Statistical analysis in metabolic phenotyping

Supplementary File 1

Benjamin J. Blaise,[1,2,3*] Gonçalo D. S. Correia,[1,4*] Gordon A. Haggart,[1,4,*] Izabella Surowiec,[5,6] Caroline Sands,[1,4] Matthew R. Lewis,[1,4] Jake T. M. Pearce,[1,4] Johan Trygg,[5,6] Jeremy K. Nicholson,[1,7,9] Elaine Holmes[1,8] and Timothy M. D. Ebbels.[1]

1 : Division of Systems Medicine, Department of Metabolism, Digestion & Reproduction, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK
2 : Department of Paediatric Anaesthetics, Evelina London Children's Hospital, Guy's and St Thomas' NHS Foundation Trust, London SE1 7EH, UK
3 : Centre for the Developing Brain, King's College London, SE1 7EH, UK
4 : National Phenome Centre, Department of Metabolism, Digestion & Reproduction , Faculty of Medicine, Imperial College London, London SW7 2AZ, UK
5 : Computational Life Science Cluster, Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden
6: Sartorius Corporate Research, Sartorius Stedim Data Analytics, 903 33, Umeå, Sweden
7 : Australian National Phenome Centre, Health Futures Institute, Murdoch University, Perth, Western Australia 6150, Australia
8: Centre for Computational & Systems Medicine Institute of Health Futures, Murdoch University, Perth, Western Australia, Australia.
9 : Institute of Global Health Innovation, Imperial College London, Level 1, Faculty Building South Kensington Campus, London, SW7 2NA, UK.

*: These authors contributed equally to this work

Correspondence should be addressed to Timothy M. D. Ebbels (t.ebbels@imperial.ac.uk)

Data analysis of an LC-MS dataset from a human urine biofluid cohort study.

This supplementary file contains two extra examples of data analysis workflows, applied to a liquid chromatography mass spectrometry dataset (LC-MS). The LC-MS dataset comes from a metabolic phenotyping investigation of human urine biofluid samples from a dementia cohort. In this sample set, baseline spot urine samples (first sample collected after recruitment to the study) were collected as part of the AddNeuroMed[1] and ART/DCR study consortia, with the aim of identifying biomarkers of neurocognitive decline and Alzheimer's disease. These samples were analysed by LC-MS and ^{1}H NMR, using the methods described by Lewis *et al*[2] and Dona *et al*[3]. Detailed information about this cohort and other available phenotypic measurements can be found in Lovestone and the ANMERGE[4] repository, which can be accessed via the Sage BioNetworks portal (https://doi.org/10.7303/syn22252881). Information about the metabolic profiling experiments can be found in the study's MetaboLights entry: https://www.ebi.ac.uk/metabolights/MTBLS719.

1.1 - Discrimination of biological sex from urinary metabolic profiles using Partial Least Squares – Discriminant Analysis

PROCEDURE

The workflow can be divided in multiple segments. The exact steps are written in the Jupyter Notebook with the title "Discrimination of biological sex from urinary metabolic profiles using Partial Least Squares – Discriminant Analysis". Run-time estimates for a desktop workstation with 8 cores and 16GB RAM are reported for each section.

Software installation - timing ~10 min

1. Download and install the Python 3 Anaconda distribution (with Python version 3.7 or above). Directions for installation can be found here: www.anaconda.com
2. Download the complete dataset (https://doi.org/10.5281/zenodo.4053167) as a .zip file. Links to more documentation on how to use the Jupyter notebook are listed there in the Readme file.
3. Launch Anaconda
4. From the Anaconda Navigator, launch a Qt Console. Type the command line "conda install plotly" and then press enter. This will install the display and plotting package.
5. From the Anaconda Navigator, launch a Jupyter Notebook session. In the browser window that opens, navigate to the work folder where the notebooks were downloaded. Open the folder chemometrics-tutorials-urineLCMS. The Jupyter Notebook for this tutorial can now be read.
6. NOTE: You can click the icon Run in the command line to execute the different parts of the code. It is important to run each section one after the other, as some sections might need elements calculated in previous sections to run correctly. Please note on the left side, the blue characters In [ ] before the execution of a code. When a code is being run, this will change to In [*] and then a number will replace the star when the code has been executed. Allow enough time for your computer to run the codes. Display windows might need to be closed after plotting many of them. Just press on the blue power sign on the top right corner of the display. Plot will stay on the screen, but interactions (zoom, move, export) will be removed.

Supervised Analysis with PLS-DA - timing ~ 30 min

7. Open the notebook with the title “Discrimination of biological sex from urinary metabolic profiles using Partial Least Squares – Discriminant Analysis”.
8. Import the reversed phase positive mode LC-MS assay (RPOS) dataset “*Dementia_RPOS_XCMS.csv*”. This file contains the probabilistic quotient normalised LC-MS metabolic profiles and the demographic variables (e.g., gender).
9. Start by selecting the type of scaling and run a Partial Least Squares-Discriminant Analysis model to see if your samples can be discriminated with a supervised approach. In this case, we suggest log-transforming the dataset and using mean centering only. The effect of log-transformation of high intensity values has the advantage of attenuating the potential impact of outliers on the robustness of the models obtained[5].
10. NOTE: We suggest performing an exploratory PCA analysis to assess the overall data quality and remove analytical outliers, before carrying out supervised analyses. A worked example of such an analysis can be found in the “Multivariate Analysis – PCA” tutorial described in the main protocol (available from https://github.com/Gscorreia89/chemometrics-tutorials)
11. Assess the scree plot of R^2 and Q^2 versus number of components to estimate the optimal model complexity (Figure S1A). By default, this estimate will be based on the stabilization of the AUC measure with increasing component number (increase compared to the previous component of 5% or less) NOTE: Since this is a classification problem, the AUC measure is more suitable for PLS component choice than the Q^2_Y.
12. Assess the score plot to see if there are any outliers, decide on their exclusion, and refit the model if necessary (Figure S1B).

Model prediction capacity and validation - timing ~40min

13. OPTIONAL: Use double cross-validation to select the best-performing combination of parameters and obtain a more reliable (compared to the cross-validated scree-plot procedure above) cross-validation estimate of the model generalization performance on separate external test-data. Set up two K-Fold cross-validation loops, an inner CV loop to optimise the parameters over a predefined grid of parameter values, and an outer CV loop to obtain model performance estimates which are independent of the parameter

optimisation. NOTE: Double cross-validation schemes might not be suitable with small sample sizes. For those cases, the simpler scree plot method is suggested.

14. OPTIONAL: Refit and cross-validate the model with the number of components chosen using double cross-validation. Plot a cross-validated ROC curve and obtain the area under the curve (AUC) measure (Figure S1C).
15. Run a permutation test to simulate the null hypothesis and check that the AUC for the ROC curve obtained in the fitted model cannot be easily obtained by chance alone (Figure S1D). CRITICAL STEP: Verify the empirical p-value obtained for AUC, if it is higher than the significance threshold, i.e., $p > 0.05$, the classifier is unreliable.

Selection of important variables - timing ~20min

16. Inspect the model weights (w) for the first PLS component and the regression coefficients (β) to find which variables are associated with each class. These values are interpreted in the same manner as univariate correlation for w, and as linear regression coefficients for β. Match the sign with the direction of the Y vector containing outcomes. Alternatively, compute the variable influence on projection (VIP) coefficients, which are based on all the weight vectors instead of only the first. The VIP values are lower bounded at 0 (non-relevant variable), and the higher the value, the more relevant a variable is for the model.

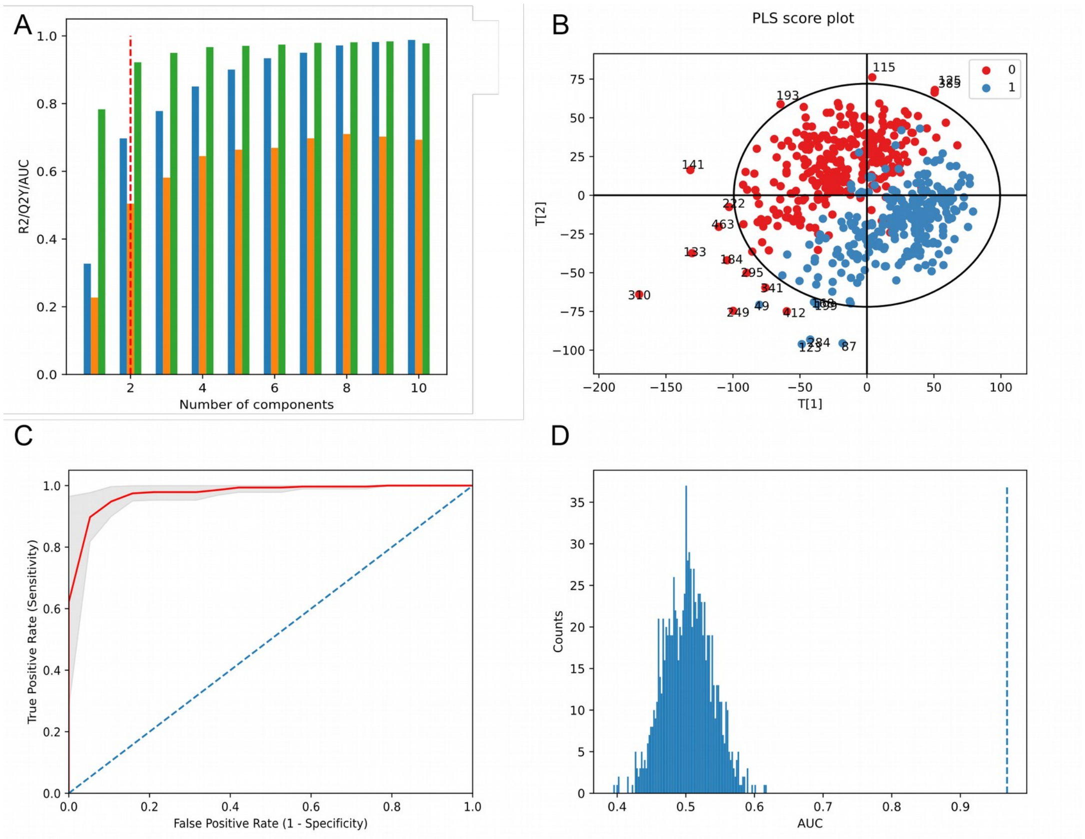

Figure S1 – Classification and model diagnostic outputs for the biological sex discrimination with PLS-DA. A – Scree plot used to suggest the number of PLS components to use. R^2_Y – blue, Q^2_Y – orange, ROC AUC - green. The AUC measure stabilizes for the model with 2 PLS components. B – PLS-DA scores plots for the discrimination between female (0, red) and male (1, blue). C – Mean (red line) cross-validated ROC curve, with ± 1 standard deviation confidence band (shaded grey). D – Permutation test results. The AUC obtained in the non-permuted dataset (AUC = 0 0.974, vertical dashed line) is much higher than those from the permuted null distribution (blue histogram), and therefore the classifier appears to be reliable (p-value < 0.001).

1.2 - Discrimination of biological sex from urinary metabolic profiles using random forests

Metabolic phenotyping data can be analysed with multiple statistical and machine learning methods, as mentioned in Box 2 in the main text. As an example, in this section we provide a template implementation for a discrimination analysis with a random forest classifier. This is exemplified by replicating with random forests the biological sex discrimination in the urine LC-MS dataset performed in the previous section with PLS-DA. The following procedure details the general steps and validation workflow recommended when analysing a metabolic phenotyping dataset with random forest models.

PROCEDURE

The workflow can be divided in multiple segments. The exact steps are written in the Jupyter Notebook with the title “Discrimination of biological sex from urinary metabolic profiles using Random Forests”. Run-time estimates for a desktop workstation with 8 cores and 16GB RAM are reported for each section.

Software installation - timing ~10 min

1. Download and install the Python 3 Anaconda distribution (with Python version 3.7 or above). Directions for installation can be found here: www.anaconda.com
2. Download the complete dataset (https://doi.org/10.5281/zenodo.4053167) as a .zip file. Links to more documentation on how to use the Jupyter notebook are listed there in the Readme file.
3. Launch Anaconda
4. From the Anaconda Navigator, launch a Qt Console. Type the command line “conda install plotly” and then press enter. This will install the display and plotting package.
5. From the Anaconda Navigator, launch a Jupyter Notebook session. In the browser window that opens, navigate to the work folder where the notebooks were downloaded. Open the folder chemometrics-tutorials-urineLCMS. The Jupyter Notebook for this tutorial can now be read.
6. NOTE: You can click the icon Run in the command line to execute the different parts of the code. It is important to run each section one after the other, as some sections might need elements calculated in previous sections to run correctly. Please note on the left side, the blue characters In [ ] before the execution of a code. When a code is being run, this will change to In [*] and then a number will replace the star when the code has been executed. Allow enough time for your computer to run the codes. Display windows might need to be closed after plotting many of them. Just press on the blue power sign on the top right corner of the display. Plot will stay on the screen, but interactions (zoom, move, export) will be removed.

Supervised Analysis with Random Forests - timing ~ 20 min

1. Open the notebook with the title “Discrimination of biological sex from urinary metabolic profiles using Random Forests”.

2. Import the reversed phase positive mode LC-MS assay (RPOS) dataset "*Dementia_RPOS_XCMS.csv*". This file contains the probabilistic quotient normalised LC-MS metabolic profiles and the demographic variables (e.g., gender).
3. Fit a random forest classifier model to the entire dataset to discriminate biological sex.
4. Explore the impact of the main random forest model parameters on the complexity of decision trees generated and impact on predictive capacity. Fit a model using the default recommendations and assess the prediction performance using out-of-bag (samples not used for training) estimates.

Model prediction capacity and validation - timing ~ 1h30 hour

5. Use double cross-validation to select the best-performing combination of parameters and obtain an unbiased cross-validation estimate of the model generalisation performance on separate external test-data. Set up two K-Fold cross-validation loops, an inner CV loop to optimize the parameters over a predefined grid of parameter values, and an outer CV loop to obtain model performance estimates which are independent of the parameter optimization.
6. Plot a cross-validated ROC curve and obtain the area under the curve (AUC) measure from the final model (Figure S2A).
7. Run a permutation test to simulate the null hypothesis and check that the AUC obtained for from the fitted RF classifier cannot be easily obtained based on chance alone (Figure S2B). CRITICAL STEP: Verify the empirical *p-value* obtained for AUC, if it is higher than the significance threshold, i.e., $p > 0.05$, the classifier is unreliable.

Selection of important variables - timing ~ 15min

8. Use the Gini importance measure to rank all features based on their importance for the model prediction.
9. OPTIONAL: Compare the ranked feature list derived with the random forest classifier with those obtained with the corresponding PLS-DA model.

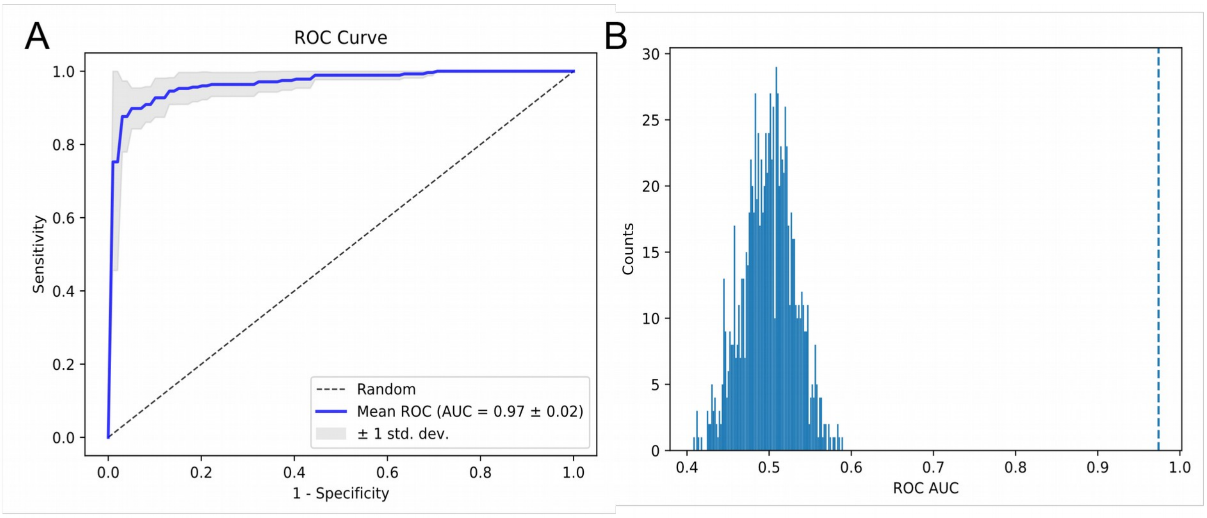

Figure S1 – Classification and model diagnostic outputs for the biological sex discrimination with random forests. A – Mean (blue line) cross-validated ROC curve, with ± 1 standard deviation confidence band (shaded grey). B – Permutation test results for random forest classifier. The AUC obtained in the non-permuted dataset (AUC = 0.97, vertical dashed line) is much higher than those from the permuted null distribution (blue histogram), and therefore the classifier appears to be reliable (p-value < 0.001).

References:

1. Lovestone, S. *et al.* AddNeuroMed - The european collaboration for the discovery of novel biomarkers for alzheimer's disease. in *Annals of the New York Academy of Sciences* (2009). doi:10.1111/j.1749-6632.2009.05064.x.

2. Lewis, M. R. *et al.* Development and Application of UPLC-ToF MS for Precision Large Scale Urinary Metabolic Phenotyping. *Anal. Chem.* **88**, acs.analchem.6b01481 (2016).

3. Dona, A. C. *et al.* Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal. Chem.* **86**, 9887–94 (2014).

4. Birkenbihl, C. *et al.* ANMerge: A comprehensive and accessible Alzheimer's disease patient-level dataset. *medRxiv* (2020) doi:10.1101/2020.08.04.20168229.

5. Baxter, M. J. Standardization and Transformation in Principal Component Analysis, with Applications to Archaeometry. *Appl. Stat.* (1995) doi:10.2307/2986142.