation, the three-step feature selection procedure was conducted and metabolites and clinical variables were selected in the training data. The set of selected metabolites and clinical variables and the reference predictors were used to develop respective prediction models with the three approaches in the training data. The AUC values were computed for the test data only. The ten AUC values of each model of each approach were averaged to produce a single estimate that was displayed in boxplots. The procedure of 10-fold cross-validation was randomly repeated 100 times, which generated 100 cross-validation AUC values of each prediction model for each approach. **Abbreviations**: AUC, area under the receiver operating characteristic curve.

Fig. 1

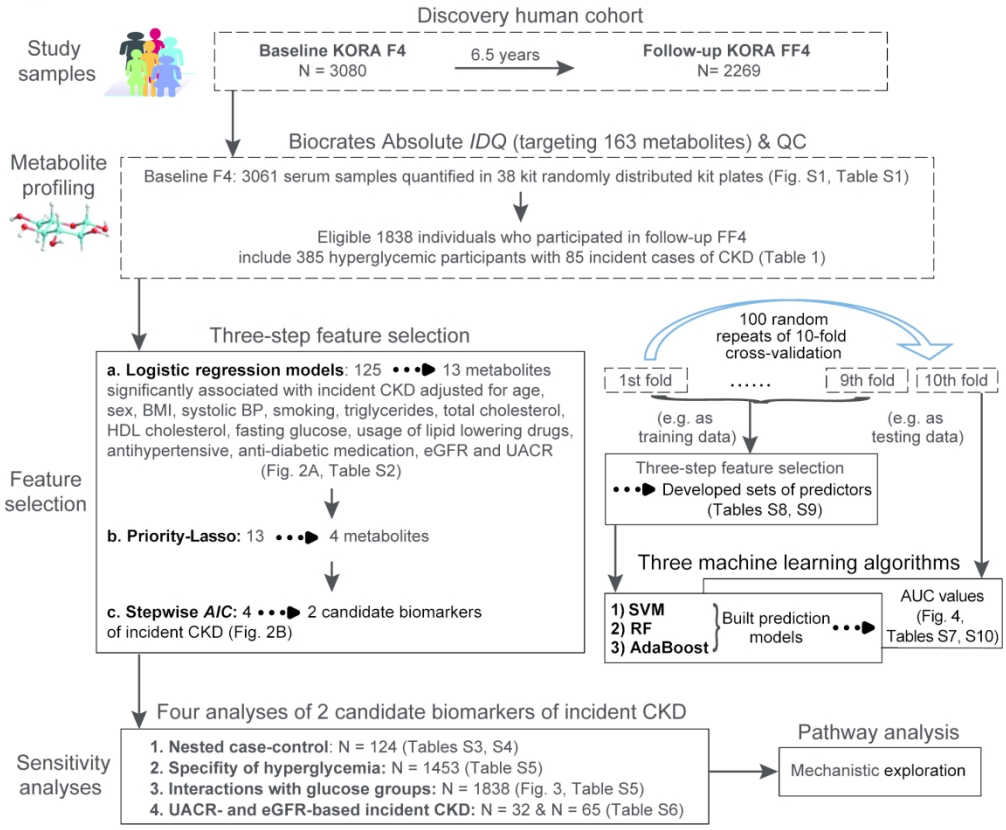

Figure 1. Study design Abbreviations: CKD, chronic kidney disease; BP, blood pressure; eGFR, estimated glomerular filtration rate; UACR, urinary albumin-to-creatinine ratio; AIC, Akaike information criterion; SVM, support vector machine; RF, random forest; AdaBoost, adaptive boosting.

Fig. 2

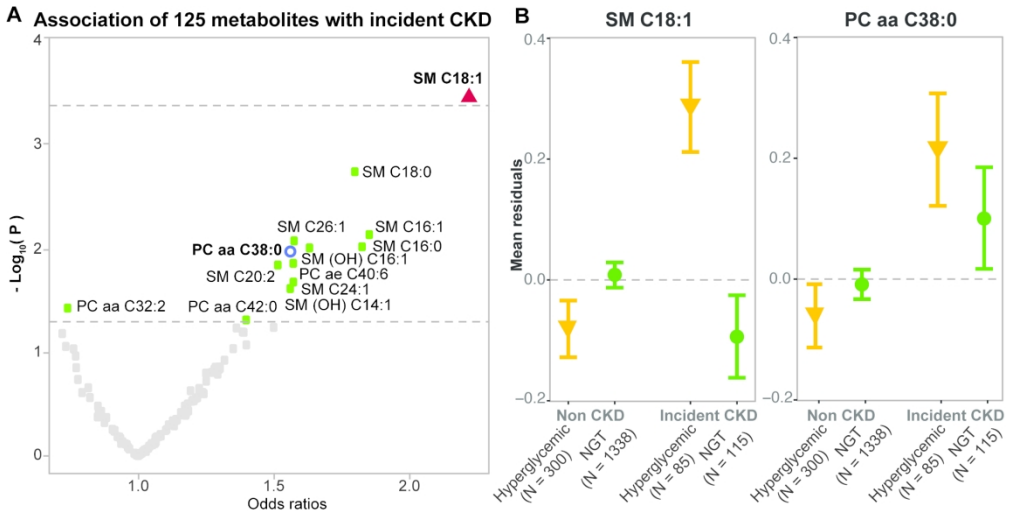

Figure 2. Serum metabolite associations with incident chronic kidney disease
A, Volcano plot of the association results for 125 metabolites with incident CKD in hyperglycemic individuals. Odds ratios and P-values are from logistic regression analysis adjusted for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, fasting glucose, use of lipid-lowering drugs, antihypertensive and anti-diabetic medication, and baseline values of estimated glomerular filtration rate and urinary albumin-to-creatinine ratio. The upper and the lower interrupted lines represent Bonferroni-corrected and uncorrected (P = 0.05) significance levels, respectively. B, Mean residuals (with standard errors) of SM C18:1 and PC aa C38:0 for non-CKD and incident CKD in hyperglycemic and NGT individuals, respectively. Metabolite residuals were calculated with linear regression models adjusted for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, and fasting glucose. Abbreviations: CKD, chronic kidney disease; SM, sphingomyelin; PC aa, phosphatidylcholine diacyl; NGT, normal glucose tolerance.

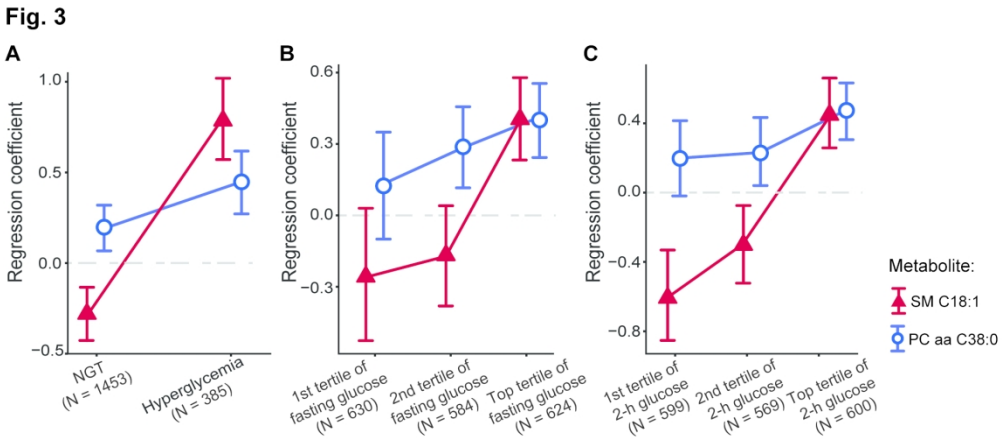

Figure 3. Associations of candidate biomarkers with incident CKD stratified according to glucose status

Fig. 4

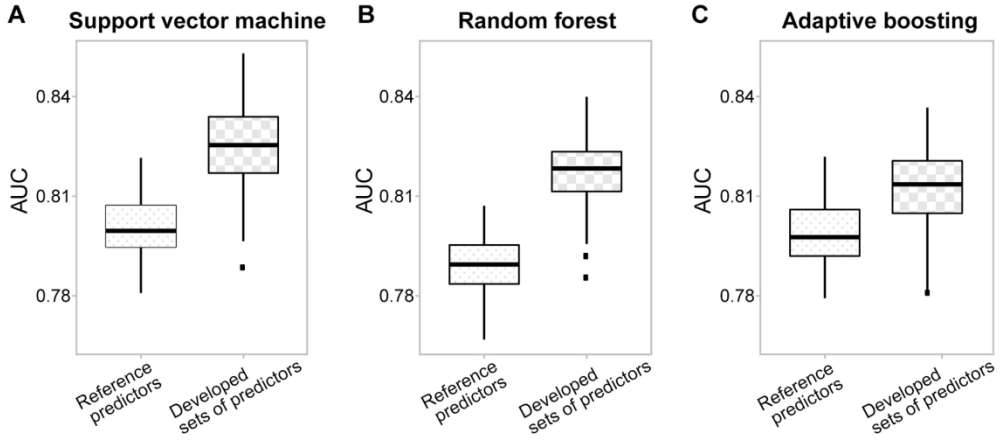

Figure 4. Prediction performance of incident CKD in hyperglycemic individuals in three machine learning approaches

156x77mm (300 x 300 DPI)

Machine learning approaches revealed metabolic signatures of incident chronic kidney disease in persons with pre- and type 2 diabetes

Supplementary Tables

Table S1. Metabolite panel of baseline KORA F4 study

The abbreviations and biochemical names of 163 metabolites are shown in the first and second column, respectively. The third column shows the missing rate of each metabolite among 3,061 KORA F4 individuals. The non-detectable rate was defined as the number of the non-detectable values divided by the number of all measured values. The fourth column presents the arithmetic means of the coefficients of variance (CV) of 114 quality controls samples (i.e. three on each kit plate). The percentage of individuals above the limit of detection (LOD) among 3,061 KORA F4 participants is shown in the fifth column. The sixth column presents the mean value of metabolite concentration (μM) in 3,061 KORA F4 participants after adjusting for plate effects. The last column shows the status (used/excluded) for each metabolite.

Metabolite	Biochemical name	Non-Detectable Rate (%)	CV (%)	Above LOD (%)	Mean Concentration (μM)	Application
C0	Carnitine	0.0	7.50	99.97	35.89	Used
C10	Decanoylcarnitine	0.0	12.40	98.30	0.36	Used
C10:1	Decenoylcarnitine	0.0	10.45	36.20	0.17	Excluded
C10:2	Decadienylcarnitine	0.0	15.61	58.58	0.04	Used
C12	Dodecanoylcarnitine	0.0	10.63	89.51	0.13	Used
C12:1	Dodecenoylcarnitine	0.0	13.51	2.16	0.15	Excluded
C12-DC	Dodecanedioylcarnitine	0.0	15.71	0.00	0.06	Excluded
C14	Tetradecanoylcarnitine	0.0	11.80	47.60	0.05	Excluded
C14:1	Tetradecenoylcarnitine	0.0	20.10	99.97	0.15	Used
C14:1-OH	Hydroxytetradecenoylcarnitine	0.0	17.88	76.54	0.02	Used
C14:2	Tetradecadienylcarnitine	0.0	11.19	99.44	0.03	Used
C14:2-OH	Hydroxytetradecadienylcarnitine	0.0	24.24	44.10	0.01	Excluded
C16	Hexadecanoylcarnitine	0.0	10.02	99.97	0.12	Used
C16:1	Hexadecenoylcarnitine	0.0	10.39	2.48	0.04	Excluded
C16:1-OH	Hydroxyhexadecenoylcarnitine	0.0	17.20	1.31	0.01	Excluded
C16:2	Hexadecadienylcarnitine	0.0	19.46	77.56	0.01	Used
C16:2-OH	Hydroxyhexadecadienylcarnitine	0.0	20.19	1.08	0.01	Excluded
C16-OH	Hydroxyhexadecanoylcarnitine	0.0	21.99	3.23	0.01	Excluded
C18	Octadecanoylcarnitine	0.0	12.52	99.90	0.05	Used
C18:1	Octadecenoylcarnitine	0.0	13.30	99.93	0.13	Used
C18:1-OH	Hydroxyoctadecenoylcarnitine	0.0	25.50	1.14	0.01	Excluded
C18:2	Octadecadienylcarnitine	0.0	11.00	99.97	0.05	Used
C2	Acetylcarnitine	0.0	9.62	99.97	8.26	Used
C3	Propionylcarnitine	0.0	10.28	99.97	0.40	Used
C3:1	Propenonylcarnitine	0.0	37.84	0.49	0.01	Excluded
C3-OH	Hydroxypropionylcarnitine	0.0	98.90	7.64	0.03	Excluded
C4	Butyrylcarnitine	0.0	11.20	99.97	0.23	Used
C4:1	Butenylcarnitine	0.0	35.99	10.42	0.02	Excluded
C4-OH (C3-DC)	Hydroxybutyrylcarnitine	0.0	34.81	9.64	0.09	Excluded
C5	Valerylcarnitine	0.0	15.83	99.97	0.12	Used
C5:1	Tiglylcarnitine	0.0	26.40	1.83	0.03	Excluded
C5:1-DC	Glutaconylcarnitine	0.0	51.54	13.92	0.02	Excluded
C5-DC (C6-OH)	Glutarylcarnitine	0.0	36.29	58.05	0.03	Excluded
C5-M-DC	Methylglutarylcarnitine	0.0	48.62	3.82	0.03	Excluded
C5-OH	Hydroxyvalerylcarnitine	0.0	24.31	14.05	0.04	Excluded
C6 (C4:1-DC)	Hexanoylcarnitine (Fumarylcarnitine)	0.0	14.19	87.62	0.07	Used
C6:1	Hexenoylcarnitine	0.0	36.13	3.50	0.02	Excluded
C7-DC	Pimelylcarnitine	0.0	29.31	73.21	0.05	Excluded
C8	Octanoylcarnitine	0.0	9.73	50.38	0.23	Used

C8:1	Octenoylcarnitine	0.0	8.45	99.22	0.09	Used
C9	Nonaylcarnitine	0.0	33.00	92.98	0.05	Excluded
Arg	Arginine	0.0	7.58	99.97	115.89	Used
Gln	Glutamine	0.0	14.28	99.97	619.01	Used
Gly	Glycine	0.0	8.35	99.97	307.70	Used
His	Histidine	0.0	10.50	99.97	98.28	Used
Met	Methionine	0.0	14.82	99.97	32.03	Used
Orn	Ornithine	0.0	11.33	99.97	81.47	Used
Phe	Phenylalanine	0.0	8.87	99.97	62.25	Used
Pro	Proline	0.0	10.15	100.00	176.09	Used
Ser	Serine	0.0	9.34	99.97	128.46	Used
Thr	Threonine	0.0	11.20	99.97	106.03	Used
Trp	Tryptophan	0.0	7.45	99.97	82.62	Used
Tyr	Tyrosine	0.0	8.61	99.97	85.47	Used
Val	Valine	0.0	15.51	100.00	277.00	Used
xLeu	Leucine/Isoleucine	0.0	9.48	100.00	213.92	Used
PC aa C24:0	Phosphatidylcholine diacyl C24:0	0.0	24.13	78.93	0.15	Used
PC aa C26:0	Phosphatidylcholine diacyl C26:0	0.0	38.23	11.43	1.08	Excluded
PC aa C28:1	Phosphatidylcholine diacyl C28:1	0.0	9.78	99.97	3.38	Used
PC aa C30:0	Phosphatidylcholine diacyl C30:0	0.0	12.24	99.97	4.74	Used
PC aa C30:2	Phosphatidylcholine diacyl C30:2	95.5	75.42	99.87	0.06	Excluded
PC aa C32:0	Phosphatidylcholine diacyl C32:0	0.0	12.23	99.97	15.21	Used
PC aa C32:1	Phosphatidylcholine diacyl C32:1	0.0	12.32	99.97	21.98	Used
PC aa C32:2	Phosphatidylcholine diacyl C32:2	0.1	20.80	99.97	3.95	Used
PC aa C32:3	Phosphatidylcholine diacyl C32:3	0.0	9.92	99.97	0.48	Used
PC aa C34:1	Phosphatidylcholine diacyl C34:1	0.0	11.63	99.97	240.68	Used
PC aa C34:2	Phosphatidylcholine diacyl C34:2	0.0	16.87	99.97	392.77	Used
PC aa C34:3	Phosphatidylcholine diacyl C34:3	0.0	14.83	99.97	18.07	Used
PC aa C34:4	Phosphatidylcholine diacyl C34:4	0.0	10.15	99.97	2.27	Used
PC aa C36:0	Phosphatidylcholine diacyl C36:0	0.0	19.81	99.97	2.72	Used
PC aa C36:1	Phosphatidylcholine diacyl C36:1	0.0	9.14	99.97	53.89	Used
PC aa C36:2	Phosphatidylcholine diacyl C36:2	0.0	8.32	99.97	232.62	Used
PC aa C36:3	Phosphatidylcholine diacyl C36:3	0.0	10.63	99.97	150.39	Used
PC aa C36:4	Phosphatidylcholine diacyl C36:4	0.0	11.24	100.00	220.61	Used
PC aa C36:5	Phosphatidylcholine diacyl C36:5	0.0	13.45	99.97	29.52	Used
PC aa C36:6	Phosphatidylcholine diacyl C36:6	0.0	15.22	99.97	1.13	Used
PC aa C38:0	Phosphatidylcholine diacyl C38:0	0.0	15.09	99.97	3.29	Used
PC aa C38:1	Phosphatidylcholine diacyl C38:1	0.1	19.94	99.93	0.87	Used
PC aa C38:3	Phosphatidylcholine diacyl C38:3	0.0	7.21	99.97	54.08	Used
PC aa C38:4	Phosphatidylcholine diacyl C38:4	0.0	6.64	99.97	119.83	Used
PC aa C38:5	Phosphatidylcholine diacyl C38:5	0.0	9.96	99.97	62.43	Used
PC aa C38:6	Phosphatidylcholine diacyl C38:6	0.0	10.27	99.97	90.66	Used
PC aa C40:1	Phosphatidylcholine diacyl C40:1	0.0	15.62	9.05	0.47	Excluded
PC aa C40:2	Phosphatidylcholine diacyl C40:2	0.0	13.75	99.97	0.36	Used
PC aa C40:3	Phosphatidylcholine diacyl C40:3	0.0	12.85	99.97	0.66	Used
PC aa C40:4	Phosphatidylcholine diacyl C40:4	0.0	7.60	100.00	4.17	Used
PC aa C40:5	Phosphatidylcholine diacyl C40:5	0.0	6.43	99.97	11.53	Used
PC aa C40:6	Phosphatidylcholine diacyl C40:6	0.0	6.22	100.00	28.76	Used
PC aa C42:0	Phosphatidylcholine diacyl C42:0	0.0	13.59	99.97	0.60	Used
PC aa C42:1	Phosphatidylcholine diacyl C42:1	0.0	15.38	99.97	0.30	Used
PC aa C42:2	Phosphatidylcholine diacyl C42:2	0.0	15.10	99.97	0.21	Used
PC aa C42:4	Phosphatidylcholine diacyl C42:4	0.0	12.77	99.97	0.22	Used
PC aa C42:5	Phosphatidylcholine diacyl C42:5	0.0	10.74	99.97	0.43	Used
PC aa C42:6	Phosphatidylcholine diacyl C42:6	0.0	10.85	62.53	0.63	Used
PC ae C30:0	Phosphatidylcholine acyl-alkyl C30:0	0.0	31.78	99.71	0.48	Excluded
PC ae C30:1	Phosphatidylcholine acyl-alkyl C30:1	4.6	46.30	98.66	0.24	Excluded
PC ae C30:2	Phosphatidylcholine acyl-alkyl C30:2	0.0	17.44	92.22	0.16	Used
PC ae C32:1	Phosphatidylcholine acyl-alkyl C32:1	0.0	10.34	99.97	2.85	Used
PC ae C32:2	Phosphatidylcholine acyl-alkyl C32:2	0.0	12.20	99.97	0.75	Used
PC ae C34:0	Phosphatidylcholine acyl-alkyl C34:0	0.0	11.28	99.97	1.73	Used
PC ae C34:1	Phosphatidylcholine acyl-alkyl C34:1	0.0	11.88	99.97	10.56	Used
PC ae C34:2	Phosphatidylcholine acyl-alkyl C34:2	0.0	12.38	99.97	12.67	Used

PC ae C34:3	Phosphatidylcholine acyl-alkyl C34:3	0.0	9.93	99.97	8.38	Used
PC ae C36:0	Phosphatidylcholine acyl-alkyl C36:0	0.0	40.89	99.97	1.10	Excluded
PC ae C36:1	Phosphatidylcholine acyl-alkyl C36:1	0.0	12.61	99.97	8.40	Used
PC ae C36:2	Phosphatidylcholine acyl-alkyl C36:2	0.0	13.72	99.97	15.19	Used
PC ae C36:3	Phosphatidylcholine acyl-alkyl C36:3	0.0	12.59	99.97	8.59	Used
PC ae C36:4	Phosphatidylcholine acyl-alkyl C36:4	0.0	11.60	99.97	20.88	Used
PC ae C36:5	Phosphatidylcholine acyl-alkyl C36:5	0.0	9.39	99.97	13.85	Used
PC ae C38:0	Phosphatidylcholine acyl-alkyl C38:0	0.0	12.57	99.97	2.48	Used
PC ae C38:1	Phosphatidylcholine acyl-alkyl C38:1	0.0	14.05	99.97	0.82	Used
PC ae C38:2	Phosphatidylcholine acyl-alkyl C38:2	0.0	13.49	99.97	2.15	Used
PC ae C38:3	Phosphatidylcholine acyl-alkyl C38:3	0.0	10.85	99.97	4.34	Used
PC ae C38:4	Phosphatidylcholine acyl-alkyl C38:4	0.0	12.38	99.97	15.73	Used
PC ae C38:5	Phosphatidylcholine acyl-alkyl C38:5	0.0	11.10	100.00	19.96	Used
PC ae C38:6	Phosphatidylcholine acyl-alkyl C38:6	0.0	9.18	99.97	8.70	Used
PC ae C40:0	Phosphatidylcholine acyl-alkyl C40:0	0.0	8.03	1.14	10.25	Excluded
PC ae C40:1	Phosphatidylcholine acyl-alkyl C40:1	0.0	12.62	99.97	1.68	Used
PC ae C40:2	Phosphatidylcholine acyl-alkyl C40:2	0.0	11.32	99.97	2.10	Used
PC ae C40:3	Phosphatidylcholine acyl-alkyl C40:3	0.0	10.64	99.97	1.14	Used
PC ae C40:4	Phosphatidylcholine acyl-alkyl C40:4	0.0	10.30	99.97	2.59	Used
PC ae C40:5	Phosphatidylcholine acyl-alkyl C40:5	0.0	8.88	99.97	3.57	Used
PC ae C40:6	Phosphatidylcholine acyl-alkyl C40:6	0.0	11.23	99.97	5.06	Used
PC ae C42:0	Phosphatidylcholine acyl-alkyl C42:0	0.0	18.33	14.80	0.52	Excluded
PC ae C42:1	Phosphatidylcholine acyl-alkyl C42:1	0.0	13.91	99.97	0.38	Used
PC ae C42:2	Phosphatidylcholine acyl-alkyl C42:2	0.0	17.58	99.97	0.68	Used
PC ae C42:3	Phosphatidylcholine acyl-alkyl C42:3	0.0	11.87	99.97	0.87	Used
PC ae C42:4	Phosphatidylcholine acyl-alkyl C42:4	0.0	9.99	100.00	1.01	Used
PC ae C42:5	Phosphatidylcholine acyl-alkyl C42:5	0.0	7.27	99.93	2.36	Used
PC ae C44:3	Phosphatidylcholine acyl-alkyl C44:3	0.0	13.32	99.97	0.11	Used
PC ae C44:4	Phosphatidylcholine acyl-alkyl C44:4	0.0	11.71	99.97	0.43	Used
PC ae C44:5	Phosphatidylcholine acyl-alkyl C44:5	0.0	7.15	99.97	2.12	Used
PC ae C44:6	Phosphatidylcholine acyl-alkyl C44:6	0.0	7.73	99.97	1.38	Used
lysoPC a C14:0	lysoPhosphatidylcholine acyl C14:0	0.0	26.82	42.21	3.21	Excluded
lysoPC a C16:0	lysoPhosphatidylcholine acyl C16:0	0.0	10.69	99.97	94.07	Used
lysoPC a C16:1	lysoPhosphatidylcholine acyl C16:1	0.0	10.01	99.97	2.90	Used
lysoPC a C17:0	lysoPhosphatidylcholine acyl C17:0	0.0	13.05	99.97	1.72	Used
lysoPC a C18:0	lysoPhosphatidylcholine acyl C18:0	0.0	10.27	99.97	25.96	Used
lysoPC a C18:1	lysoPhosphatidylcholine acyl C18:1	0.0	11.29	99.97	19.22	Used
lysoPC a C18:2	lysoPhosphatidylcholine acyl C18:2	0.0	9.42	99.97	27.22	Used
lysoPC a C20:3	lysoPhosphatidylcholine acyl C20:3	0.0	10.95	99.97	2.38	Used
lysoPC a C20:4	lysoPhosphatidylcholine acyl C20:4	0.0	9.34	99.97	6.77	Used
lysoPC a C24:0	lysoPhosphatidylcholine acyl C24:0	0.0	21.21	8.04	0.36	Excluded
lysoPC a C26:0	lysoPhosphatidylcholine acyl C26:0	0.0	32.22	59.85	0.54	Excluded
lysoPC a C26:1	lysoPhosphatidylcholine acyl C26:1	0.0	10.71	0.00	2.02	Excluded
lysoPC a C28:0	lysoPhosphatidylcholine acyl C28:0	0.0	27.17	46.46	0.48	Excluded
lysoPC a C28:1	lysoPhosphatidylcholine acyl C28:1	0.0	22.50	99.84	0.62	Used
lysoPC a C6:0	lysoPhosphatidylcholine acyl C6:0	0.0	43.89	25.51	0.02	Excluded
SM (OH) C14:1	Hydroxysphingomyeline C14:1	0.0	12.85	100.00	6.18	Used
SM (OH) C16:1	Hydroxysphingomyeline C16:1	0.0	8.72	99.97	3.35	Used
SM (OH) C22:1	Hydroxysphingomyeline C22:1	0.0	14.23	99.97	13.43	Used
SM (OH) C22:2	Hydroxysphingomyeline C22:2	0.0	13.12	99.97	11.40	Used
SM (OH) C24:1	Hydroxysphingomyeline C24:1	0.0	17.05	99.97	1.34	Used
SM C16:0	Sphingomyelin C16:0	0.0	12.92	99.97	105.98	Used
SM C16:1	Sphingomyelin C16:1	0.0	11.64	99.97	15.97	Used
SM C18:0	Sphingomyelin C18:0	0.0	9.29	99.97	23.16	Used
SM C18:1	Sphingomyelin C18:1	0.0	10.86	100.00	11.25	Used
SM C20:2	Sphingomyelin C20:2	0.1	15.99	99.97	0.38	Used
SM C22:3	Sphingomyelin C22:3	43.6	60.99	99.51	0.22	Excluded
SM C24:0	Sphingomyelin C24:0	0.0	14.33	99.97	21.68	Used
SM C24:1	Sphingomyelin C24:1	0.0	15.01	100.00	52.40	Used
SM C26:0	Sphingomyelin C26:0	0.0	57.33	99.97	0.32	Excluded
SM C26:1	Sphingomyelin C26:1	0.0	22.75	99.97	0.42	Used
H1	Sum of Hexoses	0.0	6.33	99.97	5197.44	Used

Table S2. List of 26 metabolites significantly associated with incident chronic kidney disease in either basic or full model in hyperglycemic individuals

Odds ratios (*OR*s) with 95% *CI* and *P*-values of multivariable logistic regression are shown. The basic model was adjusted for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, and fasting serum glucose. The full model was additionally adjusted for use of lipid lowering drugs, antihypertensive and anti-diabetic medication, baseline estimated glomerular filtration rate and urinary albumin-to-creatinine ratio. *P*-values shown in bold represent statistical significance at 0.05 level. **Abbreviations**: SM, sphingomyelin; PC aa, phosphatidylcholine diacyl; PC ae, phosphatidylcholine acyl-alkyl.

Metabolites	Basic Model		Full Model	
	***OR* (95% *CI*)**	***P*-value**	***OR* (95% *CI*)**	***P*-value**
C10	1.42 (1.03 - 1.98)	**3.317E-02**	1.24 (0.86 - 1.80)	2.495E-01
C12	1.49 (1.09 - 2.05)	**1.268E-02**	1.35 (0.95 - 1.92)	9.131E-02
C14:1	1.37 (1.04 - 1.83)	**2.919E-02**	1.36 (0.99 - 1.89)	5.751E-02
C18	1.44 (1.06 - 1.98)	**2.331E-02**	1.30 (0.92 - 1.84)	1.376E-01
C18:1	1.44 (1.07 -1.97)	**1.892E-02**	1.39 (0.99 - 1.96)	6.293E-02
C6 (C4:1-DC)	1.41 (1.05 -1.89)	**2.244E-02**	1.25 (0.90 - 1.75)	1.884E-01
C8	1.39 (1.02 -1.90)	**3.948E-02**	1.21 (0.85 - 1.71)	2.919E-01
Arginine	1.40 (1.07 -1.89)	**2.154E-02**	1.25 (0.93 - 1.73)	1.577E-01
Proline	1.38 (1.01 -1.89)	**4.453E-02**	1.39 (0.98 - 1.97)	6.337E-02
PC aa C32:2	0.72 (0.56 - 0.93)	**1.275E-02**	0.74 (0.56 - 0.99)	**3.690E-02**
PC aa C38:0	1.51 (1.12 - 2.07)	**8.059E-03**	1.56 (1.12 - 2.21)	**1.043E-02**
PC aa C42:0	1.41 (1.04 - 1.92)	**2.686E-02**	1.40 (1.01 - 1.96)	**4.801E-02**
PC ae C38:6	1.41 (1.01 - 1.99)	**4.573E-02**	1.40 (0.96 - 2.06)	8.386E-02
PC ae C40:5	1.42 (1.04 - 1.95)	**3.009E-02**	1.32 (0.94 - 1.88)	1.181E-01
PC ae C40:6	1.54 (1.12 - 2.14)	**9.600E-03**	1.57 (1.10 - 2.27)	**1.358E-02**
PC ae C42:5	1.43 (1.06 - 1.96)	**2.234E-02**	1.29 (0.92 - 1.81)	1.457E-01
SM (OH) C14:1	1.50 (1.06 - 2.13)	**2.277E-02**	1.56 (1.07 - 2.32)	**2.382E-02**
SM (OH) C16:1	1.59 (1.14 - 2.24)	**6.923E-03**	1.63 (1.14 - 2.39)	**9.614E-03**
SM (OH) C22:2	1.58 (1.09 - 2.33)	**1.880E-02**	1.50 (1.00 - 2.30)	5.674E-02
SM C16:0	1.91 (1.29 - 2.91)	**1.811E-03**	1.82 (1.17 - 2.91)	**9.378E-03**
SM C16:1	1.91 (1.29 - 2.88)	**1.557E-03**	1.85 (1.19 - 2.94)	**7.145E-03**
SM C18:0	1.86 (1.34 - 2.63)	**2.839E-04**	1.80 (1.26 - 2.63)	**1.754E-03**
SM C18:1	2.25 (1.54 - 3.39)	**4.976E-05**	2.22 (1.46 - 3.49)	**3.315E-04**
SM C20:2	1.40 (1.05 - 1.93)	**3.045E-02**	1.51 (1.10 - 2.14)	**1.411E-02**
SM C24:1	1.62 (1.15 - 2.31)	**7.066E-03**	1.57 (1.08 - 2.33)	**2.061E-02**
SM C26:1	1.41 (1.05 - 1.93)	**2.564E-02**	1.57 (1.13 - 2.23)	**8.215E-03**

Table S3. Baseline characteristics of propensity scores matched case-control hyperglycemic individuals

Clinical variables of incident CKD patients (= cases) matched with non-CKD participants (= controls) are shown. Mean ± standard deviation is provided when appropriate; *P*-values were calculated by univariate conditional logistic regression. *P*-values shown in bold represent statistical significance at 0.05 level. **Abbreviations:** CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; UACR, urinary albumin-to-creatinine ratio.

Clinical variables	**Incident CKD N = 62**	**Non-CKD N = 62**	***P*-value**
Age, years	65.81 ± 9.3	65.48 ± 7.62	0.777
Sex, Male, n (%)	54.84	64.52	0.261
BMI, kg/m^2	30.53 ± 4.84	29.79 ± 3.97	0.335
Fasting glucose, mg/dl	112.68 ± 27.31	114.32 ± 19.32	0.676
Systolic blood pressure, mmHg	130.03 ± 19.79	130.83 ± 16.38	0.819
Triglyceride, mg/dl [a]	136.5 [99.5 - 186]	129 [93.5 - 182.75]	0.784
Total cholesterol, mg/dl	215 ± 38.05	211 ± 33.11	0.481
HDL cholesterol, mg/dl	51.81 ± 11.59	51.66 ± 14.29	0.951
eGFR, mL/min/1.73 m²	80.17 ± 14.79	81.95 ± 10.92	0.339
UACR, mg/g [a]	8.89 [4.44 - 13.41]	6.8 [4.85 - 14.36]	0.842
Smoking, %			
Non-smoker	43.55	41.94	Reference
Former smoker	50	53.23	0.704
Current smoke	6.45	4.84	0.729
Medication usage, %			
Lipid-lowering	19.35	25.81	0.396
Antihypertensive	62.9	61.29	0.842
Anti-diabetic	14.52	16.13	0.796

[a] values are presented as median [25th- 75th percentile].

Table S4. Results of sensitivity analyses - the two metabolites significantly associated with incident chronic kidney disease in the propensity scores matched case-control hyperglycemic individuals

Odds ratios (*OR*s) per standard deviation (SD) with 95% *CI* and *P*-values of conditional logistic regression results are shown. *P*-values shown in bold represent statistical significance at 0.05 level. **Abbreviations**: SM, sphingomyelin; PC aa, phosphatidylcholine diacyl.

	SM C18:1	**PC aa C38:0**
***OR* (95% *CI*), per SD**	1.77 (1.14 - 2.73)	1.71 (1.12 - 2.62)
***P*- value**	**0.011**	**0.014**

Table S5. Results of sensitivity analyses - interaction effects of the two metabolites with different glucose subgroups

Odds ratios (*OR*s) with 95% *CI* and *P*-values of multivariate logistic regression results are shown. $P_{interaction}$ represents *P*-value of multiplicative interaction effects between metabolite and different glucose groups. *P*-values shown in bold represent statistical significance at 0.05 level. **Abbreviations**: SM, sphingomyelin; PC aa, phosphatidylcholine diacyl; NGT, normal glucose tolerance; 2-h glucose, two hour post load glucose.

Group	SM C18:1			PC aa C38:0		
	***OR* (95% *CI*)**	***P* - values**	**$P_{interaction}$**	***OR* (95% *CI*)**	***P* - values**	**$P_{interaction}$**
Glycemic status			**1.774E-03**[c]			0.417[c]
NGT[a]	0.76 (0.57 - 1.01)	0.057		1.21 (0.95 - 1.55)	0.124	
Hyperglycemia[b]	2.22 (1.46 - 3.49)	**3.315E-04**		1.56 (1.12 - 2.21)	**0.010**	
Fasting glucose			0.241[d]			0.609[d]
1st tertile[a]	0.78 (0.46 - 1.36)	0.372		1.13 (0.73 - 1.77)	0.579	
2nd tertile[a]	0.84 (0.56 - 1.27)	0.412		1.33 (0.94 - 1.88)	0.106	
Top tertile[b]	1.50 (1.08 - 2.11)	**0.019**		1.49 (1.10 - 2.03)	**0.010**	
2-h glucose			**0.010**[e]			0.538[e]
1st tertile[a]	0.55 (0.33 - 0.92)	**0.023**		1.22 (0.79 - 1.87)	0.369	
2nd tertile[a]	0.74 (0.48 - 1.14)	0.172		1.27 (0.87 - 1.88)	0.231	
Top tertile[b]	1.58 (1.07 - 2.37)	**0.022**		1.60 (1.17 - 2.23)	**0.004**	

[a] with adjustments for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, fasting glucose, use of lipid lowering drugs, antihypertensive medication, baseline estimated glomerular filtration rate and baseline urinary albumin-to-creatinine ratio.

[b] with adjustment for the covariates shown in [a] as well as use of anti-diabetic medication.

[c] The model setting : $\text{logit}(P) = \beta_0 + \beta_1$*metabolite + β_2*glycemic status + β_3* metabolite * glycemic status + β_4*covariates + ε. The covariates including the covariates shown in [a] as well as use of anti-diabetic medication.

[d] The model setting : $\text{logit}(P) = \beta_0 + \beta_1$*metabolite + β_2* three tertiles group of fasting glucose + β_3* metabolite * three tertiles group of fasting glucose + β_4*covariates + ε. The covariates included the covariates shown in [a] as well as use of anti-diabetic medication except fasting glucose.

[e] The model setting : $\text{logit}(P) = \beta_0 + \beta_1$*metabolite + β_2* three tertiles group of 2-h glucose + β_3* metabolite * three tertiles group of 2-h glucose + β_4*covariates + ε. The covariates included the covariates shown in [a] except fasting glucose.

Table S6. Results of sensitivity analyses - association of two candidate biomarkers with UACR- and eGFR- based incident CKD in hyperglycemic participants

Odds ratios (*OR*s) with 95% *CI* and *P*-values of each metabolite with UACR-based and eGFR-based incident CKD in basic and full multivariable logistic regression models are shown, respectively. UACR-based incident CKD was defined as UACR ≥ 30 mg/g at follow-up (FF4). eGFR-based incident CKD was defined as eGFR < 60 ml/min/1.73 m^2 at follow-up (FF4). Basic model was adjusted for age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol and fasting glucose. Full model was additionally adjusted for use of lipid lowering drugs, antihypertensive and anti-diabetic medication, baseline eGFR and UACR. *P*-values shown in bold represent statistical significance at 0.05 level. **Abbreviations**: CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; UACR, urinary albumin-to-creatinine ratio; SM, sphingomyelin; PC aa, phosphatidylcholine diacyl.

	SM C18:1		**PC aa C38:0**	
	Basic model	**Full model**	**Basic model**	**Full model**
UACR- based incident CKD (N = 32) & non-CKD (N = 353)				
P-value	**0.024**	**0.040**	**0.022**	**0.004**
OR (95 % *CI*) , per SD	1.79 (1.10 - 3.03)	1.80 (1.05 - 3.25)	1.66 (1.08 - 2.58)	2.17 (1.31 - 3.76)
eGFR- based incident CKD (N = 65) & non-CKD (N = 320)				
P-value	**0.008**	0.107	0.061	0.247
OR (95 % *CI*) , per SD	1.77 (1.17 - 2.75)	1.50 (0.93 - 2.5)	1.38 (0.99 - 1.94)	1.25 (0.86 - 1.85)

Table S7. Comparison of the predictive performances of two sets of predictors of incident chronic kidney disease in hyperglycemic individuals with three machine learning approaches

The median AUC (95% *CI*) of three machine learning approaches over 100 random repeats of 10-fold cross validation are shown. Reference predictors consists of baseline age, sex, estimated glomerular filtration rate and urinary albumin-to-creatinine ratio. Developed sets includes combined metabolites and clinical variables that were selected by the three-step feature selection in each round. **Abbreviation**: AUC, area under the receiver operating characteristic curve.

Algorithms	**Models**	**Median AUC (95% *CI*)**	**Absolute increase in median prediction**	**Outperform times over 100 times**
Support Vector Machine	Reference predictors	0.800 (0.783 - 0.816)	2.5%	97
	Developed sets	0.825 (0.801 - 0.849)		
Random Forest	Reference predictors	0.789 (0.771 - 0.807)	2.9%	100
	Developed sets	0.818 (0.794 - 0.836)		
Adaptive Boosting	Reference predictors	0.798 (0.781 - 0.813)	1.6%	87
	Developed sets	0.814 (0.787 - 0.832)		

Table S8. The total selected times for three most frequently selected sets of metabolites and clinical variables over 1000 selection rounds in 100 times of 10-fold cross validation

The three most frequently selected sets of metabolites and clinical variables, as well as their total selected times over 1000 selection rounds are shown. **Abbreviations:** eGFR, estimated glomerular filtration rate; UACR, urinary albumin-to-creatinine ratio; SM, sphingomyelin; PC aa, phosphatidylcholine diacyl.

Sets of metabolites and clinical variables	Selected times
SM C18:1, PC aa C38:0, age, total cholesterol, fasting glucose, eGFR, UACR	113
SM C18:1, age, total cholesterol, fasting glucose, eGFR, UACR	78
SM C18:1, PC aa C38:0, proline, age, total cholesterol, fasting glucose, eGFR, UACR	67

Table S9. The selected times for 15 most important variables over 1000 selection rounds in 100 times of 10-fold cross validation

Out of 125 metabolites and 14 clinical variables, 15 most frequently selected variables and their total selected times over 1000 selection rounds are shown. **Abbreviations:** UACR, urinary albumin-to-creatinine ratio; eGFR, estimated glomerular filtration rate; SM, sphingomyelin; PC aa, phosphatidylcholine diacyl.

Variables	Selected times
UACR	1000
eGFR	1000
Age	999
Total cholesterol	996
Fasting glucose	942
SM C18:1	857
PC aa C38:0	593
Triglyceride	270
Proline	229
PC aa C32:2	156
Tyrosine	129
SM C26:1	109
C18:1	108
PC aa C36:4	92
Use of lipid lowering drugs	81

Table S10. Predictive performance of the best set of predictors and the full model of incident CKD in hyperglycemia

Mean AUC values of the best set of predictors and the full model of incident CKD in hyperglycemia are shown. The mean AUC value of the best set of predictors was the average value of the AUC values of the 113 selected times, in which the models were fitted with support vector machine. The average AUC value of the full model was obtained using logistic regression with 10 times of 10-fold cross validation. **Abbreviations:** CKD, chronic kidney disease; AUC, area under the receiver operating characteristic curve; UACR, urinary albumin-to-creatinine ratio; eGFR, estimated glomerular filtration rate; SM, sphingomyelin; PC aa, phosphatidylcholine diacyl.

Models	Mean AUC	Absolute increase in mean prediction
The best set of predictors (i.e., SM C18:1, PC aa C38: 0, age, total cholesterol, fasting glucose, eGFR and UACR)	0.857	4.8%
The full model (i.e., age, sex, BMI, systolic blood pressure, smoking status, triglyceride, total cholesterol, HDL cholesterol, fasting glucose, use of lipid lowering drugs, antihypertensive and anti-diabetic medication, eGFR and UACR	0.809	

Supplementary Figures

Figure S1. Technical normalization across the study

Comparison of before and after normalization of plate effect of metabolite data using phosphatidylcholine diacyl (PC aa) C34:2 as an example. Metabolite concentration drifts at 38 plates were independently corrected by conducting plate effect normalization in quality controls samples (QCs, shown in plots A and B) and KORA F4 individual samples (plots C and D).

Fig. S1

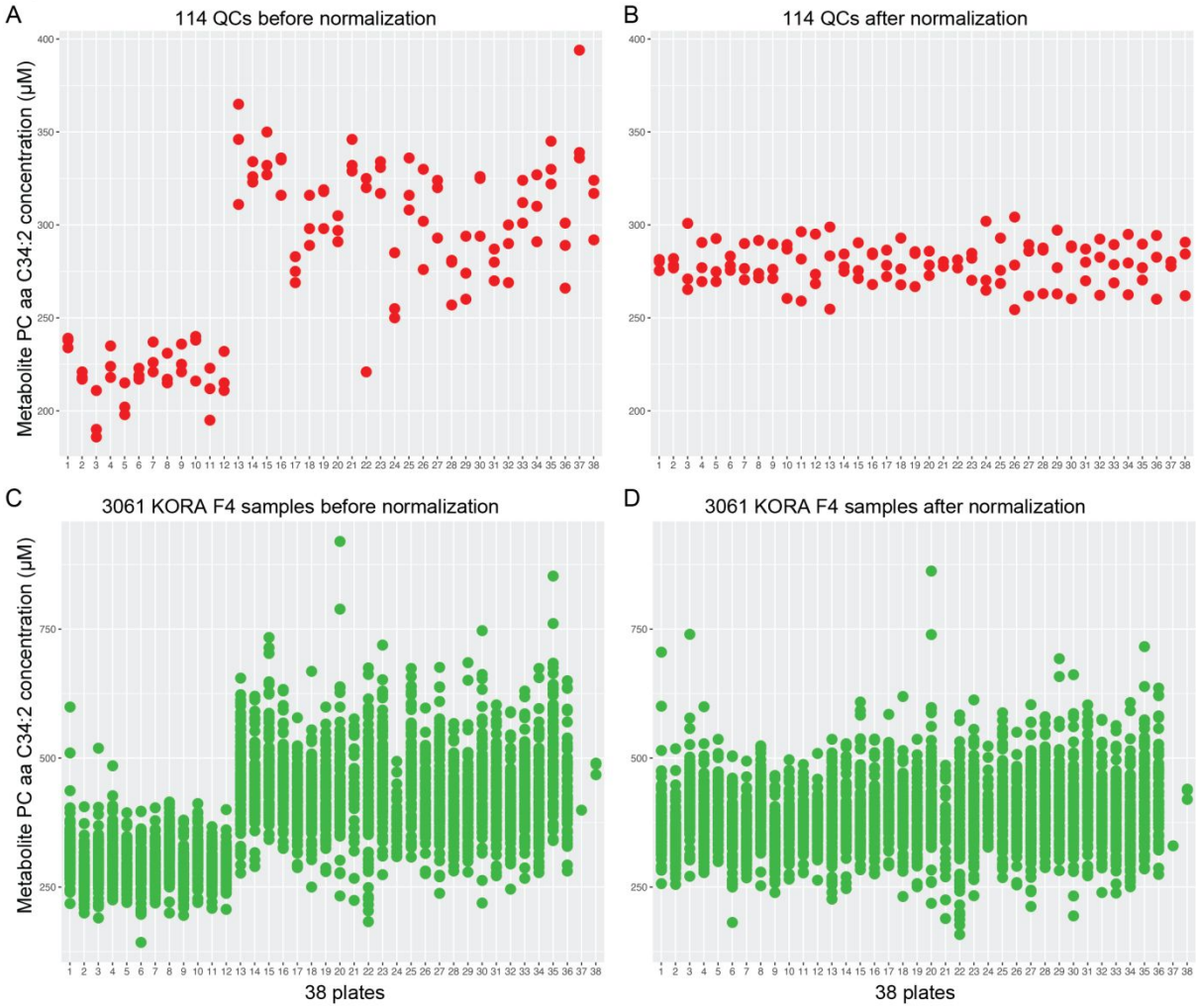

Figure S2. Correlation of nine sphingomyelins in 385 hyperglycemic participants

Pearson's correlation coefficients values of nine sphingomyelins (SMs) in 385 participants with pre-diabetes and T2D are shown. Both the size of the cycle and intensity of color indicate the degree of correlation between the metabolites. The numeric values of Pearson's correlation coefficients are shown in the bottom triangle.