**Supplementary Appendix**

Dietrich S, Floegel A, Boeing H, Schulze MB, Illig T, Pischon T, Knüppel S, Drogan D. ‘Random Survival Forest in practice – a method for modelling complex metabolomics data in time to event analysis’

This supplementary material has been provided by the authors to give readers additional information about their work.

**Supplementary methods**

Random Survival Forest

The RSF method computes an ensemble of binary decision trees to construct a RSF model. For each decision trees a random bootstrap sample is drawn that includes on average two thirds of the original data. The remaining one third of the data is excluded and called out-of-bag (OOB) data. To grow a decision tree based on the bootstrap samples, a random node splitting process is applied. The random node splitting process works as follows: At each node a random set of candidate variables is chosen for random node splitting. The number of candidate variables corresponds with the square root of the total number of variables available in the data. Within the candidate variables, the one variable with a split point that maximizes the survival differences between daughter nodes is used for node splitting. As splitting rule, the log-rank statistic can be applied (1, 2). The log-rank statistic maximized the survival differences between daughter nodes by the following formula (1, 2):

Where L(x,c) is the log rank measure of node separation, c is the splitting point, x the predictor, N is the number of individuals in the parent node, di is the number of deaths at time ti in the daughter nodes, Yi are the individuals at risk at time ti and i is number of observations.

To calculate the log-rank statistics for each candidate variable several random split points are drawn for the respective candidate variable. The number of random split points that are drawn for each candidate variable, can be pre-defined before computation of a RSF model. This parameter is named *nsplit* in the RSF package of the statistic software R and was set equal to ten in the present study. The growth of a decision tree is continued until all terminal nodes contain only a minimal number of unique events which prevents further node splits.

Minimal depth measurement

Once a RSF model is computed, it can be assessed how informative a variable is regarding time until event using the so called minimal depth measurement (3). This measurement was described in detail by Ishwaran et al. (3). Briefly, the minimal depth of a respective variable is determined as follows: In each decision tree of a RSF the distance from the root node to the closest node were the respective variable splits first is determined. This procedure is illustrated in Figure S1. The metabolite 5 in Figure S1splits at a node with a minimal depth of 1 and also of 2. The minimal depth of 1 represents the closest split to the root node and is thus assigned to the variable for this decision tree. The value is recorded for the respective variable and in the following repeated in any further computed decision tree and finally averaged over the whole RSF.

RSF prediction error rate

To determine the prediction accuracy of a RSF model, the RSF prediction error rate can be calculated based on Harrell’s concordance index (C-index) (4). For this purpose, the OOB samples of each decision tree which have not been selected for the respective bootstrap sample are used and dropped down the respective decision tree. According to Harrell’s C-index, the probability is then estimated, that within a randomly selected pair of OOB samples with an event, the OOB sample with the shorter follow-up time has the worst predictive outcome (4, 5). The RSF prediction error rate is conform to 1 - C-index with values between 0 and 1, where a lower RSF prediction error rate correspond to a RSF models with more precise prediction accuracy (4). RSF prediction error rates of 0.5 refer to RSF models which based on chance (4).

Partial (dependence) plots

Non-linearity between selected metabolites and five-year T2D-free survival was assessed graphically, using partial (dependence) plots. Partial plots represent the effect of each selected variable on predicted event-free survival after accounting for the average effects of the remaining variables of the respective RSF model (2, 6). The computation of partial plots based on an existing RSF model. To generate the partial plot of a respective variable, several new dataset are created. In each new dataset the values of the respective variable are replaced by one constant value (partial value) of the respective variable. The partial value of each new dataset represents always another value of the respective original variable. Now the OOB samples traverse the RSF using first the original data and then the converted data. The difference in predicted event-free survival is recorded and used to draw the partial plots of the respective variable which is thus also adjusted for all other included variables. The computation of partial plots is conducted by an automatically function of the R-package randomForestSRC (7).

R-code of the RSF backward elimination process

# Data used for the R-code have to include all metabolites to be tested, the covariates, a

# censoring time variable, and the censoring status.

library(randomForestSRC)

covar <- c(“list of covariates”)

count <- number of metabolites

rsf.err <- rep(0,count) # list to save error rates

rm.var <- rep(0,count) # list to save removed variables

ntree <- 1000# number of bootstrap samples

nsplit <- 10# number of node splits

for(k in 1 : count){

# grow the RSF

rsf.out <- rfsrc(Surv(time, status) ~ **.**, data = data, ntree = ntree, nsplit = nsplit,

forest = T)

# save error rate of current RSF

rsf.err[k] <- rsf.out$err.rate[ntree]

# get list of variables ordered by their minimal depth values

v.max <- max.subtree(rsf.out)

d <- sort(round(v.max$order[, 1], 3))

#-------------------------------------------------------------------

# get the metabolite with worst minimal depth value, if a covariate has

# the worst minimal depth value than ignore it and consider only metabolites

i <- 1

m <- 0

while(i > 0){

outvar<-names(d[length(d)-m])

# if outvar is a covariate then look next variable

i <- length(i <- grep(outvar, covar))

m <- m + 1

}

#-------------------------------------------------------------------

# save name of removed variable in a list

rm.var[k]<-outvar

# delete metabolite with worst minDepth from dataset

data<-data[,!(names(data) %***in***% outvar)]

# remove old RSF to ensure free working memory

rm(rsf.out)

}

output<-cbind(rm.var,rsf.err)

# check the output to find the set of metabolites with the lowest RSF prediction error rate

output

**Supplementary tables**

**Table S1:** List of analysed serum metabolites in EPIC-Potsdam with labelling of metabolites selected by Random Survival Forest backward elimination or Cox proportional hazards regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metabolite** | **Biochemical name** | **RSF with backward elimination** | **Cox Proportional Hazards Regression** | |
| **Step 1 a)** | **Step 2 b)** |
| **Acylcarnitines** | | | | |
| C0 | DL-Carnitine |  |  |  |
| C10 | Decanoyl-L-carnitine |  |  |  |
| C10:2 | Decadienyl-L-carnitine |  |  |  |
| C14:1 | Tetradecenoyl-L-carnitine |  |  |  |
| C14:2 | Tetradecadienyl-L-carnitine |  |  |  |
| C16 | Hexadecanoyl-L-carnitine | + |  |  |
| C16:2 | Hexadecadienyl-L-carnitine |  |  |  |
| C18 | Octadecanoyl-L-carnitine |  |  |  |
| C18:1 | Octadecenoyl-L-carnitine |  |  |  |
| C18:2 | Octadecadienyl-L-carnitine |  |  |  |
| C2 | Acetyl-L-carnitine |  |  |  |
| C3 | Propionyl-L-carnitine |  | + |  |
| C3-DC-M /  C5-OH | Methylmalonyl-L-carnitine / Hydroxyvaleryl-L-carnitine |  |  |  |
| C5-DC /  C6-OH | Glutaryl-L-carnitine / Hydroxyhexanoyl-L-carnitine |  |  |  |
| C7-DC | Pimelyl-L-carnitine |  |  |  |
| C8:1 | Octenoyl-L-carnitine |  |  |  |
| C9 | Nonayl-L-carnitine |  |  |  |
| **Amino acids** | | | | |
| Arginine | Arginine |  |  |  |
| Glutamine | Glutamine |  |  |  |
| Glycine | Glycine | - | - | - |
| Histidine | Histidine |  |  |  |
| Methionine | Methionine |  |  |  |
| Ornithine | Ornithine |  |  |  |
| Phenylalanine | Phenylalanine |  | + | + |
| Proline | Proline |  |  |  |
| Serine | Serine |  |  |  |
| Threonine | Threonine |  |  |  |
| Tryptophan | Tryptophan |  | + |  |
| Tyrosine | Tyrosine | + | + |  |
| Valine | Valine | + | + |  |
| xLeu | Leucine/Isoleucin |  | + |  |
| **Glycerophospholipids** | | | | |
| PC aa C28:1 | PC diacyl C28:1 |  |  |  |
| PC aa C30:0 | PC diacyl C30:0 |  |  |  |
| PC aa C32:0 | PC diacyl C32:0 | + |  |  |
| PC aa C32:1 | PC diacyl C32:1 |  | + | + |
| PC aa C32:2 | PC diacyl C32:2 |  |  |  |
| PC aa C32:3 | PC diacyl C32:3 |  |  |  |
| PC aa C34:1 | PC diacyl C34:1 |  |  |  |
| PC aa C34:2 | PC diacyl C34:2 |  |  |  |
| PC aa C34:3 | PC diacyl C34:3 |  |  |  |
| PC aa C34:4 | PC diacyl C34:4 |  |  |  |
| PC aa C36:0 | PC diacyl C36:0 |  |  |  |
| PC aa C36:1 | PC diacyl C36:1 |  | + | + |
| PC aa C36:2 | PC diacyl C36:2 |  |  |  |
| PC aa C36:3 | PC diacyl C36:3 |  | + |  |
| PC aa C36:4 | PC diacyl C36:4 |  |  |  |
| PC aa C36:5 | PC diacyl C36:5 |  |  |  |
| PC aa C36:6 | PC diacyl C36:6 |  |  |  |
| PC aa C38:0 | PC diacyl C38:0 |  |  |  |
| PC aa C38:1 | PC diacyl C38:1 |  |  |  |
| PC aa C38:3 | PC diacyl C38:3 | + | + | + |
| PC aa C38:4 | PC diacyl C38:4 |  |  |  |
| PC aa C38:5 | PC diacyl C38:5 |  |  |  |
| PC aa C38:6 | PC diacyl C38:6 |  |  |  |
| PC aa C40:2 | PC diacyl C40:2 |  |  |  |
| PC aa C40:3 | PC diacyl C40:3 |  |  |  |
| PC aa C40:4 | PC diacyl C40:4 |  | + |  |
| PC aa C40:5 | PC diacyl C40:5 |  | + | + |
| PC aa C40:6 | PC diacyl C40:6 |  |  |  |
| PC aa C42:0 | PC diacyl C42:0 | - | - |  |
| PC aa C42:1 | PC diacyl C42:1 | - | - |  |
| PC aa C42:2 | PC diacyl C42:2 |  |  |  |
| PC aa C42:4 | PC diacyl C42:4 |  |  |  |
| PC aa C42:5 | PC diacyl C42:5 |  |  |  |
| PC aa C42:6 | PC diacyl C42:6 |  |  |  |
| PC ae C30:0 | PC acyl alkyl C30:0 |  |  |  |
| PC ae C30:1 | PC acyl alkyl C30:1 |  |  |  |
| PC ae C30:2 | PC acyl alkyl C30:2 |  |  |  |
| PC ae C32:1 | PC acyl alkyl C32:1 |  | - |  |
| PC ae C32:2 | PC acyl alkyl C32:2 |  | - |  |
| PC ae C34:0 | PC acyl alkyl C34:0 |  |  |  |
| PC ae C34:1 | PC acyl alkyl C34:1 |  |  |  |
| PC ae C34:2 | PC acyl alkyl C34:2 |  | - |  |
| PC ae C34:3 | PC acyl alkyl C34:3 | - | - | - |
| PC ae C36:0 | PC acyl alkyl C36:0 |  |  |  |
| PC ae C36:1 | PC acyl alkyl C36:1 |  |  |  |
| PC ae C36:2 | PC acyl alkyl C36:2 |  | - |  |
| PC ae C36:3 | PC acyl alkyl C36:3 |  | - |  |
| PC ae C36:4 | PC acyl alkyl C36:4 |  |  |  |
| PC ae C36:5 | PC acyl alkyl C36:5 |  |  |  |
| PC ae C38:0 | PC acyl alkyl C38:0 |  |  |  |
| PC ae C38:1 | PC acyl alkyl C38:1 |  |  |  |
| PC ae C38:2 | PC acyl alkyl C38:2 |  |  |  |
| PC ae C38:3 | PC acyl alkyl C38:3 |  |  |  |
| PC ae C38:4 | PC acyl alkyl C38:4 |  |  |  |
| PC ae C38:5 | PC acyl alkyl C38:5 |  |  |  |
| PC ae C38:6 | PC acyl alkyl C38:6 |  |  |  |
| PC ae C40:1 | PC acyl alkyl C40:1 |  |  |  |
| PC ae C40:2 | PC acyl alkyl C40:2 |  |  |  |
| PC ae C40:3 | PC acyl alkyl C40:3 |  |  |  |
| PC ae C40:4 | PC acyl alkyl C40:4 |  |  |  |
| PC ae C40:5 | PC acyl alkyl C40:5 |  | - |  |
| PC ae C40:6 | PC acyl alkyl C40:6 |  | - | - |
| PC ae C42:1 | PC acyl alkyl C42:1 |  |  |  |
| PC ae C42:2 | PC acyl alkyl C42:2 |  |  |  |
| PC ae C42:3 | PC acyl alkyl C42:3 |  | - |  |
| PC ae C42:4 | PC acyl alkyl C42:4 | - | - |  |
| PC ae C42:5 | PC acyl alkyl C42:5 | - | - | - |
| PC ae C44:3 | PC acyl alkyl C44:3 |  |  |  |
| PC ae C44:4 | PC acyl alkyl C44:4 | - | - | - |
| PC ae C44:5 | PC acyl alkyl C44:5 | - | - | - |
| PC ae C44:6 | PC acyl alkyl C44:6 | - | - |  |
| lyso-PC a C14:0 | lyso-PC acyl C14:0 |  |  |  |
| lyso-PC a C16:0 | lyso-PC acyl C16:0 |  |  |  |
| lyso-PC a C16:1 | lyso-PC acyl C16:1 |  |  |  |
| lyso-PC a C17:0 | lyso-PC acyl C17:0 |  | - |  |
| lyso-PC a C18:0 | lyso-PC acyl C18:0 |  |  |  |
| lyso-PC a C18:1 | lyso-PC acyl C18:1 |  |  |  |
| lyso-PC a C18:2 | lyso-PC acyl C18:2 | - | - | - |
| lyso-PC a C20:3 | lyso-PC acyl C20:3 |  |  |  |
| lyso-PC a C20:4 | lyso-PC acyl C20:4 |  |  |  |
| lyso-PC a C28:1 | lyso-PC acyl C28:1 |  |  |  |
| **Sphingolipids** | | | | |
| SM (OH) C14:1 | Hydroxysphingomyelin C14:1 |  |  |  |
| SM (OH) C16:1 | Hydroxysphingomyelin C16:1 |  |  |  |
| SM (OH) C22:1 | Hydroxysphingomyelin C22:1 |  |  |  |
| SM (OH) C22:2 | Hydroxysphingomyelin C22:2 |  | - |  |
| SM (OH) C24:1 | Hydroxysphingomyelin C24:1 |  |  |  |
| SM C16:0 | Sphingomyelin C16:0 |  |  |  |
| SM C16:1 | Sphingomyelin C16:1 |  | - | - |
| SM C18:0 | Sphingomyelin C18:0 |  |  |  |
| SM C18:1 | Sphingomyelin C18:1 |  |  |  |
| SM C20:2 | Sphingomyelin C20:2 |  |  |  |
| SM C24:0 | Sphingomyelin C24:0 |  |  |  |
| SM C24:1 | Sphingomyelin C24:1 |  |  |  |
| SM C26:0 | Sphingomyelin C26:0 |  |  |  |
| SM C26:1 | Sphingomyelin C26:1 |  |  |  |
| Hexose\* | Hexose | + | + | + |

127 analysed metabolites are listed. Metabolites with a plus symbol were associated with increased and metabolites with a minus symbol with decreased risk of incident T2D. Step a) Each metabolite was tested individually in a Cox proportional hazards regression model. Those metabolites are highlighted that had an uncorrected p-value<0.05 before adjustment for multiple testing. Step b) Metabolites were highlighted that were independently associated with incident T2D risk after a stepwise Cox proportional hazards regression procedure was applied. Data derived from Flögel et al. (8). The RSF backward algorithm and the previously applied Cox proportional hazards regression steps of Floegel et al. were adjusted for age, sex, BMI, waist circumference, alcohol intake from beverages, smoking, cycling and sports, level of education, coffee intake, red meat intake, whole-grain bread intake, and prevalent hypertension. Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl; C, carbon; PC, phosphatidylcholine; RSF, Random survival forest; T2D, type two diabetes.

**Table S2:** Summarized findings of other metabolomics investigations for metabolites identified by Random survival forest in EPIC-Potsdam regarding association with incident type 2 diabetes.

|  |  |  |  |
| --- | --- | --- | --- |
| Metabolite | Metabolic Pathway | Positive Association to T2D Reported | Inverse Association to T2D Reported |
| AC C16 | Fatty acid metabolism | - | - |
| Glycine | Glycine, serine and threonine metabolism | (9) # | (10-12) |
| Tyrosine | Tyrosine metabolism  Phenylalanine, tyrosine and tryptophan biosynthesis | (10, 13, 14) | (15, 16) |
| Valine | Valine. Leucine and isoleucine degradation  Valine. Leucine and isoleucine biosynthesis | (10, 11, 13, 14, 17, 18) | (15, 16) |
| PC aa C32:0 | Glycerophospholipid metabolism | - | - |
| PC aa C38:3 | Glycerophospholipid metabolism | (8) | - |
| PC aa C42:0 | Glycerophospholipid metabolism | - | - |
| PC aa C42:1 | Glycerophospholipid metabolism | - | - |
| PC ae C34:3 | Glycerophospholipid metabolism | - | - |
| PC ae C42:4 | Glycerophospholipid metabolism | - | - |
| PC ae C42:5 | Glycerophospholipid metabolism | - | - |
| PC ae C44:4 | Glycerophospholipid metabolism | - | - |
| PC ae C44:5 | Glycerophospholipid metabolism | - | - |
| PC ae C44:6 | Glycerophospholipid metabolism | - | - |
| lyso-PC a C18:2 | Glycerophospholipid metabolism | (19) | (12, 20) |
| Hexose \* | Sugar | (21) |  |

Numbers in brackets refer to references from studies. # Inaccurate measurement in study (9), metabolite was defined as gylcine and lysine in this study. \* Hexose presents sum of C6 – sugars, other studies have examined either glucose or fructose, therefore a direct comparison failed. No literature evidences exist for rows with a minus symbol. Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl; C, carbon; PC, phosphatidylcholine; T2D, type two diabetes.

**Table S3:** RSF-derived error rates for the prediction of incident T2D by serum metabolites and/or covariates in EPIC-Heidelberg

|  |  |
| --- | --- |
| RSF model | Prediction error rate  mean (95% CI) |
| Covariates and selected metabolites | 0.1804 (0.1800 ; 0.1808) |
| Only Covariates | 0.2604 (0.2599; 0.2609) |
| All metabolites | 0.2663 (0.2657; 0.2669) |
| Covariates and all metabolites | 0.2341 (0.2335; 0.2347) |

**Table S4:** RSF-derived error rates for the prediction of incident T2D by serum metabolites and/or covariates in KORA

|  |  |
| --- | --- |
| RSF model | Prediction error rate  mean (95% CI) |
| Covariates and selected metabolites | 0.1227 (0.1223;0.1233) |
| Only Covariates | 0.2261 (0.2253;0.2269) |
| All metabolites | 0.2587 (0.2576;0.2599) |
| Covariates and all metabolites | 0.2245 (0.2234;0.2256) |

**Table S5:** Testing of metabolites for their association with Type 2 Diabetes Risk in EPIC-Heidelberg using univariate Cox regression with subsequent correction for multiple testing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metabolite | β-Coefficent (SE)‡ | HR (95% CI) | P-Value‡ | Corrected P-Value§ StepBonf |
| Hexose (1,2,3) | 0.48 (0.12) | 1.61 (1.28, 2.02) | <.0001 | 0.0045 |
| PC aa C42:0(3) | -0.6 (0.18) | 0.55 (0.38, 0.79) | 0.0012 | 0.1400 |
| PC aa C42:1(3) | -0.58 (0.19) | 0.56 (0.39, 0.82) | 0.0025 | 0.2957 |
| lyso-PC a C18:2(1,2,3) | -0.45 (0.18) | 0.64 (0.45, 0.90) | 0.01 | 1 |
| PC aa C32:1(2) | 0.33 (0.13) | 1.38 (1.08, 1.77) | 0.0101 | 1 |
| Glutamine(2) | -0.46 (0.19) | 0.63 (0.43, 0.91) | 0.0134 | 1 |
| PC ae C30:2(1) | -0.54 (0.22) | 0.58 (0.38, 0.90) | 0.0139 | 1 |
| PC aa C38:3(1,2,3) | 0.27 (0.11) | 1.32 (1.05, 1.65) | 0.0164 | 1 |
| PC aa C40:2(1) | -0.5 (0.21) | 0.61 (0.40, 0.92) | 0.0198 | 1 |
| PC aa C36:1(1,2) | 0.38 (0.16) | 1.46 (1.06, 2.02) | 0.0208 | 1 |
| Serine(1) | -0.4 (0.18) | 0.67 (0.47, 0.94) | 0.022 | 1 |
| PC ae C40:5 | -0.43 (0.19) | 0.65 (0.44, 0.95) | 0.0263 | 1 |
| PC ae C40:1 | -0.36 (0.16) | 0.70 (0.50, 0.96) | 0.0271 | 1 |
| PC ae C42:1 | -0.37 (0.17) | 0.69 (0.5, 0.96) | 0.0273 | 1 |
| PC ae C42:4(1,3) | -0.46 (0.21) | 0.63 (0.41, 0.95) | 0.0293 | 1 |
| PC ae C42:3(1) | -0.36 (0.18) | 0.70 (0.49, 0.98) | 0.0387 | 1 |
| PC ae C40:2 | -0.34 (0.17) | 0.71 (0.51, 0.98) | 0.0394 | 1 |
| PC ae C40:6(2) | -0.36 (0.18) | 0.70 (0.49, 0.99) | 0.0438 | 1 |
| PC ae C40:4 | -0.37 (0.2) | 0.69 (0.47, 1.02) | 0.062 | 1 |
| PC ae C42:5(2,3) | -0.34 (0.19) | 0.71 (0.49, 1.02) | 0.0641 | 1 |
| PC aa C40:5(1,2) | 0.21 (0.12) | 1.24 (0.98, 1.56) | 0.0671 | 1 |
| lyso-PC a C17:0 | -0.38 (0.21) | 0.68 (0.45, 1.03) | 0.0675 | 1 |
| PC aa C42:4 | -0.29 (0.16) | 0.75 (0.55, 1.02) | 0.068 | 1 |
| PC ae C40:3(1) | -0.39 (0.22) | 0.68 (0.44, 1.03) | 0.0698 | 1 |
| PC aa C40:3 | -0.35 (0.2) | 0.70 (0.47, 1.04) | 0.081 | 1 |
| Ornithine | -0.29 (0.17) | 0.75 (0.54, 1.04) | 0.0843 | 1 |
| PC aa C34:1(1) | 0.26 (0.16) | 1.30 (0.96, 1.76) | 0.0936 | 1 |
| PC aa C32:0(3) | 0.23 (0.14) | 1.26 (0.96, 1.65) | 0.0985 | 1 |
| SM (OH) C14:1 | -0.32 (0.19) | 0.73 (0.50, 1.07) | 0.1038 | 1 |
| PC ae C38:2 | -0.37 (0.23) | 0.69 (0.44, 1.09) | 0.1089 | 1 |
| PC aa C30:0 | 0.21 (0.13) | 1.23 (0.95, 1.59) | 0.1184 | 1 |
| Proline | -0.25 (0.16) | 0.78 (0.57, 1.07) | 0.1185 | 1 |
| PC aa C40:4(1) | 0.18 (0.12) | 1.20 (0.95, 1.51) | 0.123 | 1 |
| lyso-PC a C18:1 | -0.25 (0.17) | 0.78 (0.56, 1.07) | 0.125 | 1 |
| PC aa C34:3 | 0.29 (0.19) | 1.34 (0.92, 1.95) | 0.1283 | 1 |
| SM (OH) C22:2 | -0.25 (0.17) | 0.78 (0.56, 1.08) | 0.1344 | 1 |
| PC aa C38:0 | -0.29 (0.2) | 0.75 (0.51, 1.10) | 0.1411 | 1 |
| Met | -0.26 (0.18) | 0.77 (0.55, 1.09) | 0.1463 | 1 |
| PC ae C36:1 | -0.29 (0.21) | 0.75 (0.50, 1.12) | 0.1557 | 1 |
| SM (OH) C24:1 | -0.2 (0.14) | 0.82 (0.62, 1.08) | 0.1559 | 1 |
| PC ae C44:6(3) | -0.26 (0.19) | 0.77 (0.53, 1.12) | 0.1704 | 1 |
| PC aa C32:2 | 0.17 (0.13) | 1.18 (0.92, 1.53) | 0.1991 | 1 |
| PC ae C34:2(1) | -0.3 (0.23) | 0.74 (0.47, 1.17) | 0.2003 | 1 |
| Threonine | -0.2 (0.16) | 0.82 (0.60, 1.12) | 0.2015 | 1 |
| PC ae C32:2 | -0.32 (0.25) | 0.73 (0.45, 1.19) | 0.2039 | 1 |
| Isoleucine | -0.2 (0.16) | 0.82 (0.61, 1.11) | 0.2042 | 1 |
| Arginine | -0.19 (0.15) | 0.82 (0.61, 1.11) | 0.2046 | 1 |
| PC ae C34:0 | 0.18 (0.14) | 1.20 (0.91, 1.58) | 0.207 | 1 |
| PC aa C36:0 | -0.21 (0.18) | 0.81 (0.57, 1.14) | 0.2265 | 1 |
| PC ae C38:3 | -0.23 (0.19) | 0.80 (0.54, 1.17) | 0.242 | 1 |
| SM (OH) C16:1 | -0.2 (0.18) | 0.82 (0.58, 1.17) | 0.269 | 1 |
| Glycine(3) | -0.24 (0.22) | 0.78 (0.51, 1.21) | 0.2712 | 1 |
| PC ae C36:2 | -0.24 (0.23) | 0.79 (0.51, 1.23) | 0.29 | 1 |
| PC aa C34:4 | 0.14 (0.14) | 1.15 (0.87, 1.53) | 0.3193 | 1 |
| AC C18:0 | -0.15 (0.15) | 0.86 (0.64, 1.17) | 0.3399 | 1 |
| PC aa C42:2 | -0.2 (0.21) | 0.82 (0.54, 1.24) | 0.3414 | 1 |
| SM C24:1 | -0.16 (0.17) | 0.85 (0.61, 1.19) | 0.3497 | 1 |
| AC C7-DC | -0.14 (0.16) | 0.87 (0.64, 1.17) | 0.3533 | 1 |
| PC ae C36:3(1) | -0.22 (0.24) | 0.80 (0.50, 1.28) | 0.3575 | 1 |
| AC C16:2 | -0.15 (0.17) | 0.86 (0.62, 1.20) | 0.3648 | 1 |
| AC C2:0 | 0.15 (0.16) | 1.16 (0.84, 1.60) | 0.3775 | 1 |
| SM C26:1 | -0.14 (0.16) | 0.87 (0.63, 1.19) | 0.3805 | 1 |
| PC ae C44:5(2,3) | -0.16 (0.19) | 0.85 (0.58, 1.24) | 0.3929 | 1 |
| Leucine | -0.13 (0.15) | 0.88 (0.65, 1.18) | 0.4012 | 1 |
| AC C18:1 | -0.12 (0.14) | 0.89 (0.67, 1.17) | 0.4036 | 1 |
| PC ae C38:5 | -0.18 (0.21) | 0.84 (0.55, 1.28) | 0.4094 | 1 |
| AC C9:0 | -0.14 (0.17) | 0.87 (0.62, 1.21) | 0.4114 | 1 |
| Tyrosine(3) | 0.11 (0.13) | 1.11 (0.86, 1.44) | 0.4139 | 1 |
| SM C18:0 | 0.14 (0.17) | 1.15 (0.82, 1.60) | 0.4204 | 1 |
| PC ae C38:0 | -0.12 (0.15) | 0.89 (0.66, 1.19) | 0.4235 | 1 |
| lyso-PC a C18:0 | -0.12 (0.15) | 0.89 (0.67, 1.18) | 0.424 | 1 |
| lyso-PC a C20:4 | -0.12 (0.16) | 0.88 (0.65, 1.20) | 0.4354 | 1 |
| AC C16:0(3) | -0.13 (0.17) | 0.88 (0.63, 1.23) | 0.4465 | 1 |
| PC ae C38:4 | -0.16 (0.21) | 0.85 (0.57, 1.29) | 0.4511 | 1 |
| His | -0.14 (0.2) | 0.87 (0.59, 1.27) | 0.4658 | 1 |
| PC aa C38:4 | 0.1 (0.14) | 1.10 (0.85, 1.44) | 0.4704 | 1 |
| SM C16:1(2) | -0.15 (0.2) | 0.86 (0.58, 1.29) | 0.4716 | 1 |
| PC ae C34:3(2,3) | -0.14 (0.2) | 0.87 (0.59, 1.28) | 0.4759 | 1 |
| SM C24:0 | 0.11 (0.16) | 1.12 (0.82, 1.54) | 0.4762 | 1 |
| C10:2 | -0.11 (0.16) | 0.90 (0.65, 1.22) | 0.4892 | 1 |
| C14:1 | -0.11 (0.16) | 0.90 (0.65, 1.23) | 0.5014 | 1 |
| PC ae C44:4(2,3) | -0.12 (0.18) | 0.89 (0.62, 1.26) | 0.5035 | 1 |
| SM C16:0 | -0.11 (0.17) | 0.89 (0.64, 1.25) | 0.5119 | 1 |
| C18:2 | -0.11 (0.17) | 0.90 (0.65, 1.24) | 0.5129 | 1 |
| PC ae C38:6 | -0.13 (0.2) | 0.88 (0.59, 1.30) | 0.5133 | 1 |
| PC ae C30:0 | -0.12 (0.19) | 0.89 (0.61, 1.28) | 0.5135 | 1 |
| PC aa C42:5 | -0.09 (0.14) | 0.92 (0.69, 1.21) | 0.5303 | 1 |
| PC aa C28:1 | -0.1 (0.16) | 0.91 (0.67, 1.23) | 0.5304 | 1 |
| AC C3:0 | -0.1 (0.17) | 0.90 (0.65, 1.25) | 0.5443 | 1 |
| lyso-PC a C16:1 | 0.06 (0.11) | 1.06 (0.85, 1.33) | 0.5839 | 1 |
| PC aa C38:5 | 0.08 (0.14) | 1.08 (0.82, 1.43) | 0.5849 | 1 |
| AC C10:0 | -0.09 (0.16) | 0.92 (0.67, 1.27) | 0.605 | 1 |
| C5-DC (C6-OH) | -0.07 (0.15) | 0.93 (0.70, 1.24) | 0.6122 | 1 |
| SM (OH) C22:1 | -0.08 (0.16) | 0.92 (0.68, 1.26) | 0.6193 | 1 |
| lyso-PC a C20:3 | -0.09 (0.18) | 0.92 (0.65, 1.29) | 0.6271 | 1 |
| lyso-PC a C28:1 | -0.08 (0.17) | 0.93 (0.67, 1.29) | 0.647 | 1 |
| PC aa C34:2 | 0.06 (0.13) | 1.06 (0.82, 1.38) | 0.6504 | 1 |
| Valine(3) | -0.06 (0.13) | 0.94 (0.73, 1.22) | 0.6516 | 1 |
| lyso-PC a C16:0 | -0.06 (0.14) | 0.95 (0.72, 1.24) | 0.6881 | 1 |
| PC aa C36:5 | 0.06 (0.15) | 1.06 (0.79, 1.44) | 0.692 | 1 |
| AC C0 | -0.08 (0.2) | 0.93 (0.63, 1.36) | 0.7019 | 1 |
| PC aa C36:2 | 0.06 (0.17) | 1.06 (0.77, 1.47) | 0.7209 | 1 |
| Tryptophan | 0.05 (0.13) | 1.05 (0.81, 1.36) | 0.7296 | 1 |
| PC ae C42:2(1) | -0.07 (0.21) | 0.93 (0.62, 1.41) | 0.7358 | 1 |
| C14:2(1) | -0.05 (0.15) | 0.95 (0.71, 1.28) | 0.749 | 1 |
| PC aa C36:4 | 0.04 (0.15) | 1.04 (0.77, 1.41) | 0.7874 | 1 |
| PC ae C36:0 | -0.05 (0.2) | 0.95 (0.64, 1.40) | 0.7963 | 1 |
| PC aa C36:3 | 0.04 (0.16) | 1.04 (0.76, 1.42) | 0.819 | 1 |
| Phenylalanine(2) | -0.03 (0.15) | 0.97 (0.73, 1.29) | 0.823 | 1 |
| PC aa C32:3 | 0.04 (0.2) | 1.04 (0.71, 1.54) | 0.8235 | 1 |
| PC aa C40:6 | 0.02 (0.13) | 1.02 (0.80, 1.31) | 0.866 | 1 |
| PC ae C44:3 | 0.03 (0.18) | 1.03 (0.72, 1.47) | 0.8731 | 1 |
| PC ae C36:4 | -0.02 (0.19) | 0.98 (0.67, 1.43) | 0.9158 | 1 |
| PC aa C36:6 | 0.02 (0.15) | 1.02 (0.76, 1.36) | 0.9165 | 1 |
| PC ae C32:1 | 0.02 (0.2) | 1.02 (0.68, 1.51) | 0.9338 | 1 |
| PC aa C38:6 | -0.01 (0.14) | 0.99 (0.76, 1.29) | 0.9476 | 1 |
| lyso-PC a C14:0 | 0.01 (0.15) | 1.01 (0.75, 1.36) | 0.9545 | 1 |
| PC ae C34:1 | 0.01 (0.22) | 1.01 (0.66, 1.55) | 0.9583 | 1 |
| PC aa C42:6 | 0 (0.14) | 1 (0.76, 1.32) | 0.9906 | 1 |
| SM C18:1 | 0 (0.18) | 1 (0.71, 1.42) | 0.994 | 1 |
| PC ae C36:5 | 0 (0.18) | 1 (0.70, 1.43) | 0.9984 | 1 |

The stepwise Cox regression selection was adjusted for age, sex, alcohol intake from beverages (non-consumers; women: >0-6 g/d, 6-12 g/d, >12 g/d; men: >0-12 g/d, 12-24 g/d, >24 g/d), smoking (never, former, current ≤20 cigarettes/d, current >20 cigarettes/d), physical activity (cycling and sports in h/week), education (no degree/vocational training; trade/technical school; university degree), coffee intake (cups/d), red meat intake (g/d), whole grain bread intake (g/d), prevalent hypertension (yes/no), BMI (kg/m²), and waist circumference (cm). §*P*-values were corrected to account for multiple testing (n=127) using the Bonferroni-Holm procedure. Number in brackets marked metabolites that were selected: (1) in EPIC-Heidelberg with the RSF backward algorithm, (2) in EPIC-Potsdam with stepwise Cox regression, (3) in EPIC-Potsdam with RSF backward algorithm.

Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl; DC, decarboxyl; EPIC, European Prospective Investigation into Cancer and Nutrition; PC, phosphatidylcholine; SE, standard error; SM, sphingomyelin.

**Table S6:** Testing of metabolites for their association with Type 2 Diabetes Risk in KORA-study using univariate Cox regression with subsequent correction for multiple testing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metabolite | β-Coefficent (SE)‡ | HR (95% CI) | P-Value‡ | Corrected P-Value§ StepBonf |
| Hexose(1,2,3) | 0.52 (0.12) | 1.69 (1.34, 2.12) | <.0001 | 0.0009 |
| PC ae C34:1(1) | -1.03 (0.27) | 0.36 (0.21, 0.60) | 0.0001 | 0.012 |
| PC ae C42:1 | -0.73 (0.27) | 0.48 (0.28, 0.82) | 0.0073 | 0.8305 |
| PC ae C32:1(1) | -1.01 (0.38) | 0.37 (0.18, 0.76) | 0.0073 | 0.8305 |
| PC aa C34:1 | -0.72 (0.27) | 0.49 (0.29, 0.84) | 0.0091 | 1 |
| C7 dC | -0.69 (0.27) | 0.50 (0.30, 0.85) | 0.01 | 1 |
| PC ae C32:2 | -0.95 (0.38) | 0.39 (0.18, 0.81) | 0.0112 | 1 |
| PC ae C36:1 | -0.79 (0.31) | 0.46 (0.25, 0.84) | 0.0113 | 1 |
| Isoleucine(1) | 0.40 (0.16) | 1.49 (1.09, 2.04) | 0.0114 | 1 |
| PC aa C32:1(2) | -0.74 (0.31) | 0.48 (0.26, 0.88) | 0.0168 | 1 |
| PC aa C28:1 | -0.65 (0.27) | 0.52 (0.31, 0.89) | 0.017 | 1 |
| Leucine | 0.42 (0.18) | 1.52 (1.07, 2.18) | 0.0208 | 1 |
| PC ae C30:0 | -0.77 (0.33) | 0.46 (0.24, 0.89) | 0.0214 | 1 |
| SM (OH) C14:1 | -0.53 (0.25) | 0.59 (0.36, 0.96) | 0.0346 | 1 |
| PC aa C34:3 | -0.64 (0.31) | 0.53 (0.29, 0.96) | 0.0372 | 1 |
| Tryptophan(1) | 0.41 (0.20) | 1.51 (1.02, 2.23) | 0.0387 | 1 |
| SM C16:1(2) | -0.49 (0.24) | 0.61 (0.39, 0.98) | 0.0389 | 1 |
| PC ae C34:3(2,3) | -0.68 (0.33) | 0.51 (0.27, 0.97) | 0.0392 | 1 |
| SM (OH) C22:2 | -0.51 (0.25) | 0.6 (0.37, 0.98) | 0.0397 | 1 |
| AC C16:0(3) | -0.38 (0.19) | 0.68 (0.47, 0.98) | 0.0407 | 1 |
| PC aa C40:5(2) | -0.48 (0.24) | 0.62 (0.38, 0.99) | 0.0467 | 1 |
| PC ae C34:2 | -0.57 (0.29) | 0.56 (0.32, 1.00) | 0.0502 | 1 |
| PC ae C34:0 | -0.66 (0.34) | 0.52 (0.27, 1.01) | 0.0522 | 1 |
| AC C9:0 | -0.55 (0.28) | 0.58 (0.33, 1.01) | 0.0545 | 1 |
| PC ae C44:3 | -0.39 (0.20) | 0.68 (0.46, 1.01) | 0.058 | 1 |
| PC ae C42:2 | -0.56 (0.30) | 0.57 (0.32, 1.02) | 0.058 | 1 |
| PC ae C40:3 | -0.62 (0.33) | 0.54 (0.28, 1.03) | 0.0607 | 1 |
| PC aa C38:5 | -0.37 (0.20) | 0.69 (0.46, 1.02) | 0.0655 | 1 |
| sm oh C24:1 | -0.44 (0.24) | 0.65 (0.40, 1.04) | 0.072 | 1 |
| PC aa C32:3 | -0.54 (0.30) | 0.58 (0.32, 1.05) | 0.0743 | 1 |
| PC ae C36:0 | -0.57 (0.32) | 0.56 (0.3, 1.06) | 0.0757 | 1 |
| PC ae C36:2 | -0.57 (0.32) | 0.57 (0.30, 1.07) | 0.0779 | 1 |
| PC ae C38:3 | -0.6 (0.34) | 0.55 (0.28, 1.07) | 0.0784 | 1 |
| Tyrosine(1,3) | 0.34 (0.20) | 1.40 (0.95, 2.06) | 0.0845 | 1 |
| AC C14:1 | -0.39 (0.23) | 0.68 (0.43, 1.06) | 0.0889 | 1 |
| Valine(3) | 0.45 (0.27) | 1.57 (0.93, 2.65) | 0.0943 | 1 |
| PC ae C40:2 | -0.43 (0.26) | 0.65 (0.39, 1.09) | 0.1025 | 1 |
| AC C18:1 | -0.38 (0.23) | 0.69 (0.44, 1.08) | 0.1031 | 1 |
| PC aa C36:1(2) | -0.43 (0.27) | 0.65 (0.38, 1.10) | 0.1081 | 1 |
| lyso-PC a C16:1 | -0.42 (0.26) | 0.65 (0.39, 1.10) | 0.1089 | 1 |
| PC aa C32:0(3) | -0.40 (0.25) | 0.67 (0.41, 1.10) | 0.1105 | 1 |
| AC C2:0 | -0.39 (0.26) | 0.68 (0.41, 1.12) | 0.1284 | 1 |
| PC aa C30:0 | -0.41 (0.27) | 0.67 (0.39, 1.13) | 0.1331 | 1 |
| PC ae C40:1 | -0.35 (0.24) | 0.70 (0.44, 1.12) | 0.1361 | 1 |
| PC ae C36:3 | -0.51 (0.35) | 0.60 (0.30, 1.19) | 0.1422 | 1 |
| PC aa C34:2 | -0.31 (0.22) | 0.73 (0.48, 1.12) | 0.1524 | 1 |
| SM (OH) C16:1 | -0.31 (0.22) | 0.73 (0.48, 1.13) | 0.1557 | 1 |
| PC aa C36:4 | -0.30 (0.22) | 0.74 (0.48, 1.14) | 0.1698 | 1 |
| PC ae C40:5 | -0.40 (0.29) | 0.67 (0.38, 1.19) | 0.1716 | 1 |
| AC C10:0 | -0.43 (0.32) | 0.65 (0.35, 1.21) | 0.1742 | 1 |
| Glutamine | -0.27 (0.20) | 0.76 (0.51, 1.13) | 0.1746 | 1 |
| AC C3:0 | -0.31 (0.24) | 0.73 (0.46, 1.17) | 0.1897 | 1 |
| PC aa C36:3(1) | -0.40 (0.31) | 0.67 (0.36, 1.23) | 0.1952 | 1 |
| PC ae C38:4 | -0.35 (0.27) | 0.70 (0.41, 1.21) | 0.2018 | 1 |
| SM (OH) C22:1 | -0.31 (0.25) | 0.74 (0.45, 1.20) | 0.222 | 1 |
| PC ae C42:3 | -0.35 (0.30) | 0.71 (0.39, 1.27) | 0.2425 | 1 |
| PC aa C34:4 | -0.30 (0.25) | 0.74 (0.45, 1.22) | 0.2431 | 1 |
| SM C16:0 | -0.29 (0.25) | 0.75 (0.46, 1.22) | 0.2453 | 1 |
| PC aa C36:0 | -0.26 (0.23) | 0.77 (0.49, 1.21) | 0.2628 | 1 |
| PC ae C38:0 | -0.29 (0.26) | 0.75 (0.45, 1.25) | 0.2732 | 1 |
| Proline | 0.23 (0.21) | 1.25 (0.84, 1.88) | 0.2741 | 1 |
| PC aa C38:6 | -0.16 (0.14) | 0.86 (0.65, 1.13) | 0.2743 | 1 |
| PC ae C36:4 | -0.37 (0.35) | 0.69 (0.34, 1.38) | 0.2941 | 1 |
| PC aa C40:4 | -0.31 (0.30) | 0.73 (0.41, 1.32) | 0.2989 | 1 |
| AC C0 | -0.20 (0.19) | 0.82 (0.56, 1.20) | 0.3011 | 1 |
| PC ae C40:6(2) | -0.21 (0.21) | 0.81 (0.54, 1.22) | 0.308 | 1 |
| PC aa C42:4 | -0.28 (0.27) | 0.76 (0.44, 1.30) | 0.3094 | 1 |
| SM C24:1 | -0.22 (0.22) | 0.80 (0.53, 1.23) | 0.3159 | 1 |
| SM C26:1 | -0.24 (0.24) | 0.78 (0.49, 1.26) | 0.3165 | 1 |
| PC ae C38:2 | -0.29 (0.30) | 0.75 (0.41, 1.35) | 0.3362 | 1 |
| PC ae C40:4 | -0.31 (0.33) | 0.73 (0.39, 1.39) | 0.3374 | 1 |
| PC ae C38:6 | -0.20 (0.22) | 0.82 (0.54, 1.25) | 0.3541 | 1 |
| AC C18:0 | -0.26 (0.29) | 0.77 (0.43, 1.36) | 0.3673 | 1 |
| Phenylalanine(2) | 0.18 (0.21) | 1.20 (0.80, 1.81) | 0.3785 | 1 |
| PC ae C44:5(2,3) | -0.27 (0.30) | 0.77 (0.42, 1.39) | 0.3828 | 1 |
| lyso-PC a C18:1 | -0.29 (0.34) | 0.75 (0.39, 1.45) | 0.3873 | 1 |
| lyso-PC a C18:0 | 0.24 (0.3) | 0.67 (0.71, 2.27) | 0.413 | 1 |
| PC ae C36:5 | -0.21 (0.27) | 0.81 (0.48, 1.36) | 0.4236 | 1 |
| PC aa C42:6 | -0.21 (0.27) | 0.81 (0.48, 1.39) | 0.4468 | 1 |
| PC aa C42:5 | 0.17 (0.23) | 1.19 (0.76, 1.85) | 0.4485 | 1 |
| lyso-PC a C20:3(1) | -0.23 (0.33) | 0.79 (0.42, 1.50) | 0.4777 | 1 |
| PC aa C36:6 | -0.15 (0.23) | 0.86 (0.55, 1.33) | 0.4939 | 1 |
| AC C14:2 | -0.15 (0.23) | 0.86 (0.55, 1.35) | 0.522 | 1 |
| SM C20:2 | -0.20 (0.32) | 0.82 (0.44, 1.53) | 0.5357 | 1 |
| PC aa C38:3(2,3) | -0.15 (0.25) | 0.86 (0.53, 1.39) | 0.5435 | 1 |
| PC aa C40:3 | 0.15 (0.24) | 1.16 (0.72, 1.86) | 0.5445 | 1 |
| PC ae C42:5(2,3) | -0.16 (0.27) | 0.85 (0.50, 1.45) | 0.5472 | 1 |
| PC aa C42:1(3) | -0.13 (0.25) | 0.87 (0.54, 1.42) | 0.586 | 1 |
| PC aa C42:0(3) | -0.14 (0.27) | 0.87 (0.51, 1.46) | 0.5921 | 1 |
| Ornithine | -0.12 (0.23) | 0.89 (0.57, 1.38) | 0.5958 | 1 |
| PC ae C38:5 | -0.17 (0.32) | 0.85 (0.45, 1.59) | 0.605 | 1 |
| Serine | -0.14 (0.27) | 0.87 (0.51, 1.49) | 0.6194 | 1 |
| PC ae C38:1 | -0.11 (0.23) | 0.90 (0.57, 1.40) | 0.6303 | 1 |
| PC aa C42:2 | -0.13 (0.28) | 0.88 (0.51, 1.54) | 0.6557 | 1 |
| PC aa C32:2 | -0.13 (0.30) | 0.88 (0.49, 1.58) | 0.6642 | 1 |
| PC ae C44:4(2,3) | -0.11 (0.29) | 0.89 (0.51, 1.57) | 0.6965 | 1 |
| PC ae C42:4(3) | -0.10 (0.31) | 0.91 (0.50, 1.66) | 0.7497 | 1 |
| SM C18:0 | -0.06 (0.18) | 0.95 (0.66, 1.35) | 0.757 | 1 |
| Threonine | -0.07 (0.21) | 0.94 (0.61, 1.43) | 0.7571 | 1 |
| PC ae C44:6(1,3) | -0.09 (0.32) | 0.91 (0.49, 1.69) | 0.7635 | 1 |
| SM C18:1 | -0.06 (0.21) | 0.94 (0.62, 1.43) | 0.7648 | 1 |
| Histidine | -0.05 (0.17) | 0.95 (0.69, 1.32) | 0.78 | 1 |
| PC aa C36:5 | -0.07 (0.26) | 0.93 (0.56, 1.56) | 0.7871 | 1 |
| PC aa C36:2 | -0.05 (0.21) | 0.95 (0.63, 1.42) | 0.7905 | 1 |
| lyso-PC a C20:4 | 0.08 (0.31) | 1.08 (0.58, 1.99) | 0.8089 | 1 |
| Methionine | 0.04 (0.19) | 1.05 (0.71, 1.53) | 0.8174 | 1 |
| PC aa C38:0 | -0.05 (0.22) | 0.95 (0.62, 1.47) | 0.8204 | 1 |
| SM C24:0 | -0.06 (0.28) | 0.94 (0.55, 1.62) | 0.8293 | 1 |
| Glycine(2,3) | -0.05 (0.32) | 0.95 (0.51, 1.77) | 0.8724 | 1 |
| lyso-PC a C17:0 | -0.05 (0.35) | 0.95 (0.48, 1.88) | 0.887 | 1 |
| lyso-PC a C16:0 | 0.04 (0.30) | 1.04 (0.58, 1.86) | 0.8949 | 1 |
| PC aa C40:6 | -0.02 (0.16) | 0.98 (0.72, 1.34) | 0.9123 | 1 |
| AC C18:2 | 0.03 (0.23) | 1.03 (0.65, 1.62) | 0.9132 | 1 |
| AC C10:2(1) | -0.02 (0.25) | 0.98 (0.60, 1.58) | 0.9219 | 1 |
| PC aa C38:4 | -0.02 (0.24) | 0.98 (0.62, 1.55) | 0.926 | 1 |
| PC aa C40:2 | 0.01 (0.28) | 1.01 (0.58, 1.76) | 0.9657 | 1 |
| lyso-PC a C18:2(2,3) | 0.01 (0.29) | 1.01 (0.57, 1.79) | 0.9729 | 1 |
| Arginine | 0 (0.24) | 1.00 (0.62, 1.61) | 0.9934 | 1 |

The stepwise Cox regression selection was adjusted for age, sex, alcohol intake from beverages (non-consumers; women: >0-6 g/d, 6-12 g/d, >12 g/d; men: >0-12 g/d, 12-24 g/d, >24 g/d), smoking (never, former, current ≤20 cigarettes/d, current >20 cigarettes/d), physical activity (cycling and sports in h/week), education (no degree/vocational training; trade/technical school; university degree), coffee intake (cups/d), red meat intake (g/d), whole grain bread intake (g/d), prevalent hypertension (yes/no), BMI (kg/m²), and waist circumference (cm). §*P*-values were corrected to account for multiple testing (n=127) using the Bonferroni-Holm procedure. Number in brackets marked metabolites that were selected: (1) in KORA with the RSF backward algorithm, (2) in EPIC-Potsdam with stepwise Cox regression, (3) in EPIC-Potsdam with RSF backward algorithm.

Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl; DC, decarboxyl; EPIC, European Prospective Investigation into Cancer and Nutrition; PC, phosphatidylcholine; SE, standard error; SM, sphingomyelin.

**Supplementary figures**

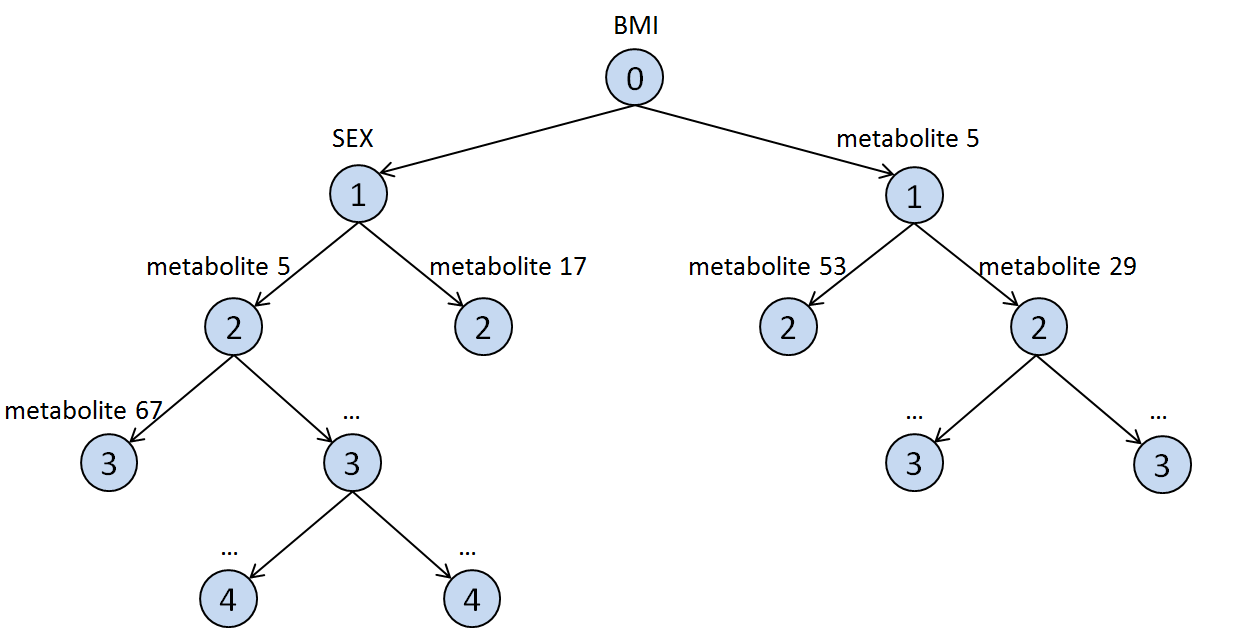


Figure S1: Illustration of minimal depth measurement. The minimal depth is indicated by an integer value inside the nodes.

**Figure S2:** Selected metabolites that are most predictive for incident type 2 diabetes ranked by the minimal depth measurement in the EPIC-Heidelberg study. Metabolites were selected using the random survival forest backward algorithm. Metabolites with lower minimal depth values are more predictive regarding incident type 2 diabetes. Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl; PC, phosphatidylcholine; AC, acylcarnitine.

**Figure S3:** Selected metabolites that are most predictive for incident type 2 diabetes ranked by the minimal depth measurement in the KORA study. Metabolites were selected using the random survival forest backward algorithm. Metabolites with lower minimal depth values are more predictive regarding incident type 2 diabetes. Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl; PC, phosphatidylcholine; AC, acylcarnitine.

**References:**

1. Segal MR. Regression Trees for Censored-Data. Biometrics. 1988;44(1):35-47.

2. Ishwaran H, Kogalur UB. Random survival forest for R. R News. 2007;7(2):25-31.

3. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-Dimensional Variable Selection for Survival Data. Journal of the American Statistical Association. 2010;105(489):205-17.

4. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Annals of Applied Statistics. 2008;2(3):841-60.

5. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA : the journal of the American Medical Association. 1982;247(18):2543-6.

6. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of statistics. 2001;29(5):1189-232.

7. Ishwaran H, Kogalur UB. randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC) <https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf>. 2015.

8. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, Joost HG, et al. Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. Diabetes. 2012.

9. Zhang J, Yan L, Chen W, Lin L, Song X, Yan X, et al. Metabonomics research of diabetic nephropathy and type 2 diabetes mellitus based on UPLC-oaTOF-MS system. Anal Chim Acta. 2009;650(1):16-22.

10. Ferrannini E, Natali A, Camastra S, Nannipieri M, Mari A, Adam KP, et al. Early metabolic markers of the development of dysglycemia and type 2 diabetes and their physiological significance. Diabetes. 2013;62(5):1730-7.

11. Fiehn O, Garvey WT, Newman JW, Lok KH, Hoppel CL, Adams SH. Plasma metabolomic profiles reflective of glucose homeostasis in non-diabetic and type 2 diabetic obese African-American women. PLoS One. 2010;5(12):e15234.

12. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, et al. Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol. 2012;8:615.

13. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. Nat Med. 2011;17(4):448-53.

14. Xu F, Tavintharan S, Sum CF, Woon K, Lim SC, Ong CN. Metabolic signature shift in type 2 diabetes mellitus revealed by mass spectrometry-based metabolomics. J Clin Endocrinol Metab. 2013;98(6):E1060-5.

15. van Doorn M, Vogels J, Tas A, van Hoogdalem EJ, Burggraaf J, Cohen A, et al. Evaluation of metabolite profiles as biomarkers for the pharmacological effects of thiazolidinediones in Type 2 diabetes mellitus patients and healthy volunteers. Br J Clin Pharmacol. 2007;63(5):562-74.

16. Zhang X, Wang Y, Hao F, Zhou X, Han X, Tang H, et al. Human serum metabonomic analysis reveals progression axes for glucose intolerance and insulin resistance statuses. J Proteome Res. 2009;8(11):5188-95.

17. Bao Y, Zhao T, Wang X, Qiu Y, Su M, Jia W, et al. Metabonomic variations in the drug-treated type 2 diabetes mellitus patients and healthy volunteers. J Proteome Res. 2009;8(4):1623-30.

18. Menni C, Fauman E, Erte I, Perry JR, Kastenmuller G, Shin SY, et al. Biomarkers for type 2 diabetes and impaired fasting glucose using a non-targeted metabolomics approach. Diabetes. 2013.

19. Zhu C, Liang QL, Hu P, Wang YM, Luo GA. Phospholipidomic identification of potential plasma biomarkers associated with type 2 diabetes mellitus and diabetic nephropathy. Talanta. 2011;85(4):1711-20.

20. Liu L, Wang M, Yang X, Bin M, Na L, Niu Y, et al. Fasting Serum Lipid and Dehydroepiandrosterone Sulfate as Important Metabolites for Detecting Isolated Postchallenge Diabetes: Serum Metabolomics via Ultra-High-Performance Liquid Chromatography/Mass Spectrometry. Clin Chem. 2013.

21. Li X, Xu Z, Lu X, Yang X, Yin P, Kong H, et al. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: Biomarker discovery for diabetes mellitus. Anal Chim Acta. 2009;633(2):257-62.