Machine Learning Approaches Reveal Metabolic Signatures of Incident Chronic Kidney Disease in Individuals With Prediabetes 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] Cornelia Prehn,[4] Li Wang,[1,2,5] Jana Nano,[2,3] Markus F. Scheerer,[6] Susanne Neschen,[6] Gabi Kastenmüller,[7] Karsten Suhre,[8] Michael Laxy,[9] Freimut Schliess,[10] Christian Gieger,[1,2,3] Jerzy Adamski,[4,11,12] Martin Hrabe de Angelis,[3,6,12] Annette Peters,[2,3] and Rui Wang-Sattler[1,2,3]

Diabetes 2020;69:2756–2765 | https://doi.org/10.2337/db20-0586

Early and precise identification of individuals with prediabetes and type 2 diabetes (T2D) at risk for 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 C18:1 and phosphatidylcholine diacyl 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 people with prediabetes and T2D. The metabolite link with hyperglycemia-related early kidney dysfunction warrants further investigation.

Chronic kidney disease (CKD) affects approximately 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).

[1]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[2]Institute of Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[3]German Center for Diabetes Research (DZD), München-Neuherberg, Germany
[4]Research Unit of Molecular Endocrinology and Metabolism, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[5]Department of Scientific Research and Shandong University Postdoctoral Work Station, Liaocheng People's Hospital, Shandong, P. R. China
[6]Institute of Experimental Genetics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[7]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[8]Department of Physiology and Biophysics, Weill Cornell Medicine - Qatar, Doha, Qatar
[9]Institute of Health Economics and Health Care Management, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[10]Profil Institut für Stoffwechselforschung GmbH, Neuss, Germany
[11]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[12]Chair of Experimental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Freising, Germany

Corresponding author: Rui Wang-Sattler, rui.wang-sattler@helmholtz-muenchen.de

Received 3 June 2020 and accepted 29 September 2020

This article contains supplementary material online at https://doi.org/10.2337/figshare.13022624.

S.Z. is currently affiliated with the Institute for Vascular Signalling, Centre for Molecular Medicine, Goethe University, Frankfurt am Main, Germany.

M.F.S. is currently affiliated with Medical Affairs & Pharmacovigilance, Bayer AG, Berlin, Germany.

S.N. is currently affiliated with Sanofi Aventis Deutschland GmbH, Frankfurt am Main, Germany.

Among the established risk factors for CKD, diabetes 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 prediabetes have been related to high prevalence of CKD in U.S., European, and Asian populations (4–7). Early screening of hyperglycemic individuals at 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 in combination 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 prediabetes and T2D 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, prediabetes, and T2D (12–19). Several studies have investigated the metabolite profiles of CKD, both in the general population and populations with T2D (20–22). However, to the best of our knowledge, none of them have explored the metabolites associated with future development of CKD in people with prediabetes or 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], and 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 prediabetes or 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. For the second follow-up (FF4), 2,269 participants were examined 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 2-h postload glucose (2-h glucose) values with the World Health Organization diagnostic criteria (25). The hyperglycemic group comprised participants with prediabetes and newly diagnosed T2D (i.e., fasting glucose ≥110 mg/dL 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 antidiabetes agents (13,23).

We examined 2,142 individuals who participated in both KORA F4 and FF4. Exclusion criteria were *1*) nonfasting 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), and *3*) diagnosis for type 1 diabetes ($n = 6$ at F4), unclear type of diabetes ($n = 21$ at F4), or CKD ($n = 173$ at F4). The remaining data set comprised 385 hyperglycemic participants and 1,453 individuals with NGT (Fig. 1 and 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) (isotope dilution mass spectrometry–standardized and International Federation of Clinical Chemistry and Laboratory Medicine–standardized values) using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation (26). Non-CKD was defined as an eGFR ≥60 mL/min/1.73 m^2 and an UACR <30 mg/g at both F4 and FF4 (9). Incident cases of CKD were defined as no CKD at baseline (F4) but reduced kidney function (eGFR <60 mL/min/1.73 m^2) or kidney damage (UACR ≥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 AbsoluteIDQ 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 (Supplementary Table 1). Each plate also contained three quality control (QC) samples (sex-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% and *2*) 50% of all measured

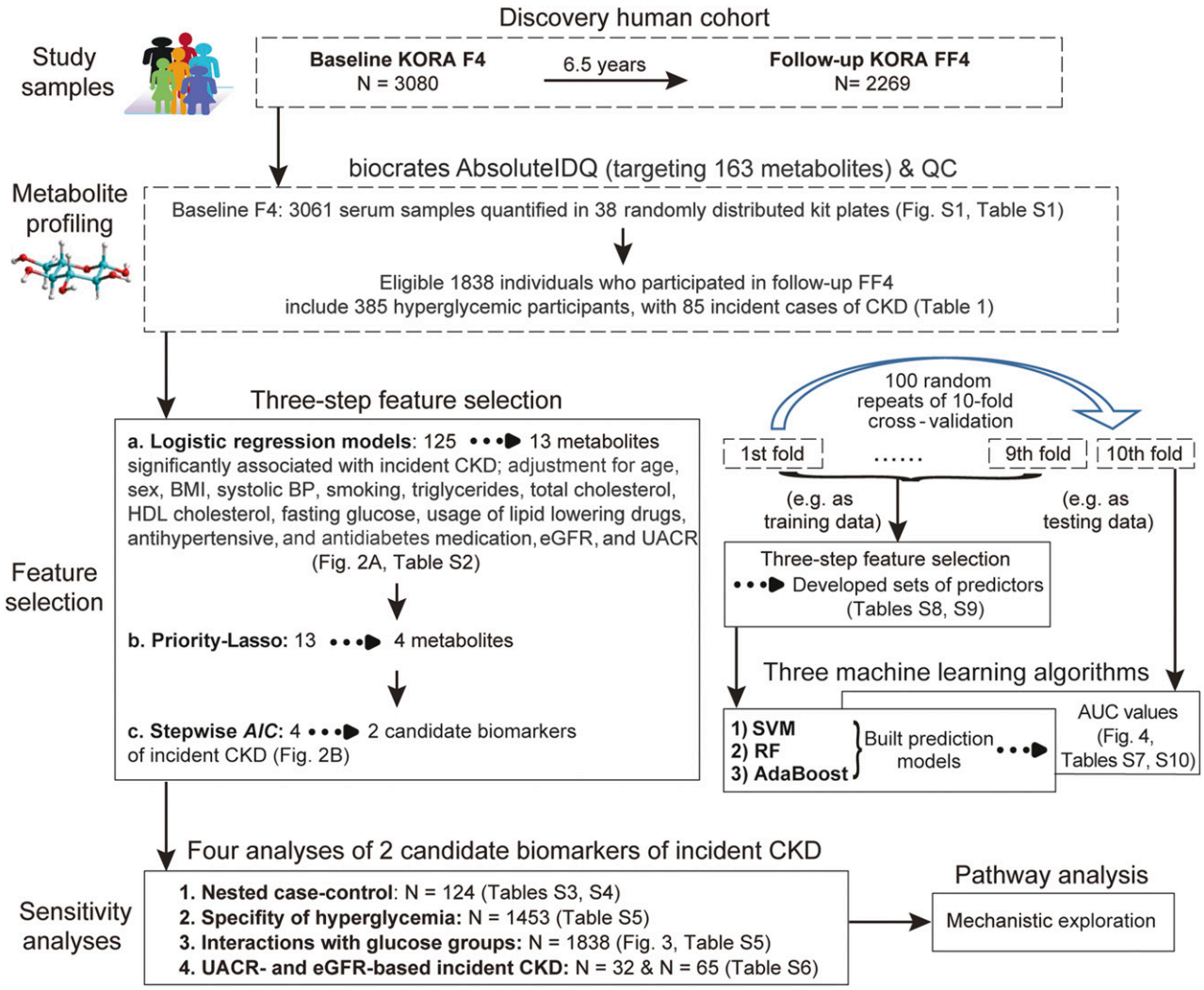

Figure 1—Study design. Fig. S1 and Tables S1–S10 refer to Supplementary Fig. 1 and Supplementary Tables 1–10 and are available in the Supplementary Material.

sample concentrations equal to or above three times the median of the 38 zero samples. In total, 125 metabolites passed the criteria and were used in the subsequent analysis (Supplementary Table 1). For minimization of the plate effect, metabolite concentrations were adjusted for the plate normalization factors. For each metabolite, the plate normalization factors were calculated by division of the mean of QC sample values in each plate with the mean of all QC sample values in 38 plates. As shown in Supplementary Fig. 1, plate normalization efficiently corrected the inter-plate variations in metabolite concentration.

For comparability between different metabolites, their concentrations were natural-log transformed and scaled to a mean value of 0 and SD of 1.

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). The basic model was adjusted for age, sex, BMI, systolic blood pressure (BP), smoking status, triglyceride, total cholesterol, HDL cholesterol, and fasting glucose. The full model was additionally adjusted for the use of lipid-lowering, antihypertensive, and antidiabetes 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 multicollinearity of included variables and to retain metabolite and clinical variables with nonzero coefficients. Priority-Lasso is a Lasso-based intuitive procedure that uses 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

Table 1—Characteristics of the KORA study population

	Hyperglycemic participants			NGT participants		
Clinical variables	Incident CKD, $N = 85$	Non-CKD, $N = 300$	P	Incident CKD, $N = 115$	Non-CKD, $N = 1,338$	P
Age, years	67.78 ± 8.78	59.44 ± 9.39	**1.29E−10**	60.97 ± 12	50.05 ± 10.82	**4.81E−20**
Male sex, %	55.29	58.00	0.656	46.09	46.64	0.910
BMI, kg/m^2	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[a]	159.82 ± 39.87[a]	**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	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 m^2	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 m^2	57.5 ± 18.3	81.67 ± 13.12		66.68 ± 19.32	89.5 ± 13.48	
Baseline UACR, mg/g	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	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
Nonsmoker	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**
Antidiabetes	16.47	11.33	0.208	0	0	—

Data are means ± SD for quantitative variables or median (25th–75th percentile) unless otherwise indicated. KORA participants were classified according to their hyperglycemic status at baseline (F4) and incident CKD status at follow-up (FF4). Unless indicated, variables show baseline measurements. P values were calculated by univariate logistic regression. P values shown in boldface type represent statistical significance at 0.05 level. [a]In the hyperglycemic participants, 2-h glucose levels were only available in 61 individuals with incident CKD and 254 individuals without CKD.

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 case and control subjects 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 one-to-one 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 1,453 normoglycemic participants by MLR. *3)* We explored the interaction effects of candidate biomarkers with glucose levels for incident CKD in 1,838 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 ≥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 nonoverlapping 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 data sets. 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 data set. The AUC values of the 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. For assessment of 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). 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 (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 (type of boosting algorithm to perform), and iter (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 1,000 times of three-step feature selection. The most frequently selected set of metabolites and clinical variables among these 1,000 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 on 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 and 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 CKD case subjects in hyperglycemic and NGT groups were significantly older and had significantly higher baseline values of HbA_{1c}, fasting and 2-h glucose, and UACR, whereas their baseline eGFR values were significantly lower. They also self-reported a significantly higher intake of antihypertensive 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. 2*A* and Supplementary Table 2). Among the 13 metabolites, nine corresponded to sphingomyelins (SMs) and SM C18:1 remained significant after stringent Bonferroni correction (Fig. 2*A* and Supplementary Fig. 2). Of the 13 metabolites, 4 metabolites were selected by priority-Lasso and 2 (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 case subjects in comparison with 300 non-CKD individuals (Fig. 2*B*). 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 $P = 3.315E-04$) (Supplementary Table 2).

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 case and matched

control subjects (Supplementary Table 3), and the two candidate biomarkers showed significant risk associations with incident CKD (Supplementary Table 4).

Both metabolites were not significantly associated with incident CKD in 1,453 normoglycemic individuals, i.e., when 115 incident CKD case subjects were compared with 1,338 non-CKD individuals who were both NGT at baseline (Table 1, Supplementary Table 5, and Fig. 2*B*). 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 (Supplementary Table 5). Moreover, SM C18:1 demonstrated significant multiplicative interaction effects with glycemic status and 2-h glucose (Fig. 3 and Supplementary Table 5).

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 in full MLR (Supplementary Table 6). Moreover, SM C18:1 was a significant predictor for eGFR-based incident CKD in the basic MLR (Supplementary Table 6).

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 >0.813 (Fig. 4 and Supplementary Table 7). In comparisons with 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]) (Supplementary Table 7), thereby outperforming the reference predictors in 97 out of 100 times of 10-fold cross validation (Supplementary Table 7). The improvement remained consistent after application of the other two machine learning approaches, RF (2.9% absolute increase in median AUC value) and AdaBoost (1.6%) (Supplementary Table 7). 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

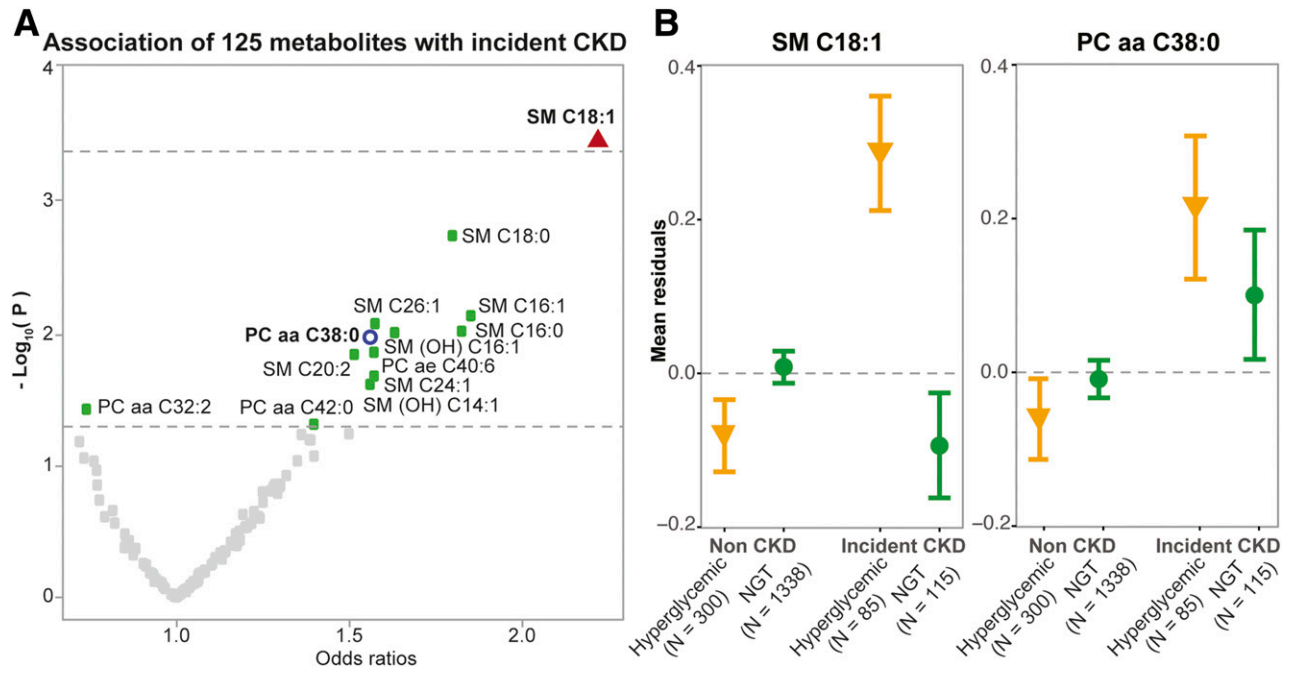

Figure 2—Serum metabolite associations with incident CKD. *A*: Volcano plot of the association results for 125 metabolites with incident CKD in hyperglycemic individuals. Odds ratios and *P* values are from logistic regression analysis adjusted for age, sex, BMI, systolic BP, smoking status, triglyceride, total cholesterol, HDL cholesterol, fasting glucose, use of lipid-lowering drugs, antihypertensive, and antidiabetes medication, and baseline values of eGFR and UACR. The upper and the lower interrupted lines represent Bonferroni-corrected and uncorrected ($P = 0.05$) significance levels, respectively. *B*: Mean residuals (with SEs) 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 BP, smoking status, triglyceride, total cholesterol, HDL cholesterol, and fasting glucose.

C18:1 and PC aa C38:0) and five clinical variables (age, total cholesterol, fasting glucose, eGFR, and UACR). This set was the most frequently selected set: 113 times over 1,000 selection rounds (Supplementary Table 8). 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 1,000 rounds (Supplementary Table 9). 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 (Supplementary Table 10).

DISCUSSION

This longitudinal study revealed significant accumulation of sphingo- and glycerophospholipids (SM C18:1 and PC aa C38:0) in individuals with prediabetes and T2D 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 and PC aa C38:0) and five clinical variables (age, total cholesterol, fasting glucose, eGFR, and 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 SM accumulation in glomerular diseases of genetic and nongenetic origin (32). Out of 125 analyzed metabolites comprising amino acids, acylcarnitines, hexoses, and glycerophospho- and sphingolipids (Supplementary Table 1), SMs represented the majority of metabolites associated with incident CKD in hyperglycemic participants ($P < 0.05$) (Fig. 2*A*). Increased SM levels in relation to CKD were also reported in individuals with type 1 diabetes (33) and T2D (34), except for the nontargeted lipidomic study of type 1 diabetes (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, and 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. 2*A*).

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 to 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

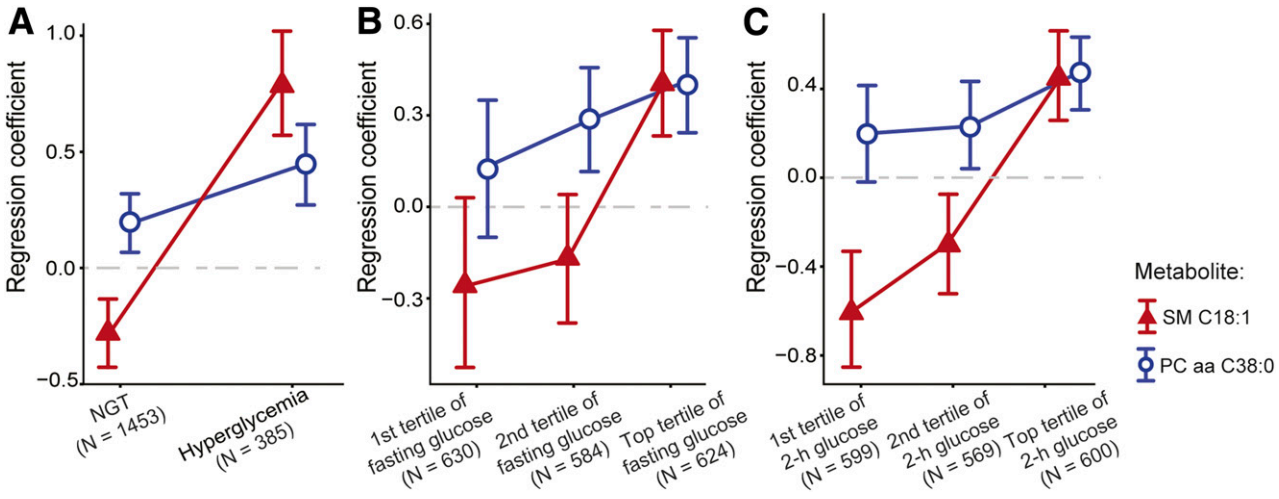

Figure 3—Stratified associations of candidate biomarkers with incident CKD according to glucose status. Associations of SM C18:1 and PC aa C38:0 with incident CKD stratified by hyperglycemic status (*A*) and each tertile of fasting glucose (*B*) and 2-h glucose (*C*) values. Regression coefficients in NGT and first and second tertile of fasting and 2-h glucose were adjusted for age, sex, BMI, systolic BP, smoking status, triglyceride, total cholesterol, HDL cholesterol, fasting glucose, use of lipid-lowering drug and antihypertensive medication, and baseline values of eGFR and UACR. Regression coefficients in the hyperglycemic group and the top tertile of fasting and 2-h glucose were additionally adjusted for antidiabetes medication.

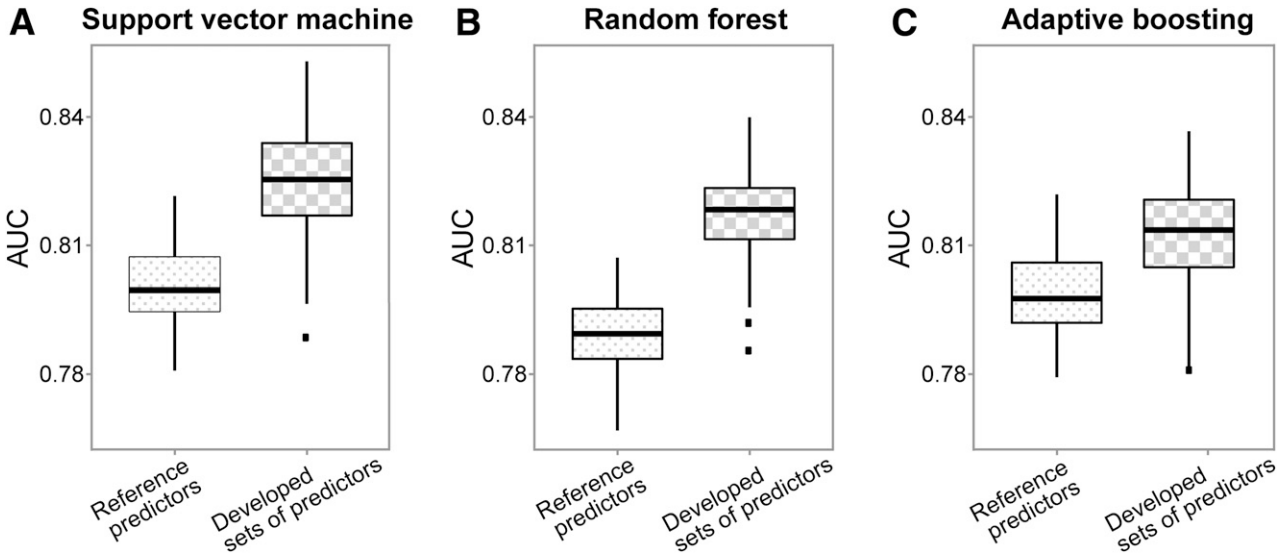

Figure 4—Prediction performance of incident CKD in hyperglycemic individuals in three machine learning approaches. The box plots show the AUC values of two models applying three machine learning approaches over 100 times of 10-fold cross validation. Reference predictors: baseline age, sex, eGFR, and UACR. 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 for 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 10 AUC values of each model of each approach were averaged to produce a single estimate that was displayed in box plots. 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.

cardiovascular disease, for which CKD is an independent risk factor (41).

Circulatory accumulation in SMs and saturated PCs in individuals with prediabetes and T2D may also reflect early stages of diabetic nephropathy such as mesangial matrix expansion, podocyte injury, and glomerular enlargement (42). The 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). The 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. 2*A*) and highly correlated with our top hit, SM C18:1 (Pearson correlation coefficient = 0.66, $P < 2.2e{-}16$) (Supplementary Fig. 2). Renal accumulation in SM (d18:1/16:0) was related to reduced enzyme activity of AMPK in the diabetic kidney glomeruli, mitochondrial dysfunction, and CKD progression (43).

The altered levels of certain SM and PC species in hyperglycemic individuals at 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 monounsaturated fatty acids (46), and enzymes involved in de novo synthesis and the ceramide-SM homeostasis such as SM synthase 2 (SMS2) show fatty acyl chain specificity and may determine the regional expression of SM species in the kidney (47). Reduction of 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 the testing

data set anymore (53). In our study, we used 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 for the training data and the performance evaluation was only performed for 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 significant progress.

Our study has several additional advantages. We used a well-characterized, population-based human cohort that allows for adjustment 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 the missing replication (of 10 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 prediabetes 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. The authors express appreciation to all KORA study participants for donating their blood and time. The authors thank the field staff in Augsburg conducting the KORA studies. The authors are grateful to the staff (J. Scarpa, K. Faschinger, N. Lindemann, A. Ludolph, S. Jelic, and B. Langer) from the Institute of Epidemiology and the Genome Analysis Center Metabolomics Platform, and KORA-PASST Platform at Helmholtz Zentrum München – German Research Center for Environmental Health, who helped in the sample and data logistics, and metabolomics measurements. Additionally, the authors thank Dr. Anne-Laure Boulesteix from the Institute of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University of Munich, Germany, for tips on statistical methods.

Funding. The KORA study was initiated and financed by 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 European Union Seventh Framework Programme (EU FP7) grant HEALTH-2013-2.4.2-1/602936 (Project CarTarDis) and the European Institute of Innovation and Technology (EIT) Health–supported 19076 and 20679 iPDM-GO "Integrated Personalized Diabetes Management Goes Europe" innovation project. 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 Medicine - Qatar, a program funded by the Qatar Foundation.

Duality of Interest. 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. No other potential conflicts of interest relevant to this article were reported.

Bayer AG was not involved in work related to data and manuscript generation.

Author Contributions. J.H. conceived the study, analyzed the data, and wrote the manuscript. C.H. researched cohort data and edited the manuscript. M.C. contributed to pathway analysis and wrote the manuscript. M.T. researched data and edited the manuscript. J. Adam edited the manuscript. S.Z. researched data. C.P. researched metabolomics data. L.W. edited the manuscript. J.N. edited the manuscript. M.F.S. researched data and edited the manuscript. S.N. researched data. G.K. researched metabolomics data. K.S. researched metabolomics data. M.L. reviewed the manuscript. F.S. edited the manuscript. C.G. researched cohort data. J.Adam. researched metabolomics data. M.H.d.A. researched data. A.P. researched cohort data. R.W.-S. designed the study, researched metabolomics data, and wrote the manuscript. R.W.-S. 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.

Prior Presentation. Parts of this study were presented in abstract form at the 15th Annual Conference of the Metabolomics Society The Hague, the Netherlands, 23–27 June 2019, and at the 7th DZD Diabetes Research School, Barcelona, Spain, 14–16 September 2019.

References

1. Bikbov B, Purcell CA, Levey AS, et al.; GBD Chronic Kidney Disease Collaboration. 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, et al.; CDC CKD Surveillance Team. 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, et al. 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, et al. 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, et al. 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:e000955

8. Ceriello A, Barkai L, Christiansen JS, et al. 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, et al.; Kidney Disease Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int Suppl 2013;3:1–150

10. Tangri N, Stevens LA, Griffith J, et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA 2011;305:1553–1559
11. Dunkler D, Gao P, Lee SF, et al.; ONTARGET and ORIGIN Investigators. 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, et al. 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, et al. Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol 2012;8:615
14. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and the risk of developing diabetes. Nat Med 2011;17:448–453
15. Chen GC, Chai JC, Yu B, et al. 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, et al. 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, et al. Untargeted metabolomics identifies a plasma sphingolipid-related signature associated with lifestyle intervention in prepubertal children with obesity. Int J Obes 2018;42:72–78
18. Razquin C, Toledo E, Clish CB, et al. 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, et al. 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, et al. 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, et al. 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, et al. Ageing investigation using two-time-point metabolomics data from KORA and CARLA studies. Metabolites 2019;9:44
25. World Health Organization, International Diabetes Federation. *Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycaemia: Report of a WHO/IDF Consultation*. Geneva,World Health Org., 2006
26. Inker LA, Schmid CH, Tighiouart H, et al.; CKD-EPI Investigators. 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. J Stat Softw 2006;17:1–27
32. Merscher S, Fornoni A. Podocyte pathology and nephropathy - sphingolipids in glomerular diseases. Front Endocrinol (Lausanne) 2014;5:127
33. Mäkinen VP, Tynkkynen T, Soininen P, et al. 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, et al. 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 2016;2:470–480
35. Tofte N, Suvitaival T, Ahonen L, et al. 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 isomericlipid species analyzed using theMxP®Quant 500 Kit. Available from https://www.biocrates.com/wp-content/uploads/2020/02/Biocrates_Q500_isomers_isobars.pdf
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, Kühn T, Sookthai D, et al. 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, et al. 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, et al. 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, et al. 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;20:5901
45. Johnson EL, Heaver SL, Waters JL, et al. Sphingolipids produced by gut bacteria enter host metabolic pathways impacting ceramide levels. Nat Commun 2020;11:2471
46. Czumaj A, Śledziński T, Carrero JJ, et al. Alterations of fatty acid profile may contribute to dyslipidemia in chronic kidney disease by influencing hepatocyte metabolism. Int J Mol Sci 2019;20:2470
47. Sugimoto M, Wakabayashi M, Shimizu Y, et al. 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, et al. Reducing plasma membrane sphingomyelin increases insulin sensitivity. Mol Cell Biol 2011;31:4205–4218
49. Fan Y, Shi F, Liu J, et al. 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, et al. 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, et al. 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, König IR. Statistical learning approaches in the genetic epidemiology of complex diseases. Hum Genet 2020;139:73–84