Supplemental Text 2. Correlation filtering, machine learning algorithms and hyperparameter tuning

Total HDL, LDL and TRL particle concentrations and their changes were removed due to collinearity. In the baseline dataset, acetone was removed due to linear dependencies. By utilizing the $|r| > 0.8$ analytes correlation filter, LDL-C, medium HDL, H2P, LDL size, TRL-TG, TRL-C and ketone bodies were removed, while the more conservative $|r| > 0.6$ analyte correlation filter resulted in the removal of TC, HDL-C, ApoB, ApoA1, large HDL, medium HDL, H2P, large TRL, medium TRL, HDL size, LDL size, TRL-TG, TRL-C, LDL PPD, leucine, ketone bodies and BHB. In the dataset considering baseline and delta analytes, acetone and Δacetone were removed due to linear dependencies. By utilizing the $|r| > 0.8$ analyte correlation filter, TC, medium HDL, H2P, HDL size, TRL-TG, TRL-C, ketone bodies, Δmedium HDL, ΔH2P, ΔTRL-C, ΔBHB and ΔAcAc were removed, while the more conservative $|r| > 0.6$ analyte correlation filter resulted in the removal of LDL-C, HDL-C, ApoB, ApoA1, large HDL, medium HDL, H4P, H2P, large TRL, medium TRL, HDL size, LDL size, TRL-TG, TRL-C, LDL PPD, leucine, ketone bodies, BHB, Δlarge HDL, Δmedium HDL, Δsmall HDL, Δsmall LDL, ΔHDL size, ΔLDL size, ΔTRL-TG, ΔTRL-C, Δketone bodies and ΔAcAc.

In this paper, eight statistical and machine learning algorithms were utilized and compared. The seven algorithms were: Logistic Regression (GLM), Cox proportional hazards models (COX) using the *survival* and *rms* R packages, Stochastic Gradient Boosting (SGB) using the *gbm* R package, Random Forest (RF) using the *ranger* R package, Support Vector Machines using the *kernlab* R package with a linear kernel (SVM-L), polynomial kernel (SVM-P) and radial kernel (SVM-R) and Artificial Neural Network (ANN) using the *nnet* R package. The *caret* R package was used for the grid search / hyperparameter tuning.

The possible hyperparameters for these methods were:

GLM:

none

COX:

none

SGB:

number of boosting iterations: 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000, 1050, 1100, 1150, 1200, 1250, 1300, 1350, 1400, 1450, 1500.

maximum tree depth: 1, 2, 4

shrinkage: 0.001, 0.05, 0.1

minimal terminal node size: 20

RF:

number of randomly selected predictors: 2, 3, 4, 5

splitting rule: gini, extratrees

minimal node size: 1, 3, 5

SVM-L:

cost: 0.01, 0.1, 1, 2, 4, 8, 16, 32

SVM-P:

polynomial degree: 2, 3, 4, 5

scale: 0.1, 0.01, 0.005, 0.001

cost: 0.01, 0.1, 1, 2, 4, 8, 16, 32

SVM-R:

sigma: $1/2^{25}$, $1/2^{20}$, $1/2^{15}$, $1/2^{10}$, $1/2^{5}$, 1

cost: 0.01, 0.1, 1, 2, 4, 8, 16, 32

ANN:

number of hidden units: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

weight decay: 0.01, 0.1, 1, 3, 5, 10, 20

The tuned hyperparameters in the five separate test sets for short- and long-term T2D (Model 8, correlation filter $|r| > 0.6$) are shown in the table below:

		T2D short - test1	**T2D short - test2**	**T2D short - test3**	**T2D short - test4**	**T2D short - test5**	**T2D long - test1**	**T2D long - test2**	**T2D long - test3**	**T2D long - test4**	**T2D long - test5**
SGB	number of boosting iterations	250	50	300	100	50	300	50	100	50	550
	maximum tree depth	1	4	1	2	4	1	4	1	2	2
	shrinkage	0.05	0.05	0.05	0.05	0.05	0.10	0.10	0.10	0.10	0.05
	minimal terminal node size	20	20	20	20	20	20	20	20	20	20
RF	number of randomly selected predictors	5	3	2	2	2	2	5	5	5	4
	splitting rule	gini	gini	gini	gini	gini	gini	gini	gini	extratrees	gini
	minimal node size	3	5	3	3	1	1	5	5	5	5
SVM-L	cost	0.10	0.01	2.00	0.01	16.0	0.10	0.01	2.00	0.01	0.01
SVM-P	polynomial degree	5	5	4	5	2	4	5	4	4	5
	scale	0.01	0.01	0.01	0.01	0.10	0.001	0.01	0.001	0.005	0.001
	cost	0.10	0.10	0.01	0.10	0.01	32.00	0.1	16.00	8.00	8.00
SVM-R	sigma	2^{-5}	2^{-5}	2^{-5}	2^{-5}	2^{-10}	2^{-5}	2^{-5}	2^{-5}	2^{-5}	2^{-10}
	cost	0.01	0.01	0.01	0.01	1.00	1.0	0.1	2.0	1.0	32.0
ANN	number of hidden units	14	3	10	4	3	14	6	3	10	3
	weight decay	5	10	3	3	5	1	1	1	20	1