International Journal of Epidemiology, 2016, 1406–1420
doi: 10.1093/ije/dyw145
Advance Access Publication Date: 1 September 2016
Original article

Metabolomics

Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis

Stefan Dietrich[1,*], Anna Floegel[1], Martina Troll[2,3], Tilman Kühn[4], Wolfgang Rathmann[5,6], Anette Peters[3,6,7], Disorn Sookthai[4], Martin von Bergen[8], Rudolf Kaaks[4], Jerzy Adamski[6,9,10], Cornelia Prehn[9], Heiner Boeing[1], Matthias B Schulze[6,11], Thomas Illig[2,12], Tobias Pischon[1,13], Sven Knüppel[1], Rui Wang-Sattler[2,3,6] and Dagmar Drogan[1]

[1]Department of Epidemiology, German Institute of Human Nutrition, Nuthetal, Germany, [2]Research Unit of Molecular Epidemiology, [3]Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany, [4]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, [5]Institute for Biometrics and Epidemiology, Leibniz Center for Diabetes Research at Heinrich Heine University, Germany, [6]German Center for Diabetes Research (DZD), München-Neuherberg, Germany, [7]Department of Environmental Health, Harvard School of Public Health, Boston, MA, USA and, [8]Department of Molecular Systems Biology, Helmholtz Centre for Environmental Research (UFZ), Institute of Biochemistry, Faculty of Biosciences, Pharmacy and Psychology, University of Leipzig, Leipzig, Germany and Department of Chemistry and Bioscience, University of Aalborg, Aalborg East, Denmark, [9]Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, German Research Center for Environmental Health, München-Neuherberg, Germany, [10]Lehrstuhl für Experimentelle Genetik, Technische Universität München, Freising-Weihenstephan, Germany, [11]Department of Molecular Epidemiology, German Institute of Human Nutrition, Nuthetal, Germany, [12]Hannover Unified Biobank, and Institute for Human Genetics, Hannover, Germany, [13]Molecular Epidemiology Group, Max Delbruck Center for Molecular Medicine (MDC) Berlin-Buch, Berlin, Germany

*Corresponding author. Department of Epidemiology, German Institute of Human Nutrition (DIfE), Arthur-Scheunert-Allee 114-116, DE-14558 Nuthetal, Germany. E-mail: stefan.dietrich@dife.de

Accepted 20 May 2016

Abstract

Background: The application of metabolomics in prospective cohort studies is statistically challenging. Given the importance of appropriate statistical methods for selection of disease-associated metabolites in highly correlated complex data, we combined random survival forest (RSF) with an automated backward elimination procedure that addresses such issues.

Methods: Our RSF approach was illustrated with data from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study, with concentrations of 127 serum metabolites as exposure variables and time to development of type 2 diabetes mellitus (T2D) as outcome variable. Out of this data set, Cox regression with a stepwise selection method was recently published. Replication of methodical comparison (RSF and Cox regression) was conducted in two independent cohorts. Finally, the R-code for implementing the metabolite selection procedure into the RSF-syntax is provided.

Results: The application of the RSF approach in EPIC-Potsdam resulted in the identification of 16 incident T2D-associated metabolites which slightly improved prediction of T2D when used in addition to traditional T2D risk factors and also when used together with classical biomarkers. The identified metabolites partly agreed with previous findings using Cox regression, though RSF selected a higher number of highly correlated metabolites.

Conclusions: The RSF method appeared to be a promising approach for identification of disease-associated variables in complex data with time to event as outcome. The demonstrated RSF approach provides comparable findings as the generally used Cox regression, but also addresses the problem of multicollinearity and is suitable for high-dimensional data.

Key words: Cox proportional hazards regression, exploratory survival analysis, multicollinearity, random survival forest, right-censored data, metabolomics, type 2 diabetes mellitus, variable selection

Key Messages

- The implemented random survival forest backward selection algorithm enables metabolite selection and thus detection of potential novel disease biomarkers while taking into account possible confounders.
- The application of the random survival forest backward selection resulted in the identification of 16 metabolites which are associated with type 2 diabetes risk and slightly improved risk prediction compared with RSF on all metabolites.
- Non-linear relationships between identified metabolites and predicted 5-year event-free survival were displayed, thereby improving the interpretability of random survival forest results.
- Epidemiologists are encouraged to consider the provided random survival forest backward selection as a sensible complement to conventional regression-based selection methods for variable selection when analysing highly correlated complex survival data.

Introduction

Metabolite profiling offers the opportunity to discover new disease biomarkers, thereby potentially improving our understanding of disease aetiology.[1–4] However, the exploratory analysis of large metabolomic data sets containing hundreds of variables with regression approaches has unique statistical challenges including correction for multiple testing and handling of multicollinearity. These challenges have only partially been solved so far and can be considered as limitations inherent in current statistical methods. In this context, multivariate classification methods may overcome such limitations. Among them, random survival forest (RSF) could be a powerful method,[5] especially if an automated variable selection procedure could be linked with the possibility to retain a fixed set of potential confounding factors in the model.

RSF is specifically suitable for exploratory analysis of right-censored highly correlated complex survival data of prospective cohorts where the outcome is a time-dependent variable.[5] RSF uses a collection of decision trees for prediction and to rank variables by their importance for time to event, which was recently, successfully applied to identify risk factors of different diseases.[6–8] Consequently, RSF seems also suitable to reduce the data dimension of highly correlated metabolomic data in prospective cohorts by selecting the most important metabolites that are linked with event time of interest.

The importance of exploratory analysis of complex data sets in epidemiological studies will increase in the future and thus appropriate methods must be used. Hence, we illustrate the applicability of the RSF approach for exploratory identification of metabolites associated with disease risk. We applied RSF to a subcohort of the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study, using concentrations of 127 serum metabolites as exposure variables and time to development of type 2 diabetes mellitus (T2D) as endpoint.[3] An RSF backward elimination procedure was implemented to restrict the number of metabolites to informative metabolites only, while retaining forcing a set of potential confounding factors into the model. To investigate the impact of metabolite selection and potential confounders on prediction, prediction error rates were compared between several RSF models computed based on different data (identified or all metabolites, traditional T2D risk factors and classical biomarkers). In addition, in two independent cohorts, methodical replications were conducted. Finally, we discuss the performance of RSF in general and in comparison with Cox proportional hazards regression (CR), which has been recently applied to identify metabolites related to T2D in the same data.[3]

Materials and Methods

Study population

The EPIC-Potsdam study is part of the ongoing large-scale European-wide prospective cohort study, the European Prospective Investigation into Cancer and Nutrition (EPIC). From 1994 to 1998, 16644 women and 10904 men, aged mainly 35 to 65 years, were recruited from the general population in Potsdam and surrounding areas.[9] At baseline, participants underwent examinations including anthropometric and blood pressure measurements, filled in self-administered questionnaires on diet and lifestyle and answered personal computer-assisted interviews. Blood samples (30 ml) were collected at baseline and immediately fractionated, aliquoted into straws and stored at −196 °C until measurement of serum metabolites.[3] Baseline blood samples were also used for measurement of the classical biomarkers HbA1c, triglycerides, HDL-cholesterol, adiponectin and high sensitive CRP as described before.[10,11] Information about cases of incident T2D and diabetes-specific medication was recorded every 2 to 3 years by self-administered questionnaires. All self-reports had been verified by the treating physician (ICD-10: E11). In total, 849 cases of incident T2D have been recorded between baseline examination and August 2005. Using a nested case-cohort design, we randomly selected a subcohort of 2500 individuals from the EPIC-Potsdam cohort, which served as the reference group. The case-cohort design is an efficient and well-established subsampling mechanism for investigating biomarker-disease associations in prospective studies.[12] After excluding participants with self-reported prevalent T2D or antidiabetic medication at baseline, missing blood samples or missing data on follow-up, metabolites and covariates, the analytical study population consists of 800 cases of incident T2D and 2197 non-cases. For computation of RSF models including classical biomarkers, the study population consists of 690 cases of incident T2D and 2067 non-cases due to the exclusion of participants with missing classical biomarker measurements.

Replication cohorts

The implemented RSF approach and the previously used stepwise CR approach[3] were also applied in two independent German cohort studies: the Cooperative Health Research in the Region of Augsburg (KORA) study and in the EPIC-Heidelberg study.

KORA is a population-based cohort study conducted in Southern Germany.[13] The baseline survey 4 (KORA S4) consists of 4261 individuals (aged 25-74 years) examined between 1999 and 2001. During the years of 2006 to 2008, 3080 participants took part in the follow-up survey 4 (KORA F4). Clinical data for each participant were retrieved from medical records. Based on physician-validated and self-reported diagnosis, fasting glucose and 2-h post-glucose load and information on medications, incident T2D cases were identified.[13,14] After exclusion of participants with prevalent T2D at baseline S4 and incident T2D cases with earliest diagnosis in the follow-up survey 4 (2006-08), missing data on follow-up, metabolite profiles and clinical parameters, the current analysed KORA study population consists of 21 incident T2D cases and 779 non-cases. The sampling procedure and metabolite measurement of the KORA S4 have been described in detail elsewhere.[14]

The EPIC-Heidelberg study is part of the European-wide EPIC-study[15] with 25540 participants aged 35-65 years who were recruited between 1994 and 1998.[16] Recruitment and follow-up procedures and verification of prevalent and incident T2D cases in EPIC-Heidelberg were the same as in EPIC-Potsdam.[9,17] For measurements of serum metabolites, a random subcohort including only participants free of diabetes at baseline was established in 2006.[18] After excluding participants with antidiabetic medication at baseline, missing blood samples and missing data on follow-up, metabolites and covariates, the analytical study population consists of 45 cases of incident T2D and 716 non-cases.

EPIC-Potsdam study procedures were approved by the Ethics Committee of the Medical Association of the State of Brandenburg (Germany) and all participants provided written informed consent. The EPIC-Heidelberg study was approved by the Ethics Committee of the Medical Faculty of the University of Heidelberg (Heidelberg, Germany). The study participants provided written consent for the use of their blood samples and data. All KORA participants gave written informed consent, and the studies were

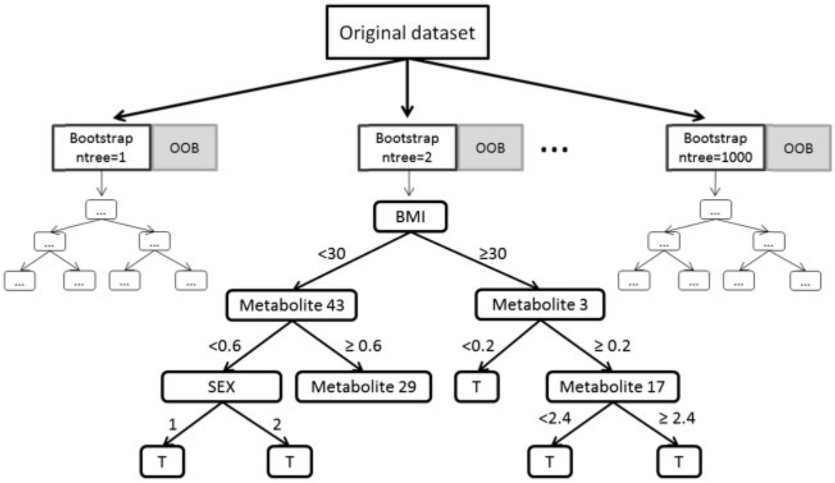

Figure 1. Schematic exemplary illustration of the computation of a RSF using 1000 bootstrap samples. The splitting process was illustrated in one possible decision tree. Numbers on edges represents possible cut points used for splitting the respective parent node into two daughter nodes. Based on the split points (e.g. 30 for BMI) observations are assigned to the left or right daughter node. To determine a split point random node splitting and the log rank statistic is used. Abbreviation: T, terminal node.

approved by the Ethics Committee of the Bavarian Medical Association.

Measurement of metabolites

In EPIC-Potsdam, metabolite concentrations of baseline blood serum samples were determined using AbsoluteIDQ^{TM} p150 Kits (Biocrates Life Sciences AG, Innsbruck, Austria) which is based on the flow injection analysis tandem mass spectrometry (FIA-MS/MS) technique as described in detail by Römisch-Margl *et al.*[19] Of 163 quantified metabolites in EPIC-Potsdam, 36 with an intraclass correlation coefficient of < 0.40 (indicating poor reliability) were excluded, leaving 127 quantified metabolites for statistical analyses:[20] one hexose (sum of six-carbon monosaccharides without distinction of isomers), 14 amino acids, 14 spingomyelins, 17 acylcarnitines (Cx:y; x = number of carbon atoms, y = number of double bonds), 37 acyl-alkyl-, 34 diacyl- and 10 lyso-phosphatidylcholines (PC).

The serum blood samples from participants in the baseline KORA S4 study were quantified with the AbsoluteIDQ^{TM}p180 Kit (Biocrates Life Sciences AG, Innsbruck, Austria),[14] of which 118 metabolites were used for the present study. In the EPIC-Heidelberg study, baseline blood serum samples were quantified using the MetaDisIDQ^{TM} Kit (Biocrates Life Sciences AG, Innsbruck, Austria).[18] Of quantified metabolites, 122 were used in the present study. In the replication analyses, only metabolites that were also available in EPIC-Potsdam study were included.

Random survival forest

An RSF is computed by an ensemble of binary decision trees which can be used for selecting the most important variables that are linked with time to event. As previously described in detail,[5] bootstrapping and random node splitting are applied to grow an ensemble of independent decision trees that form the RSF (Figure 1). Once an RSF model is computed, it can be assessed how informative a variable is regarding time until event, using the so-called minimal depth measurement (Supplementary Figure S1, available as Supplementary data at *IJE* online). For this purpose, for each variable the distance (counting the nodes) from the root node to the node where a variable splits first is determined in each decision tree. By averaging over all trees, a reliable measure of importance of a variable regarding time to event can be obtained.[21] The lower the minimal depth values, the more predictive is a variable for the outcome of interest. To determine the prediction accuracy of an RSF model, an RSF prediction error rate is computed based on Harrell's concordance index (C-index). The RSF prediction error rate is conformed to 1 minus C-index, which implies that lower values reflect an RSF model with better predictability. A detailed description of the RSF method is provided in the supplement, available as Supplementary data at *IJE* online. AN RSF can be computed automatically using the R-package randomForestSRC.[22]

Statistical analysis

To obtain a reduced set of informative metabolites associated with incident T2D taking into account covariates, we systematically removed noise metabolites by implementing the following stepwise RSF backward algorithm: (i) compute an RSF using covariates and all metabolites to be tested; (ii) rank the metabolites by minimal depth and remove the metabolite with the worst minimal depth from the data; (iii) compute a new RSF with the remaining data; (iv) repeat steps (ii) and (iii) until only one metabolite vremain;. and (v) choose the set of metabolites with the smallest prediction error rate. A similar algorithm was suggested by Díaz-Uriarte *et al.* and Jiang *et al.*[23,24] in the context of the method 'random forest' but without retaining possible covariates throughout the selection process. The code for implementing the selection algorithm in the RSF-syntax is provided in the supplement, available as Supplementary data at *IJE* online.

The RSF backward algorithm was applied on the data consisting of all metabolites and covariates, to compute a final RSF with the smallest prediction error rate. With RSF there is no need for standardization of data, and thus the crude data were used. For the purpose of comparability, the same set of covariates was used as by Floegel *et al.*[3] including age, sex, body mass index (BMI) (kg/m^2), waist circumference (cm), alcohol intake from beverages (non-consumer, women $>$ 0-6, 6-12 and $>$ 12 g/day; men $>$ 0-12, 12-24 and $>$ 24 g/day), smoking (never smoker, former, current $\leq$ 20 cigarettes/day, current $>$ 20 cigarettes/day), cycling and sports (h/week), level of education (no degree/vocational training, trade/technical school, university degree), coffee intake (cups/day), red meat intake (g/day), whole-grain bread intake (g/day) and prevalent hypertension. The latter was defined as either systolic blood pressure (BP) of 140 mmHg or higher, diastolic BP of 90mm Hg or higher, self-reported hypertension diagnosis or use of antihypertensive medication.

The same covariate classification was used in EPIC-Potsdam and EPIC-Heidelberg. In the KORA study, meat and whole-grain bread was not recorded as g/day but as categories (almost every day, several times a week, about once a week, several times a month, once a month or less, never). Physical activity was classified in the KORA study as follows: regularly more than 2 h/week, regularly around 1h/week, irregularly around 1h/week, almost to none. In EPIC-Potsdam and EPIC-Heidelberg, the diagnosis date was available as day of diagnosis and in the KORA study as year of diagnosis. In addition to the metabolite selection procedure of the RSF backward algorithm, a CR procedure for metabolite selection was applied in the two replication cohorts as described previously by Floegel *et al.*[3]

Subsequently for evaluation, different RSF models were computed containing: (i) identified metabolites derived by applying the backward elimination procedure plus traditional T2D risk factors (i.e. covariates); (ii) covariates only; (iii) all metabolites only; and (iv) all metabolites plus covariates. In EPIC-Potsdam, two additional RSF-models were computed using data of (v) classical biomarkers plus covariates; and (vi) identified metabolites plus classical biomarkers and covariates. For each RSF model, 100 repetitions were computed and used to calculate means and 95% confidence intervals (CI) of prediction error rates of the respective RSF models. Furthermore, a final RSF model containing covariates plus the subset of identified metabolites was computed to derive minimal depth values and, in EPIC-Potsdam, partial (dependence) plots of identified metabolites. Partial plots represent the effect of each metabolite on predicted 5-year T2D-free survival after accounting for the average effects of the other variables.[25,26]

Additional partial plots of classical biomarkers were computed in EPIC-Potsdam from an RSF model including data of classical biomarkers and covariates only. Moreover, to investigate correlation structures in EPIC-Potsdam, Spearman correlation coefficients between each possible pair of identified metabolites were calculated with adjustment for covariates using data of non-cases of T2D. Metabolites identified by RSF and previously by CR3 were highlighted in a metabolite network based on a Gaussian graphic model. A Gaussian graphic model represents undirected probabilistic graphs useful to analyse and visualize the dependency structure of highly correlated variables.[27] Especially for acyl-alkyl-PC, we compared the correlation structure of acyl-alkyl-PC identified by RSF and previously by CR.[3] The analyses were conducted with the statistic software R (version 3.0.0), the R-package randomForestSRC (Version 1.2) and SAS version 9.4. The default values for computation of RSFs were used. Each RSF was computed using 1000 bootstrap samples and the log-rank splitting rule with 10 splits per variable. The code is available upon request.

Results

In EPIC-Potsdam, the mean age [standard deviation SD)] was 54.7 (7.3) in future T2D cases and 49.3 (8.9) in participants of the reference group (Table 1). With 41.8%, the proportion of women among cases was substantially lower than for non-cases (57.8%). In EPIC-Heidelberg, the participants were of the same ages as in EPIC-Potsdam, whereas in the KORA study the participants were approximately 10 years older.

Applying the RSF backward algorithm on the data of 127 metabolites adjusted for the covariates resulted in a reduced set of 16 metabolites in the EPIC-Potsdam study (Figure 2). This set of metabolites had the smallest prediction error rate during selection process, suggesting high relevance for incident T2D (Table 2). Among them, hexose

Table 1. Baseline characteristics of the analysed study populations[a]

Baseline characteristics	EPIC-Potsdam		EPIC-Heidelberg		KORA	
	Non-cases ($n = 2197$)	Incident type 2 diabetes cases ($n = 800$)	Non-cases ($n = 716$)	Incident type 2 diabetes cases ($n = 45$)	Non-cases ($n = 779$)	Incident type 2 diabetes cases ($n = 21$)
Age (years)[b]	49.3 (8.9)	54.7 (7.3)	50.3 (8.0)	52.4 (7.7)	62.9 (5.4)	63.8 (5.0)
Women (%)	57.8	42.2	58.1	44.4	51.22	33.33
BMI (kg/m^2)	25.9 (0.09)	30.1 (0.15)	25.2 (0.14)	28.7 (0.56)	27.9 (0.14)	32.2 (0.87)
Waist circumference, men (cm)[c]	93.3 (0.34)	103.6 (0.46)	94.3 (0.51)	100.9 (1.77)	98.8 (0.44)	109.4 (2.34)
Waist circumference, women (cm)[c]	80.1 (0.30)	93.4 (0.62)	78.9 (0.49)	90.1 (2.24)	88.6 (0.52)	98.3 (3.95)
Prevalent hypertension (%)	48.3	70.8	25.3	48.9	46.7	85.7
Education						
No degree/vocational training (%)	36.6	45.6	23.5	44.4	8.9	9.5
Trade/technical school (%)	23.9	25.4	43.4	44.4	70.0	66.7
University degree (%)	39.5	29.0	33.1	11.1	21.2	23.8
Smoking status						
Never (%)	47.4	36.2	43.3	46.7	11.6	19.1
Former (%)	32.3	42.3	35.2	26.7	34.8	42.9
Current (%)	20.3	21.5	21.5	26.7	53.7	38.1
Among smokers: number of cigarettes/day	12.5 (0.44)	16.0 (0.74)	16.5 (0.80)	16.6 (3.09)	16.1 (0.83)	27.8 (4.31)
Physical activity (h/week)[d]	2.9 (0.08)	2.2 (0.13)	2.4 (0.09)	2.5 (0.37)	47.4	38.1
Alcohol intake from beverages (g/day)	14.8 (0.42)	14.5 (0.71)	17.6 (0.69)	16.8 (2.78)	16.0 (0.67)	15.6 (4.07)
Coffee consumption (cups/day)	2.8 (0.05)	2.7 (0.08)	2.7 (0.10)	2.5 (0.38)	2.9 (0.08)	3.4 (0.50)
Whole-grain bread intake (g/day)[e]	46.4(1.12)	38.2 (1.91)	108.0 (2.70)	101.0 (10.80)	84.98	90.84
Red meat intake (g/day)[f]	43.0 (0.62)	48.9 (1.06)	31.1 (1.06)	28.9 (4.25)	92.68	100

[a] Presented are age- and sex-adjusted mean (standard error) for continuous variables or percentages for categorical variables.

[b] Unadjusted mean (standard deviation).

[c] Age-adjusted mean (standard error).'

[d] For EPIC-Potsdam and EPIC-Heidelberg, average of cycling and sports during summer and winter season; for the KORA study, the physical activity is provided as percentage ≥ 1 h/week.

[e] For the KORA study, whole-grain bread intake is given as percentage ≥ once per week (includes whole-grain bread, brown bread, crispbread).

[f] For the KORA study, meat intake is given as percentage ≥ once per week (without sausages or ham).

appeared to have the strongest influence on incident T2D according to minimal depth. Beside hexose, the metabolites acyl-alkyl-PC C34:3 and diacyl-PC C38:3 had the lowest minimal depth values. Furthermore, several acyl-alkyl-PC (C42:4, C42:5, C44:4, C44:5, C44:6), diacyl-PC (C32:0, C42:0, C42:1), aminoacids (valine, tyrosine, glycine), lyso-PC a C18:2 and acylcarnitine C16 were identified. The mean prediction error rate of 100 computed RSF models containing only covariates was 0.216 and thus higher than the mean prediction error rate obtained in RSF models containing all 127 metabolites (0.199) or all metabolites plus covariates (0.173, Table 2). Application of the RSF backward algorithm resulted in the final RSF model (identified metabolite plus covariates) with a mean prediction error rate of 0.165. The most predictive RSF model with a mean prediction error rate of 0.145 included the identified metabolites, classical biomarkers and covariates, whereas the RSF model that included only classical biomarker and covariates resulted in a mean prediction error rate of 0.155 (Table 2).

Direction and non-linearity between identified metabolites and predicted 5-year T2D-free survival was assessed visually in partial plots (Figure 3). The T2D-free survival decreased noticeably as values of hexose, diacyl-PC C38:3, valine, tyrosine and acylcarnitine C16 increased. Threshold values were approximately 5000 μmol/l of hexose and 50 μmol/l of diacyl-PC C38:3. Individuals with the lowest values of hexose had approximately 25% higher T2D-free survival compared with individuals with highest values. In contrast, increasing values of all identified acyl-alkyl-PC, lyso-PC a C18:2, glycine, diacyl-PC C42:0 and C42:1 were associated with an increase of T2D-free survival. Most of the partial plots indicate non-linear associations between the respective metabolites and T2D-free survival. Non-linear associations were also observed in partial plots of classical biomarkers (Figure 4).

The determination of the mutual correlations of identified metabolites resulted in correlations ranging from -0.05 to 0.85. As illustrated in Figure 4, a highly correlated metabolite cluster of five acyl-alkyl-PCs (C42:4, C42:5,

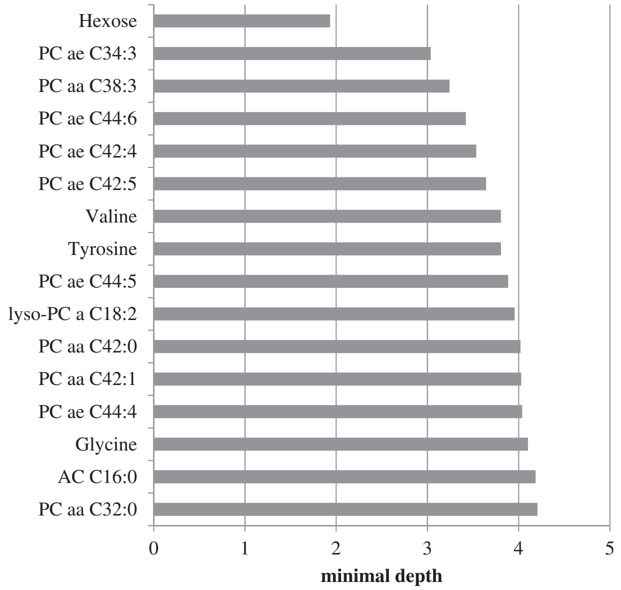

Figure 2. Identified metabolites that are most predictive for incident type 2 diabetes ranked by the minimal depth measurement. Metabolites were identified using the random survival forest backward algorithm. Metabolites with lower minimal depth values are more predictive regarding incident type 2 diabetes.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

Table 2. RSF-derived error rates for the prediction of incident T2D in different RSF models in the EPIC-Potsdam study

RSF model	Prediction error rate mean (95% CI)
Covariates and selected metabolites	0.16489 (0.16487; 0.16491)
Only covariates	0.21580 (0.21578; 0.21583)
All metabolites	0.19855 (0.19852; 0.19859)
Covariates and all metabolites	0.17314 (0.17311; 0.17317)
Covariates and biomarker	0.15515 (0.15512; 0.15517)
Covariates, selected metabolites and biomarkers	0.14533 (0.14530; 0.14535)

Covariates included 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. Biomarkers included HbA1c, triglycerides, HDL-cholesterol, adiponectin and high sensitive C-reactive protein (CRP).

C44:4, C44:5, C44:6) and two diacyl-PC (C42:0 and C42:1) was identified by the RSF backward algorithm. All seven metabolites were previously also identified by univariate CR; however, four of these metabolites lost statistical significance in the subsequently applied stepwise CR selection procedure[3] (Figure 4 and Supplementary Table S1, available as Supplementary data at *IJE* online). As demonstrated for acyl-alkyl-PCs identified by RSF and those identified previously with CR by Floegel et al.,[3] RSF tend to select metabolites with stronger correlations than CR (Figure 5).

The application of the RSF backward algorithm in the EPIC-Heidelberg study and in the KORA study resulted in the identification of 18 and 10 metabolites, respectively, which were associated with incident T2D and improved the prediction of T2D (Supplementary Figures S2 and S3, and Supplementary Tables S3 and S4, available as Supplementary data at *IJE* online). The application of a CR procedure for metabolite selection resulted in the identification of hexose only in EPIC-Heidelberg (Supplementary Table S5, available as Supplementary data at *IJE* online) and of hexose and acyl-alkyl-PC C34:1 in the KORA study (Supplementary Table S6, available as Supplementary data at *IJE* online). Most of the metabolites, which tested significant in individual tests, lost statistical significance after correction for multiple testing (Supplementary Tables S5 and S6). Nine of 18 metabolites in EPIC-Heidelberg and five of 10 metabolites in the KORA study identified with the RSF backward algorithm also tested significant with CR before correction for multiple testing. The RSF backward algorithm identified metabolites hexose, diacyl-PC C38:3, acyl-alkyl-PC C42:4 and lyso-PC a C18:2 in EPIC Heidelberg; hexose, acyl-alkyl-PC C44:6 and tyrosine in KORA were also identified in EPIC-Potsdam.

Discussion

To illustrate the applicability of RSF for exploratory data analysis in prospective cohorts, we applied an RSF backward algorithm to a well-described study population.[3] Using this approach, we were able to reduce the dimensionality of our complex data set to a subset of 16 metabolites while retaining established T2D risk factors. This was accompanied by an improvement of the prediction error rate, indicating that mainly noise metabolites were excluded. The identified metabolites also improved the prediction of T2D when classical biomarkers were available. Moreover, of identified metabolites seven metabolites form a highly correlated metabolite cluster. Partial plots, a feature of RSF, were used to display non-linear relationship between the identified metabolites and predicted 5-year T2D-free survival, thereby improving the interpretability of RSF results. In two replication cohorts with lower numbers of incident T2D cases, the RSF backward algorithm could be applied to identify incident T2D-associated metabolites. Some of these metabolites were also identified with stepwise CR, but most of them lost statistical significance after correction for multiple testing.

Many common chronic diseases of Western societies have a strong metabolic component. Therefore, the application of metabolomics in epidemiological studies is expected to expand our aetiological understanding of several

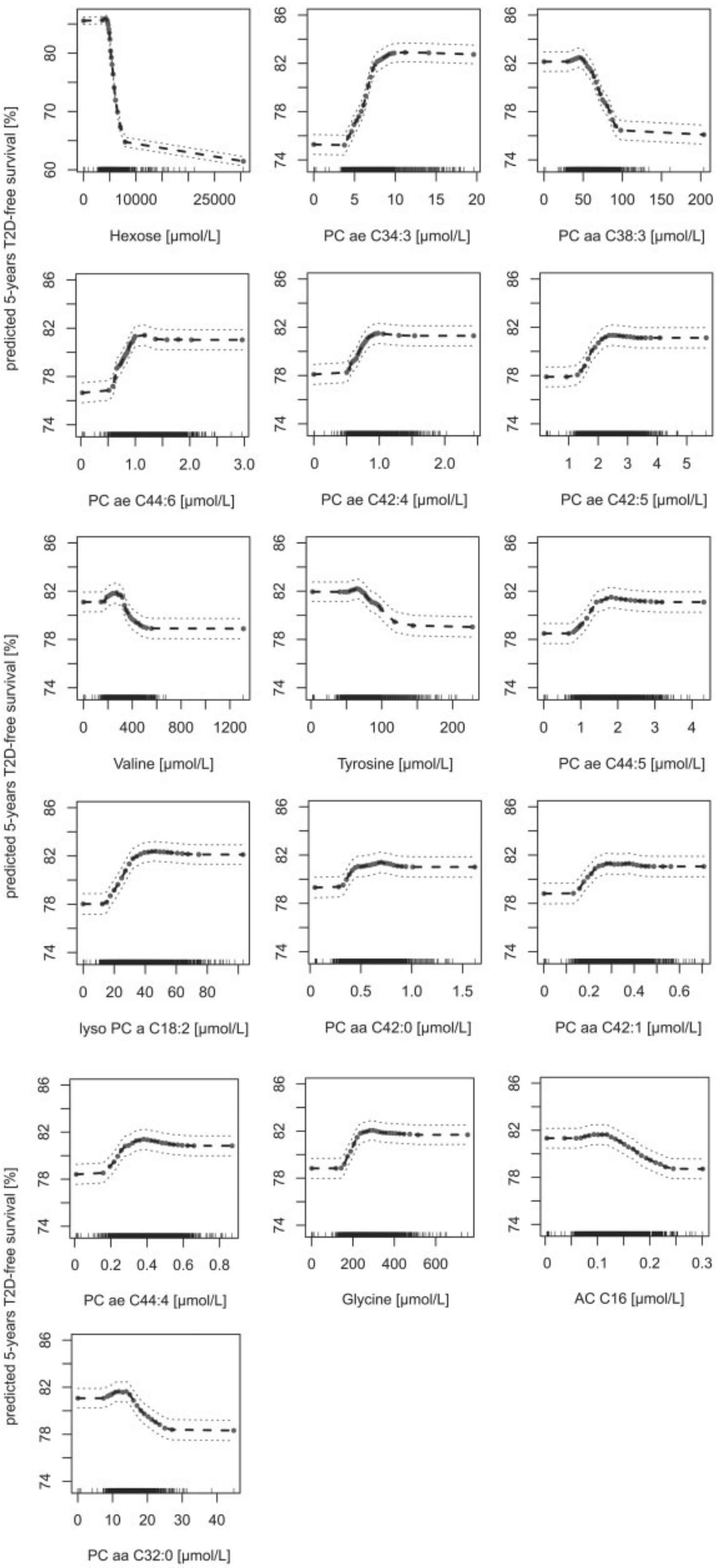

Figure 3. Partial plots of the selected metabolites including the partial values (black points) ± 2 SE (dashed grey lines). Values on the vertical axis represent predicted five-years T2D-free survival for a given variable after adjusting for all other variables (covariates and selected metabolites). Metabolite concentrations are on the horizontal axis. A lower predicted five-years T2D-free survival means a higher risk to develop type 2 diabetes within five years of follow-up time in EPIC-Potsdam. Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl, PC, phosphatidylcholine; AC, acylcarnitine.

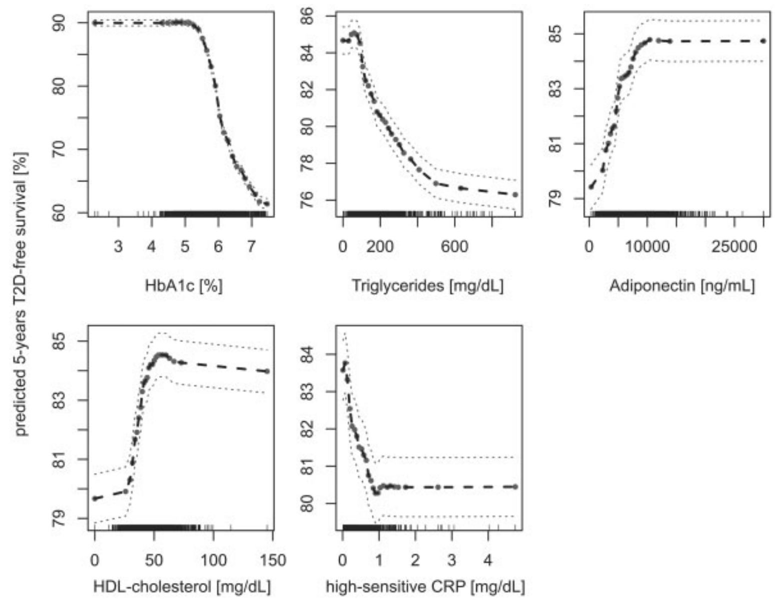

Figure 4. Partial plots of tested biomarkers in EPIC-Potsdam including the partial values (black points) ± 2 SE (dashed grey lines). The partial plots based on a RSF model that included the covariates and the five biomarkers. Values on the vertical axis represent predicted five-years T2D-free survival for a given variable after adjusting for all other variables (covariates and biomarkers). Biomarker concentrations are on the horizontal axis. A lower predicted five-years T2D-free survival means a higher risk to develop type 2 diabetes within five years of follow-up time in EPIC-Potsdam.

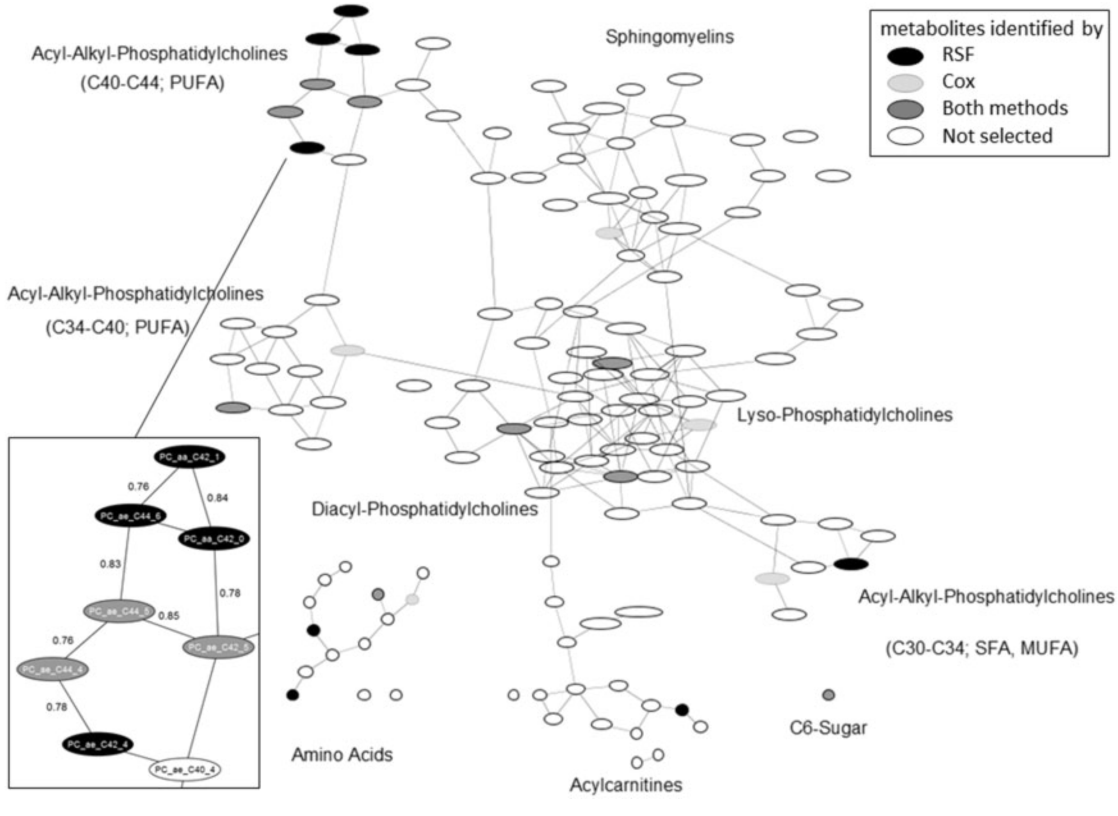

Figure 5. Gaussian graphic model of serum metabolites analysed for associations with type 2 diabetes in EPIC-Potsdam. Each node represents one metabolite and each edge between two nodes represents the partial correlation between two metabolites mutually adjusted for all other metabolites. Metabolites that were identified to be associated with incident T2D by RSF or previously by Floegel et al. with a stepwise Cox regression approach (3) are colour coded. One highly correlated metabolite pattern was resized and filled with metabolite names (nodes) and partial correlation coefficients (edges). Abbreviations: a, acyl; aa, diacyl; ae, acyl-alkyl; Cox, Cox proportional hazards regression, PC, phosphatidylcholine; RSF, Random survival forest.

diseases.[3,4,28–33] However, in general, metabolomic data consist of hundreds to thousands of metabolites and, due to the tight co-regulation of metabolic networks, metabolomics data exhibit a complex correlation structure. Therefore, identification of important predictors related with the outcome time to event remains a continuing statistical challenge.

As previously described by Floegel *et al.*,[3] a two-stage CR analysis can be applied to identify metabolites associated with disease risk. In their study,[3] the association between each metabolite and incident T2D was first assessed in a univariate test followed by a stepwise selection procedure. In compliance with the proportional hazards assumption, this approach allows calculation of hazard rate ratios and thus meaningful measures of strength and direction of risk associations. However, testing each metabolite individually–which is a frequently used approach in exploratory data analysis–increases the probability of type I error, unless appropriate methods to adjust for multiple testing are used.[34] Yet, correction for multiple testing may substantially decrease statistical power in data sets containing a large number of 'noise variables'. This seems to be the reason why most metabolites lost statistical significance after correction for multiple testing in the two replication cohorts. Compared with multiple testing, multivariable statistical analyses may provide a deeper understanding of altered metabolic pathways associated with disease development. However, a high number of predictor variables and mutual correlations increase the risk of multicollinearity in multivariable regression models. The larger the regression model, the more likely overfitting, unreliable estimation of regression coefficients, inflated standard errors or convergence problems will occur. In addition, multicollinearity increases the risk of arbitrary predictor choices in stepwise selection processes,[35,36] which altogether hampers the identification of disease-related metabolic pathways in regression models.

Compared with regression approaches, RSF has several advantages. RSF is completely data driven and thus independent of hypothesis testing. It does not test the goodness of fit of data to a hypothesis, but seeks a model that best explains the data. RSF may thus represent a suitable tool for exploratory analysis of survival data where previous knowledge is still limited. As RSF is a multivariate feature selection method, the above discussed limitations of univariate regression approaches do not apply here.

For tree growing, RSF uses random subsets of variables per node. Consequently, correlated variables will be selected independently from each other to split nodes leading to interruption of the correlation structure of variables. As a consequence, there is less competition between highly correlated variables due to the process of random node splitting, and reliable variable selection is possible even in the presence of multicollinearity.[37] This may be the reason why RSF appeared to favour correlated metabolites. Furthermore, the problem of overfitting–e.g. when multivariate regression models are performed on a high number of variables without internal validation–is largely reduced due to randomization via bootstrap sampling.[38] This feature makes RSF very appealing for explorative metabolomics research, where false-positive discoveries due to overfitting are considered to be a major problem.[39] However, the computation of an RSF represents a kind of black box, as the reduction of an RSF into one understandable decision tree is inappropriate. Instead, minimal depth measurements and partial plots can be considered for interpretation.

In light of these advantages, we applied RSF to a well-described data set including 127 serum metabolites and a number of established T2D risk factors.[3] For the sake of comparability, we used the same set of covariates as Floegel *et al.*,[3] though we acknowledge the fact that additional study characteristics such as drug use[40] or unmeasured participant characteristics could have confounded the observed metabolite-disease associations. Although RSF has been recommended for automated survival analyses,[6] methodological issues related to feature selection deserve special attention. In some recent applications, RSF has been applied on smaller sets of predictor variables.[6–8] In exploratory metabolic data analysis, however, it is necessary to identify a subset of disease-related variables among numerous additional unknown variables with no or minor association to the endpoint. Therefore, we applied a strict backward selection process under adjustment of established T2D risk factors, resulting in a reduced data set of 16 metabolites. Because the selection process resulted in a metabolite set with improved prediction error rates, it can be concluded that only noise variables were removed. With an error rate of 0.165 (equalling a C-statistic of 0.835), the predictive power of RSF coupled with backward elimination is slightly lower than what was reported by Flögel *et al.* following their two-step CR (C-statistic = 0.849).[3] The predictive accuracy of RSF and CR has been compared in several previous applications, with superiority of RSF in some[5,7,41] but not all[6,42] applications. In general, however, differences appear to be small and partly related to censoring frequency.[5]

All 16 metabolites identified by RSF also showed a nominally significant association of the same direction with the endpoint in univariate CR, though the acylcarnitine C16 lost significance in the CR following adjustment for multiple testing, as previously published.[3] Seven other metabolites–namely the acyl-alkyl-PC C42:4, C44:6, diacyl-PC C32:0, C42:0, C42:1, tyrosine and valine–were not identified in the subsequent stepwise CR,[3] since they

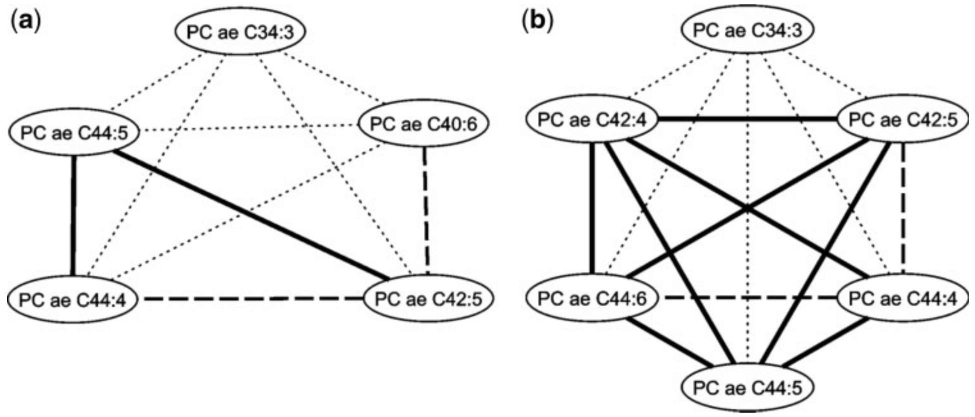

Figure 6. Correlation structure for acyl-alkyl phosphatidylcholines which were selected by (**a**) Cox proportional hazards regression analysis by Floegel *et al.* (3) and (**b**) random survival forest. Edges represents spearman correlation coefficients adjusted for age, sex, alcohol intake from beverages, smoking, cycling and sports, education, coffee intake, red meat intake, whole-grain bread intake, prevalent hypertension, BMI, and waist circumference. Dotted lines rs = 0 – 0.5, Thin dashed lines rs = 0.5 – 0.75, thick lines rs > 0.75. ae = acyl-alkyl; PC = phosphatidylcholine.

were not independent of other metabolites in the multivariate model (Supplementary Table S1). The diacyl-PC C38:3 and four acyl-alkyl-PC (C34:3, C42:5, C44:4, C44:5) were identified by both RSF and multivariate CR.[3] To our knowledge, these metabolites have not been linked to incident T2D in other cohort studies before.

However, glycerophospholipids constitute a large chemical class. The technological platforms used for metabolite profiling differ with regard to glycerophospholipid coverage, thereby limiting the comparability of previous findings. Nevertheless, alterations in diacyl-PCs and acyl-alkyl-PCs are common in the (pre)diabetic state,43–48 possibly influencing T2D risk via their impact on cell membrane integrity and cellular signal transduction.[49] Lyso-PC a C18:2–which was inversely associated with T2D in our study and in Floegel et al.[3] –is one of the few phospholipids measured in multiple human studies, with most of them also observing an inverse relation to T2D14,[48,50,51] or impaired glucose tolerance.[14,43,52] In vitro, lyso-PC a C18:2 has been shown to stimulate glucose-induced insulin release, which may partly explain the above findings.[50]

Our observation of an inverse association of glycine with T2D is in line with Floegel et al.[3] and other studies.[14,50,51] This amino acid may decrease with increasing gluconeogenesis or increasing glutathione demand as a result of oxidative stress.[53] Moreover, in the multivariate approach of *Floegel et al.* only the amino acid phenylalanine was identified to be independently associated with T2D,[3] whereas RSF did not select phenylalanine but the biochemically related aromatic amino acid tyrosine. In fact, a positive association to T2D or insulin resistance has been repeatedly observed for phenylalanine[4,54,55] as well as tyrosine.[4,50,]54–57 The (patho)biological mechanism still needs to be explored, though a competing transport mechanism of aromatic amino acids and branched-chain amino acids (BCAA) into mammalian cells has been suggested as one possible explanation.[58]

BCAA are believed to induce insulin resistance via impaired insulin signalling in skeletal muscles, and our finding of a positive association of valine to T2D is in line with several previous studies.[4,50,51,55–57,59–61] In Floegel *et al.*, an increased T2D risk was observed for the BCAA valine and isoleucine following adjustment for multiple testing, but no independent association to the endpoint was observed in the multivariate model.[3] We assume that the correlation structure and linear or non-linear associations with T2D risk contributed to the diverse selection of chemically-related metabolites by RSF and the two-stage CR. In particular, compared with the CR analysis of Floegel *et al.*,[3] RSF appears to favour metabolites that were partly highly correlated with each other, as seen in Figures 5 and 6. Some of these metabolites differ only by the number of double bonds or are likely members of the same metabolic pathway. Since RSF selects these metabolites independently of the correlation structure, bias by multicollinearity is unlikely, due to the random node-splitting process.

For use in observational epidemiology and clinical interpretation of metabolites, it is also important that RSF is able to handle the issue of confounding in metabolite-disease associations. Hence, we modified our backward selection approach so that the selection process is run on all metabolites while forcing a pre-defined set of potential covariates into the model. This allows the interpretation of the metabolites under consideration of covariates. Our finding, that the RSF prediction error rate of a model including all metabolites decreases from 0.198 to 0.173 upon additional consideration of traditional T2D risk factors, is not surprising and underlines the importance of inclusion of possible confounders when analysing metabolomic data (Table 2). Since the RSF trees are grown solely with predictor variables (i.e. metabolites) with an impact on T2D-free survival, one may speculate that noise variables do not affect prediction. However, our data show that the applied backward selection algorithm further decreases the RSF prediction error rate. Though this decrease is small, 95% CI of prediction error rates do not overlap in Table 2, indicating that mainly noise variables were excluded.

Moreover, we attempted to improve the interpretability of RSF analyses by using partial plots to determine the direction of association of each variable with the outcome of interest, including potential non-linearity. Assessment of non-linearity is also possible in CR, e.g. by cubic spline regression. Yet, partial plots are adjusted for all variables in the respective RSF model.[25,26] Therefore, they can also be obtained in the presence of multicollinearity among variables. As shown in the partial plots of Figure 4, the identified predictors appear to have a non-linear relation to predicted 5-year T2D-free survival. Besides direction of associations, partial plots are ideally suited to derive non-linear associations and possible clinical relevant cut points. This was also shown by the partial plots of the classical biomarkers. The cut points that can be derived from the partial plots of classical biomarkers are greatly in line with the guidelines of the American Diabetes Association.[62] This renders partial plots a useful tool of exploratory survival analysis in order to gain first indications for further research in the laboratory and the clinical environment.

The RSF backward algorithm and the previously applied stepwise CR approach were also applied and compared in two replication cohort studies. However, due to the low number of incident T2D cases in the two replication cohorts, most metabolites tested significantly in individual CR lost statistical significance after correction for multiple testing. Unfortunately, replication cohorts with higher numbers of incident T2D cases that measured Biocrates serum metabolites do not exist. Nevertheless, in contrast it was shown that the RSF backward algorithm can be applied to identify metabolites that are associated with incident T2D also in populations with a low number of incident cases and high numbers of variables. Some of the metabolites identified by the RSF approach also tested significant before correction of multiple testing in the individual CR tests. Interestingly, in the EPIC-Heidelberg cohort, some of the metabolites identified by the RSF approach differ only in the number of double bondings, pointing to a potential biological link. As in the EPIC-Potsdam study, identified metabolites improved the prediction of incident T2D in the replication cohorts when used together with traditional T2D risk factors (Supplementary Tables S3 and S4). However, the lower number of incident T2D cases in the replication cohorts, the older KORA population and the availability of diagnosis date only in years in the KORA cohort, limited a direct comparison of findings between the replication cohorts and the EPIC-Potsdam cohort for the RSF as well as for the CR approach.

One limitation of RSF is that this method does not immediately allow calculation of a relative risk for each variable, which is an intuitive and meaningful measure of association in epidemiological studies. Instead, the contribution of each marker to its relative relatedness with the endpoint needs to be assessed by minimal depth ranking. However, variables identified by RSF can be analysed in subsequent regression models to estimate relative risks. Yet regression models including all identified metabolites may not be appropriate, given the fact that RSF can independently select structurally related metabolites of high correlations. Furthermore, a disadvantage of tree-based methods is that they tend to prefer splits of continuous variables,[63] if the data consist of a mix of continuous and categorical variables. To minimize this bias, the number of splits chosen should be as small as possible. Accordingly, we chose the number of splits equal to 10.

Some additional methodological issues should be acknowledged when interpreting our findings. The EPIC-Potsdam study is not representative of the German general population[64] and the higher proportion of women and highly educated participants likely influences metabolite-disease associations. Moreover, the comparison of the two variable selection approaches was only possible in replication cohorts with a low number of incident T2D cases, resulting in an insufficient statistical power for the stepwise CR approach. Thus, even though it was shown that the RSF approach for variable selection can also be applied in data with a low number of cases and high number of variables, the external replication of the results obtained in the EPIC-Potsdam study was somewhat limited. Unfortunately, no other population-based studies with a comparable high number of incident T2D cases, in which metabolomics analyses using the Biocrates kit have been carried out, exist to our knowledge. Our data are also limited in terms of metabolite coverage. In fact, 75% of the covered metabolites were choline-containing phospholipids and another 18% were amino acids. Their close structural and metabolic inter-relationship is reflected by strong correlations between metabolites, which may have contributed to a preferential selection of highly correlated metabolites by the RSF algorithm. With 127 quantified serum metabolites, our study has been conducted in a small data set derived from targeted metabolomics profiling.

However, the advantages of RSF over CR may be more apparent in data sets containing a higher number of noise variables,[5] such as data derived from untargeted metabolomics.

Taken together, we believe that RSF is a sensible complement to CR. The introduced RSF backward algorithm is particularly suitable for variable selection when highly correlated complex survival data are investigated to identify unknown biomarkers associated with the disease of interest, taking into account possible confounders. Using the provided R-code, our RSF backward algorithm can be easily implemented and used to reduce the dimensionality of data derived from 'omic' sciences in order to improve the interpretability. However, partial plots are a first step to investigating the direction and potential non-linearity of individual metabolite-disease associations, and verification and translation of RSF findings into clinically understandable association measures should be a matter for future research.

Supplementary Data

Supplementary data are available at *IJE* online.

Acknowledgements

We thank the Human Study Centre (HSC) of the German Institute of Human Nutrition Potsdam-Rehbrücke, namely the trustee and the examination unit for the collection, the data hub for the processing and the participants for the provision of the data, the biobank for the processing of the biological samples, and the head of the HSC, Manuela Bergmann, for the contribution to the study design and leading the underlying processes of data generation.

Funding

The EPIC-Potsdam study was supported by a grant from the Federal Ministry of Education and Research, Germany (Bundesministerium für Bildung und Forschung) to the German Center for Diabetes Research (DZD), (grant number 01GI0922). The EPIC-Heidelberg study was supported by the Helmholtz Association (portfolio theme 'metabolic dysfunction') and the German Federal Ministry of Education and Research (BMBF, grant number 01ER0809). 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-2009-2.2.1-3/242114 (Project OPTiMiSE) and HEALTH-2013-2.4.2-1/602936 (Project CarTarDis), and by a grant from the BMBF to the German Center for Diabetes Research (DZD e. V.).

Conflict of interest: None declared.

References

1. Barderas MG, Laborde CM, Posada M *et al*. Metabolomic profiling for identification of novel potential biomarkers in cardiovascular diseases. *J Biomed Biotechnol* 2011;vol. **2011**: 790132. doi:10.1155/2011/790132.
2. Cheng S, Rhee EP, Larson MG *et al*. Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation* 2012;**125**: 2222–31.
3. 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* 2012;**62**: 639–48.
4. Wang TJ, Larson MG, Vasan RS *et al*. Metabolite profiles and the risk of developing diabetes. *Nat Med* 2011;**17**:448–53.
5. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. *Ann Appl Stat* 2008;**2**:841–60.
6. Datema FR, Moya A, Krause P *et al*. Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. *Head Neck* 2012;**34**:50–58.
7. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation* 2011;**4**:39–45.
8. Omurlu IK, Ture M, Tokatli F. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. *Expert Syst Appl* 2009; **36**:8582–88.
9. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. *Ann Nutr Metab* 1999;**43**:205–15.
10. Stefan N, Fritsche A, Weikert C *et al*. Plasma fetuin-A levels and the risk of type 2 diabetes. *Diabetes* 2008;**57**:2762–67.
11. Montonen J, Drogan D, Joost HG *et al*. Estimation of the contribution of biomarkers of different metabolic pathways to risk of type 2 diabetes. *Eur J Epidemiol* 2011;**26**:29–38.
12. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. *J Clin Epidemiol* 1999;**52**:1165–72.
13. Rathmann W, Kowall B, Heier M *et al*. Prediction models for incident type 2 diabetes mellitus in the older population: KORA S4/F4 cohort study. *Diabet Med* 2010;**27**:1116–23.
14. Wang-Sattler R, Yu Z, Herder C *et al*. Novel biomarkers for prediabetes identified by metabolomics. *Mol Syst Biol* 2012;**8**:615.
15. Riboli E, Hunt KJ, Slimani N *et al*. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;**5**(6B): 1113–24.
16. Boeing H, Wahrendorf J, Becker N. EPIC-Germany - A source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition. *Ann Nutr Metab*.1999;**43**:195–204.
17. Bergmann MM, Bussas U, Boeing H. Follow-up procedures in EPIC-Germany - Data quality aspects. *Ann Nutr Metab* 1999; **43**:225–34.
18. Kühn T, Floegel A, Sookthai D *et al*. Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study. *BMC Med* 2016;**14**:1–9.
19. 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**:133–42.

20. Floegel A, Drogan D, Wang-Sattler R *et al*. Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. *PLoS One* 2011;**6**:e21103.
21. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *J AmStat Assoc* 2010;**105**(489):205–17.
22. Ishwaran H, Kogalur UB. *RandomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC)*. 2015. https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf (15 July 2016, date last accessed).
23. Diaz-Uriarte R, de Andres SA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006;**7**:3.
24. Jiang H, Deng Y, Chen H-S *et al*. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 2004;**5**:81.
25. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001;**29**:1189–232.
26. Ishwaran H, Kogalur UB. Random survival forest for R. *R News* 2007;**7**:25–31.
27. Kramer N, Schafer J, Boulesteix AL. Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics* 2009;**10**:384.
28. Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* 2011;**40**:387–426.
29. Mamas M, Dunn WB, Neyses L, Goodacre R. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Arch Toxicol* 2011;**85**:5–17.
30. Dunn WB, Broadhurst DI, Deepak SM *et al*. Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate. *Metabolomics* 2007;**3**:413–26.
31. Kobayashi T, Nishiumi S, Ikeda A *et al*. A novel serum metabolomics-based diagnostic approach to pancreatic cancer. *Cancer Epidemiol Biomarkers Prev* 2013;**22**:571-79.
32. Abate-Shen C, Shen MM. DIAGNOSTICS. The prostate-cancer metabolome. *Nature* 2009;**457**:799–800.
33. Suhre K, Meisinger C, Döring A *et al*. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PloS One* 2010;**5**:e13953.
34. Bender R, Lange S. Adjusting for multiple testing - when and how? *J Clin Epidemiol* 2001;**54**:343–49.
35. Leigh JP. Assessing the importance of an independent variable in multiple regression: is stepwise unwise? *J Clin Epidemiol* 1988;**41**:669–77.
36. Harrell Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer, 2001.
37. Siroky DS. Navigating Random Forests and related advances in algorithmic modeling. *Stat Surv* 2009;**3**:147–63.
38. van der Schaaf A, Xu CJ, van Luijk P, Van't Veld AA, Langendijk JA, Schilstra C. Multivariate modeling of complications with data driven variable selection: guarding against overfitting and effects of data set size. *Radiother Oncol* 2012;**105**: 115–21.
39. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2006;**2**:171–96.
40. Xu T, Brandmaier S, Messias AC *et al*. Effects of metformin on metabolite profiles and LDL cholesterol in patients with type 2 diabetes. *Diabetes Care* 2015;**38**:1858–67.
41. Prosperi MC, Ingham SL, Howell A, Lalloo F, Buchan IE, Evans DG. Can multiple SNP testing in BRCA2 and BRCA1 female carriers be used to improve risk prediction models in conjunction with clinical assessment? *BMC Med Inform Decis Mak* 2014; **14**:87.
42. Gorodeski EZ, Ishwaran H, Kogalur UB *et al*. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the Women's Health Initiative. *Circulation* 2011;**4**:521–32.
43. Gall WE, Beebe K, Lawton KA *et al*. Alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. *PLoS One* 2010;**5**:e10883.
44. Ha CY, Kim JY, Paik JK *et al*. The association of specific metabolites of lipid metabolism with markers of oxidative stress, inflammation and arterial stiffness in men with newly diagnosed type 2 diabetes. *Clin Endocrinol (Oxf)* 2012;**76**:674–82.
45. Suhre K, Meisinger C, Doring A *et al*. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS One* 2010;**5**):e13953.
46. 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**:1711–20.
47. Drogan D, Dunn WB, Lin W *et al*. Untargeted metabolic profiling identifies altered serum metabolites of type 2 diabetes mellitus in a prospective, nested case control study. *Clin Chem* 2015; **61**:487–97.
48. Liu L, Wang M, Yang X *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;**59**:1338–48.
49. Cole LK, Vance JE, Vance DE. Phosphatidylcholine biosynthesis and lipoprotein metabolism. *Biochim Biophys Acta* 2012;**1821**: 754–61.
50. Ferrannini E, Natali A, Camastra S *et al*. Early metabolic markers of the development of dysglycemia and type 2 diabetes and their physiological significance. *Diabetes* 2013;**62**:1730–37.
51. 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**:e15234.
52. Zhao X, Fritsche J, Wang J *et al*. Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. *Metabolomics* 2010;**6**:362–74.
53. Sekhar RV, McKay SV, Patel SG *et al*. Glutathione synthesis is diminished in patients with uncontrolled diabetes and restored by dietary supplementation with cysteine and glycine. *Diabetes Care* 2011;**34**: 162–67.
54. Yamada C, Kondo M, Kishimoto N *et al*. Association between insulin resistance and plasma amino acid profile in non-diabetic Japanese subjects. *J Diabet Invest* 2015;**6**:408–15.
55. Wurtz P, Makinen VP, Soininen P *et al*. Metabolic signatures of insulin resistance in 7,098 young adults. *Diabetes* 2012;**61**: 1372–80.

56. 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**:E1060–65.
57. Tai ES, Tan ML, Stevens RD *et al*. Insulin resistance is associated with a metabolic profile of altered protein metabolism in Chinese and Asian-Indian men. *Diabetologia* 2010;**53**:757–67.
58. Newgard CB, An J, Bain JR *et al*. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell Metab* 2009;**9**:311–26.
59. Bao Y, Zhao T, Wang X *et al*. Metabonomic variations in the drug-treated type 2 diabetes mellitus patients and healthy volunteers. *J Proteome Res* 2009;**8**:1623–30.
60. Menni C, Fauman E, Erte I *et al*. Biomarkers for type 2 diabetes and impaired fasting glucose using a non-targeted metabolomics approach. *Diabetes* 2013;**62**:4270–76.
61. Wurtz P, Tiainen M, Makinen VP *et al*. Circulating metabolite predictors of glycemia in middle-aged men and women. *Diabetes Care* 2012;**35**:1749–56.
62. American Diabetes Association. Standards of Medical Care in Diabetes 2016. *Diabetes Care* 2016;**39**(**Suppl 1**): S52–59.
63. Loh W.Y. and Shih YS. Split selection methods for classification trees. *Statistica Sinica* 1997;**7**:815–40.
64. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. *Ann Nutr Metab* 1999;**43**:205–15.