A Novel Method for Power Analysis and Sample Size Determination in Metabolic Phenotyping

Journal:	*Analytical Chemistry*
Manuscript ID	ac-2016-00188w.R1
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Blaise, Benjamin; Hospices Civils de Lyon, HFME - reanimation neonatale; Imperial College London, Biomolecular Medicine, Division of Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine Correia, Gonçalo; Imperial College London, Surgery and Cancer Tin, Adrienne; Johns Hopkins University Bloomberg School of Public Health, Department of Epidemiology Young, J. Hunter; Johns Hopkins University Bloomberg School of Public Health, Department of Medicine Vergnaud, Anne-Claire; Imperial College London, Department of Epidemiology and Biostatistics Lewis, Matthew; Imperial College London, Biomolecular Medicine Department of Surgery and Cancer, Faculty of Medicine Pearce, Jake; Imperial College London, Biomolecular Medicine Elliott, Paul; Imperial College London, Epidemiology and Public Health Nicholson, Jeremy; Imperial College London, Surgery and Cancer; Imperial College of London, Holmes, Elaine; Imperial College, Biomedical Sciences Ebbels, Timothy; Imperial College London, Biomolecular Medicine

A Novel Method for Power Analysis and Sample Size Determination in Metabolic Phenotyping

Benjamin J. Blaise[1,2,*], Gonçalo Correia[1,*], Adrienne Tin,[3] J. Hunter Young,[4] Anne-Claire Vergnaud,[5] Matthew Lewis[1,6], Jake T.M. Pearce[1,6], Paul Elliott,[5] Jeremy K. Nicholson[1], Elaine Holmes[1] and Timothy M. D. Ebbels[1]

1 : Biomolecular Medicine, Division of Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London SW7 2AZ, UK

2 : Hospices Civils de Lyon, Service de réanimation néonatale et néonatalogie, Hôpital Femme Mère Enfant, 59 bd Pinel, 69677 Bron Cedex, France

3: Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe Street, Baltimore, MD 21205, USA

4: Johns Hopkins Bloomberg School of Public Health, Department of Medicine, The Johns Hopkins University and The Welch Center for Epidemiology and Clinical Research, 2024 E. Monument Street, Baltimore, MD 21205, USA

5: Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, St. Mary 's Campus, Norfolk Place, W2 1PG London, United Kingdom

6: MRC-NIHR National Phenome Centre, Department of Surgery and Cancer, Imperial College London, IRDB Building, Du Cane Road, London W12 0NN, UK

*: These authors contributed equally to this work

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

keywords: metabolic phenotyping, sample size, effect size, power analysis, experimental design

Abstract

Estimation of statistical power and sample size is a key aspect of experimental design. However, in metabolic phenotyping, there is currently no accepted approach for these tasks, in large part due to the unknown nature of the expected effect. In such hypothesis free science, neither the number or class of important analytes, nor the effect size are known *a priori*. We introduce a new approach, based on multivariate simulation, which deals effectively with the highly correlated structure and high-dimensionality of metabolic phenotyping data. First, a large data set is simulated based on the characteristics of a pilot study investigating a given biomedical issue. An effect of a given size, corresponding either to a discrete (classification) or continuous (regression) outcome is then added. Different sample sizes are modeled by randomly selecting data sets of various sizes from the simulated data. We investigate different methods for effect detection, including univariate and multivariate techniques. Our framework allows us to investigate the complex relationship between sample size, power and effect size for real multivariate data sets. For instance, we demonstrate for an example pilot data set, that certain features achieve a power of 0.8 for a sample size of 20 samples, or that a cross-validated predictivity Q^2_Y of 0.8 is reached with an effect size of 0.2 and 200 samples. We exemplify the approach for both Nuclear Magnetic Resonance and Liquid Chromatography – Mass Spectrometry data from humans and the model organism *C. elegans*.

Introduction

Ethical considerations and economic constraints urge scientists and physicians to properly design their experiments. In the commonly used inferential framework of null hypothesis significance testing (NHST), multiple observations are analyzed and summarized through a model and a respective test statistic, from which a *p-value* can be estimated. Different types of error can occur in NHST based inference.[1] Of particular interest are the significance level α, that represents the risk of falsely identifying truly negative results as statistically significant, and the parameter β that represents the risk of falsely rejecting truly positive results as non-significant. They are also referred as type I (false positive) and type II (false negative) errors, respectively.

Statistical power analysis relates sample size, effect size and significance level to the chance of detecting an effect in a dataset. In inferential statistics, the null hypothesis H_0 represents the absence of an effect, such as a relationship or difference between the measured phenomena, whereas the alternative hypothesis H_A represents the existence of an effect. The power is the probability 1-β of flagging a true effect as statistically significant: P(reject $H_0|H_A$ is true). For study design purposes, these analyses are usually performed by fixing power at a desired level (usually 80-90%, thus leading to false rejection of true effects in 10-20% of the cases), and estimating the sample size required, given an effect size, significance level and particular test to be used.[2] In power analysis, the effect size corresponds to the quantitative measure of the strength of a phenomenon relative to the variation in the population (*e.g.* how different two groups of samples are relative to the within-group variance). The stronger the effect, the more easily it will be detected, thus requiring a smaller number of samples to meet similar power requirements.

Statistical power analysis should be performed before the beginning of a study. It is a safeguard that estimates the probability of obtaining meaningful results, and thus success of a study.[3,4] Pilot studies are the primary source of information for power calculations, giving information on variable distributions and effect sizes.[5,6] A pilot of 20 samples has been suggested as sufficient to perform robust power analysis.[7,8,9] The protocol used in the pilot and main studies should be identical. If pilot data are not available, useful information might be retrieved from relevant literature, such as epidemiological studies. Despite some methodological debates about *a posteriori* power calculations,[10] these studies are useful to estimate variation within the population and thus to aid future experimental design.[2] As long as the sampling and analytical protocols follow standard operating procedures, data can be assumed to be reproducible. It is, for instance, possible to use the control group of a large cohort to evaluate an effect not associated to the initially studied effect.

Metabolic phenotyping is being increasingly applied in clinical and epidemiological studies,[11,12] yet meaningful statistical power calculations in this field are hard to accomplish. In common with other 'omics' techniques, metabolic phenotyping is normally used in an untargeted, hypothesis free, top-down approach, without prior knowledge of the important molecular markers (in this case, metabolites). However, some challenges are unique to the metabolic field. For example, the number of observed metabolites is unknown *a priori,* and highly dependent on the analytical platform and the dynamic range of metabolite concentrations, considerations that do not apply to genomics. The data are characterized by high-dimensionality, with typically hundreds of samples but thousands of variables[13]. Further, there is strong multicollinearity between variables, arising both from technical sources (different signals from same chemical entities, for e.g. adducts in mass spectrometry) but also due to the inter-metabolite relationships comprising the metabolic network.[14,15] Since in these studies thousands of variables can be surveyed at once, multiple hypotheses testing corrections must be performed, for example by controlling the family wise error rate (FWER) or false discovery rate (FDR).[16–18] These issues complicate the task of designing an experiment with the adequate sample size to precisely detect and estimate the magnitude of a metabolic effect.

Metabolic phenotyping studies deal with both discrete (classification) and continuous (regression) outcome variables. They usually aim at both sample classification where the focus is on prediction, and the identification of candidate biomarkers, with a focus on variable selection/importance. Despite recent efforts, there is still no widely accepted method for sample size determination in metabolic phenotyping. Power analysis is often avoided and sample size determination becomes driven by sample availability or is based on pilot data or extrapolated from the literature. Approaches developed in other fields either preclude the hypothesis-free approach, or provide a limited investigation of effect and sample sizes.[19] Moreover, they typically do not account for the strong correlation structure seen in metabolic phenotyping data arising both from the fact that multiple signals can belong to the same compound and the fact that metabolites in shared metabolic pathways are often not independent of each other. The Data driven sample size determination (DSD) algorithm[20] was developed to be used with small pilot study data and for a specific set of univariate analyses, but does not account for correlation in the data. Another algorithm, MetSizeR[21] offers calculations for some multivariate latent variable models, which are rarely used in metabolic phenotyping. Other approaches have been derived for other "omics"[19,22]; however, they rely on an *a priori* estimation of the effect size, independent from the variables under consideration. Trutschel et al. recently exemplified how pilot study data, including technical replicates and quality controls, can optimize experimental design, by evaluating the different sources of variance.[23] Recently, a sample size determination module has been implemented in MetaboAnalyst 3.0,[24] based on the Bioconductor package Sample Size and Power Analysis developed for genomics[25] but this approach doesn't take into account correlation between variables and relies on an a concept of summary average power for the dataset[24]. However, there is no reason to expect that each variable exhibits the same power, and since for most studies no preconception about which variables will be affected exists, it is preferable to set a study sample size to a number where the majority of variables reach a minimum level of power.

In this study, we introduce a new approach, which explicitly incorporates the correlations between variables characteristic of metabolic datasets, to investigate the relationship between statistical power, sample and effect size, and obtain estimates of the required sample size for metabolic phenotyping studies. Based on pilot data, we simulate new samples with marginal distributions and correlation structure similar to the ones observed in the pilot data. These can be used to study the sensitivity of power and other metrics to sample size. The data are simulated using a multivariate log-normal distribution fit to the pilot data, which allows us to maintain the long-tails and strong correlations that are typically seen in metabolic studies. Then, the desired effect size is introduced in the simulated dataset, depending on the type of outcome variables used in the data set and the statistical method intended for data analysis. This procedure is repeated multiple times with samples randomly drawn from the simulated data and the outcomes of the statistical analyses are stored and used to derive estimates and confidence intervals of performance statistics (e.g. true positive, false negative rates), from which power and other quantities of interest may be calculated. Different multiple hypothesis testing corrections can be used and their effect on power and efficiency (*e.g.* in controlling FDR) benchmarked.

Our proposed method provides a straightforward way to perform power analysis and sample size determination in metabolic phenotyping studies using any spectroscopic platform (e.g. nuclear magnetic resonance (NMR) or mass spectrometry (MS)). It captures the dependence structure between variables and allows a synchronized investigation of how sample and effect sizes affect power for each variable individually, or for multivariate approaches, identification of significant features or sample classification and prediction capability. Experimental design in metabolic phenotyping is performed by extracting the effect size from pilot data and determining the corresponding sample size to reach the desired level of power, using a specific statistical analysis method strategy and applying stringent multiple hypothesis testing corrections.

Methods

Data sets. This study uses data sets published and described elsewhere.

C. elegans data set. High-resolution magic angle spinning (HRMAS) ^{1}H NMR experiments were carried out on a Bruker Avance II spectrometer operating at 700MHz. The data set is composed of 33 spectra of entire wild type *C. elegans*.[26,27] 16k data points with 512 scans were acquired using a spectral width of 8503Hz, for a total acquisition time of approximately 25 minutes. ^{1}H HRMAS NMR spectra were phased and referenced to the β proton signal of alanine (δ = 1.48 ppm) using Topspin 2.1 (Bruker GmbH, Rheinstetten, Germany). Residual water signal (4.61 to 4.99ppm) was excluded. Spectra were divided into 0.001 ppm-wide bins over the chemical shift range [0; 10ppm] using the AMIX software (Bruker GmbH).

Atherosclerosis Risk in Communities Study (ARIC) data set. 1861 human urine spot samples from a community-based cohort of middle aged adults (55-65) in the U.S., measured by standard ^{1}H NMR.[28] The NMR pulse sequence and spectra acquisition parameters are detailed elsewhere.[29]

AIRWAVE Health Monitoring Study data set. AIRWAVE is a cohort study of police and emergency workers that has been designed to evaluate the effects of portable communication devices on health.[30] 951 standard human plasma samples were profiled by reversed phase Ultra Performance Liquid Chromatography-Mass Spectrometry (UPLC-MS), with an ACQUITY UPLC and Xevo G2-S oaToF MS in positive electrospray ionization mode (both Waters Corp., Milford, MA, USA).

Preprocessing. Different preprocessing steps were used on the data sets to accommodate analytical characteristics.

C. elegans data set. Statistical Recoupling of Variables (SRV) was used to identify NMR peaks in spectra.[31,32] Recoupling parameters were determined empirically, to obtain satisfactory identification of metabolic features, based on previous biochemical characterization of the data set (singlet base width=0.01ppm, bucketing resolution=0.001ppm, correlation threshold=0.8).

ARIC data set. 24 metabolites (creatinine, creatine, D-glucose-beta, alanine, lactate, acetate, succinate, citrate, dimethylamine, trimethylamine, betaine, glycine, fumarate, formate, 1-methylnicotinamide, N-dimethylglycine, trigonelline, hippurate, D-lactose, acetone, D-3-hydroxybutyric acid, acetoacetic acid, benzoic acid and ethanol) were quantified by Bruker BioSpin GmbH, using proprietary quantification technology. The data set was normalized by the quantified creatinine values and the other 23 metabolites used in the power analysis.

AIRWAVE data set. MS features where detected using Progenesis QI (Nonlinear Dynamics). Using pooled quality control (QC) samples injected periodically throughout the analytical run, and a QC dilution series, unreliable features were removed. Only features with a relative standard deviation below 20% in QC samples, and a Spearman correlation to the dilution factor higher than 0.9 and were kept. For the power analysis, a small number of remaining negative value artefacts were set to zero.

Data simulation. The workflow of the simulation and power estimation process is summarized in SI Figure 1. To efficiently simulate the long-tailed distributions and correlations present in biochemical measurements, pilot data were log-transformed before simulation and modeled in the log space as a multivariate normal distribution. An offset was added to avoid negative values prior to log transformation. A total of 5000 simulated profiles were sampled from a multivariate normal distribution with mean and covariance estimated from the log-transformed pilot data using the mvnrnd function from MATLAB^{TM} (Matlab R2014a, Mathworks, Natick, MA). Data were then exponentiated to the original scale, and the offset was subtracted to generate the final simulated data set corresponding to a multivariate log-normal distribution. A small number of remaining negative values were set to zero (for example, in the *C. elegans* data set, less than 5% of the total values were affected, with a maximum change in the mean of any variable of 0.72%). These negative values result from the use of a non-truncated normal distribution. The same simulation process was performed setting the off-diagonal elements of the covariance matrix to zero to study the effect of correlation on power analysis and sample size determination.

Sample and effect sizes. Sample size was investigated by selecting data sets of various sizes from the simulated data. The effect was introduced in different ways, depending on the nature of the intended statistical analysis. In the case of a two-group comparison (the classification case), two similar datasets were generated. An effect of a given size was then applied to one of them, on a single predetermined variable and its highly correlated partners. Standardized effect sizes (ES) were implemented by adding to each selected variable *X_i,* the product of the standard deviation of the variable σ and a number *ES*, *$0 \leq ES \leq 1$* (in steps of 0.05), as shown in Equation 1. To preserve the correlation structure in the effect, the same effect was implemented on all variables showing a high Pearson correlation (r> 0.8) to the effect variable.

Equation 1: $X_i \rightarrow X_i + ES \times \sigma$

For the regression case, a single data set was generated, and an effect of a given size was applied to a single predetermined variable. The effect was introduced for the selected variable X_i by simulating an outcome variable *Y* according to equation 2. The effect size β is chosen in the interval [0, 1] (with increments of 0.05). Normally distributed random noise was added (mean=0, standard deviation=1) to mimic biological and technological variability.

Equation 2: $Y = X_i \times \beta + \varepsilon \qquad \varepsilon \approx N(0,1)$

For each combination of sample and effect size, 100 repeats were conducted using different randomly selected subsets of the simulation, to generate standard errors on the mean or 95% confidence intervals for the obtained results.

Power analysis. The sample and effect sizes were then investigated in terms of prediction accuracy and identification of candidate biomarkers.

For the identification of statistically significant variables in the classification case, both univariate and multivariate approaches were considered. For the univariate case, a one-way analysis of variance (ANOVA) was used to investigate the intra and inter-group variances, and compute a corresponding *p-value* to identify statistically significant variations. Multiple hypothesis testing corrections were then applied to control type 1 error. For *m* hypothesis (*i.e.* number of tests performed), the Bonferroni (equation 3), Benjamini-Hochberg (Equation 4) and Benjamini-Yekutieli (Equation 5) procedures[16–18] are defined as follows:

Equation 3: $\alpha' = \alpha / m$

Equation 4: $\alpha' = \max_k \left(p_k \leq \frac{k}{m}\alpha \right)$

Equation 5: $\alpha' = \max_k \left(p_k \leq \frac{k}{m\sum_{i=1}^{m}\frac{1}{i}}\alpha \right)$

Here, α is the family wise error rate or false discovery rate, α' is the adjusted per-test significance level and p_k the p-value of the *k'th* most significant variable. By default the Benjamini-Yekultieli procedure was used. We advocate its use for metabolic phenotyping studies, given its ability to deal with correlated data under negative dependence. The variables highlighted as significant were then compared to the variables selected to show effects in the simulation. This allows the identification of true positives (TP, variables that were selected and identified as significant), true negatives (TN, variables not selected and not significant), false positives (FP, variables that were not selected but identified as significant) and false negatives (FN, variables which were selected, but not significant).

In the regression scenario, the outcome variable Y was regressed on each variable to obtain estimated regression coefficients β_i, and the corresponding *p-value* enabling identification of variables significantly associated to the outcome. False positive rates were controlled by multiple hypothesis testing corrections as above. True/false positives/negatives were identified by comparing the variables selected in the simulation to those deemed significant according to the regression.

For all univariate approaches, the significance threshold (FWER or FDR) was set to 0.05. The numbers of true/false positives/negatives were summarised in confusion matrices[33] from which all relevant performance statistics could be calculated. Performance statistics used were true negative/positive rates, false negative/positive, power and false discovery rates (TPR=Power=TP/(TP+FN), TNR=TN/(FP+TN), FPR=FP/(FP+TN), FNR=FN/(TP+FN), FDR=FP/(FP+TP)). In the univariate case such statistics refer to the ability to detect important variables.

In addition to univariate analysis, multivariate analyses were implemented. An Orthogonal-Partial Least Squares regression (O-PLS)[34] was performed to discriminate samples belonging to the 2 simulated groups in the classification case. The number of components was determined as the smallest number such that the Q^2Y predictive goodness-of-fit parameter did not increase by more than 5% on adding a subsequent component.[35] A confusion matrix was then computed to assess classification performance, considering the initial Y classification vector and the predicted Y obtained from the O-PLS model.

Sample size determination for each variable based on pilot data. To calculate the minimum sample size required to achieve adequate power in the *C. elegans* pilot data set, we first computed observed Cohen's *d* effect sizes for each variable (Equation 6), defined as the difference of the mean of the variable in the two groups divided by a pooled standard deviation.[36]

Equation 6: $d = \frac{\overline{x_1} - \overline{x_2}}{s}$, $s = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$

Here *s* represents the pooled standard deviation (Equation 6), σ_1 and σ_2 the standard deviations of the considered variable in the two groups under study, and n_1 and n_2 the number of samples per group.

The TPR and FNR were then computed for each variable in the pilot data, using the observed effect size *d*. For this *d*, we identified the smallest sample size such that the upper 95% confidence bound of the FNR was below 0.2 (corresponding power >0.8). This value was then considered as the necessary sample size, to achieve a statistical power of 0.8 and consequently identify a least one statistically significant variable at a significance threshold of 0.05 using the Benjamini-Yekutieli correction. Sample sizes determined for each variable were represented with a colour code on a typical spectral profile (Figure 3). The overall workflow is illustrated in SI Figure 1.

The code is implemented in MATLAB^{TM} (Matlab R2014a, Mathworks, Natick, MA) and freely available in an online repository (https://bitbucket.org/Gscorreia89/power-calculations).

Results

We propose a simulation approach based on pilot data to investigate power and sample size effects in untargeted metabolomic data. We first demonstrate our method using a pilot data set of 33 *C. elegans* ^{1}H NMR spectra. Later, we apply the approach to two larger data sets generated with NMR and LC-MS platforms.

Simulation of normal and log-normal data with and without correlations. We compared the simulations using multivariate normal or multivariate log-normal distributions to the pilot data set. While the mean and standard deviations of both approaches showed little difference, the skewness and kurtosis of the log-normal were better matched to those of the pilot data. A PCA analysis indicated a good fit of data simulated using both approaches to the multivariate envelope of the pilot data (SI Figure 2A). Overall, the log-normal simulation provided a better match with the long-tailed distributions typically seen in metabolic data, and were therefore used in all subsequent simulations.

Next we investigated the effect of correlation structure (i.e. the presence or absence of off-diagonal elements in the covariance matrix). Figure 1A shows a scores plot from a PCA model trained on data simulated without correlations. The pilot data, and data from a simulation with correlations are projected on to the same model. Clearly, the simulation using a non-zero correlation structure represents the pilot data more faithfully than that where correlations are not present. SI Figures 2B and 2C confirm that the correlation matrices resulting from the pilot data and the data simulated with correlations exhibited similar patterns of correlation, demonstrating that the simulation process can encapsulate the dependency structure among metabolic variables.

Simulation of different effect and sample sizes. Data sets of different sizes were constructed by randomly selecting samples from the large simulations (n=5000). For each sample and effect size, 100 data sets were sampled to allow estimation of the mean and variability of performance statistics. Both discrete (classification) and continuous (regression) outcome variables were considered. In the classification case, one of the groups was shifted relative to the other by randomly selecting a single variable and adding a given fraction of its standard deviation. For the regression case, the effect was implemented by modifying the regression coefficient of the variable under consideration. Figure 1 illustrates the discrete case, using all variables, for an effect size of 0 (Figure 1B) and 0.6 (Figure 1C). The effect is represented in a 3 dimensional PCA score plot, where the blue population is unaffected and the red is progressively shifted away. In Figure 1B, the two groups and their Hoteling's T^2 ellipses are superimposed. With an effect size of 0.6 (Figure 1C), the two data point clouds are already well separated in the PCA plot.

Power analysis across different variables. Investigating the changes induced by different sample and effect sizes throughout the data set allows the identification of potential variables of interest. Different variables will exhibit different levels of variation in the population, leading to different minimum sample sizes to detect an effect. This is illustrated for the regression case in SI Figure 3. At a sample size of 200 or effect size of 0.5, one can easily select variables presenting a strong and early decrease of the FNR for which an effect can be detected. There were clear differences across the variables in the effect size that can be detected for a given sample size. For example, at a sample size of 100, there is 0.8 power to detect effect sizes of 0.44, 0.4 and 0.38 for the metabolites valine, glyceryl of lipids and unsaturated lipids respectively.

A more comprehensive view for a single variable is given in Figure 2 where TPR and FNR are illustrated as a function of sample and effect sizes. The shape of the landscape allows the identification of the minimal sample size to detect a given effect size or the identification of the detectable effect size for a given sample size. For example, at an effect size of 0.5 and a sample size of 100, a TPR of 0.6 and FNR of 0.2 (0.8 power) is reached.

Sample size determination. To exemplify how this method can be used to determine sample size using the two classes in the pilot data, the power and sample size results were illustrated on the NMR spectrum. First, the observed effect size for each variable is measured using Cohen's *d*. We then used the simulations to identify the smallest sample size giving a FNR below 0.2. This value is color-coded on a typical NMR spectrum, as shown in Figure 3 presenting both the entire NMR spectrum and expansions of the aromatic, aliphatic and sugar regions. Panels E and F illustrate the differing sample sizes required to attain different powers for an aliphatic valine doublet (0.98-1.00ppm, Hγ). At sample size of 200 a power of 0.8 is achieved, but a sample size of 300 is required to obtain 0.95 power. We note that signals found to change in previous biological studies on this data set correspond to a low minimum sample size of around 20 samples (primarily lipid, glucose, glycerol and amino acid signals). Similar results were obtained for the regression case.

Effect of correlations and multiple testing corrections. To test the effect of correlations between variables on the power analysis, we compared the simulations using either the full covariance matrix estimated from the pilot data or a diagonal covariance matrix containing only the variances of each variable (no correlation). When variables are correlated, the effective number of independent variables is reduced, and therefore we might expect to see an increase in power. Results are presented in SI Figure 4 (A and B), for the regression case. As expected, the introduction of correlation seems to induce a gain in statistical power. However, these results have to be interpreted carefully, as in the simulation with correlations the number of positives is higher than the non-correlated case (because all variables highly correlated ($r>0.8$) to the chosen variable are considered to be positives). We also examined the effect on power of three commonly used multiple testing corrections when correlation is present. Corrections appeared in the expected order (SI Figure 4 C and D), with an increase of FNR going from no correction, through Benjamini-Hochberg correction, then Benjamini-Yekutieli correction and finally to the conservative and relatively low power Bonferroni correction.

Application to other statistical methods and data sets. To illustrate the general applicability of the method, sample size determination and power analysis were performed in several different contexts corresponding to common metabolic phenotyping situations. We examined different statistical methods (*e.g.* O-PLS) and different profiling platforms (e.g. UPLC-MS). Figure 4A shows the performance of an O-PLS model discriminating two groups for the *C. elegans* pilot data. The predictive capacity of the model, represented by the cross-validated goodness-of-fit parameter Q^2_Y, is shown in Figure 4B. Q^2_Y was negative for low sample and effect sizes and, as expected, rose with both of them. In this case, half of the variables were randomly selected to implement the effect. A Q^2 above 0.8 is reached with an effect size of 0.2 and a sample size of 200 samples. A further NMR example illustrating the case of a targeted assay is presented in Figure 4C. Here, 24 metabolites were quantified from ^{1}H NMR spectra of human spot urine samples from the ARIC cohort. The FNR is represented for each variable with respect to effect and sample size, for the regression case. Here it is easy to see that, for example, with 200 samples, one is able to detect an effect size of approximately 0.35 for most variables. Similar results are presented for a set of human sera from the AIRWAVE study profiled by UPLC-MS in Figure 4D, with a simulated effect size of 0.5, also for the regression case. No variable reaches a power of 0.8 for a sample size of 50 (red), only 54 for a sample size of 100 and all variables exhibit a power greater than 0.8 for a sample size of 250 samples. These examples exhibit the general utility of the simulation method in exploring power and sample size in many diverse situations.

Discussion

Robust methods for power analysis and sample size determination in metabolic phenotyping have been needed for many years. There are many reasons why these issues have not been tackled earlier, the main one relating to the intrinsic complexity of typical metabolic profiles, which is distinct from more conventional data. A key point is that since the approach is hypothesis free, we do not know size of the effect we are looking for or which metabolites might be affected. Secondly, the variability of metabolite levels, even in standard biofluids from normal healthy human populations, is not well characterized, let alone their levels and stability in disease states. Both of these reasons make it imperative to obtain pilot data in advance of designing larger studies. Further, each assay imprints its own statistical characteristics, such as noise levels and detection limits, on our ability to detect effects. In many situations, we detect chemically unidentified signals, while some signals may only appear in particular cohorts or pathophysiological scenarios. Finally, the relative youth of this technology has often led to the investigation of historical cohorts of samples that have been collected for other purposes such as GWAS rather than *a priori* applying rigorous experimental design appropriate for metabolic investigation.

We designed a new flexible approach to deal with power analysis in metabolic phenotyping data sets. It consists of three steps: 1) modeling the distribution of pilot study data, 2) introducing an artificial effect, and 3) deriving estimates and confidence intervals for performance metrics. In addition, confounding effects, such as batch effects, can be easily accommodated. The approach is particularly attractive when using large population studies performed with standardized protocols, to inform sample size determination in future studies.

To our knowledge, our method is the only one that models between variable correlations from pilot data. Concerning correlation between variables, high correlation seems to increase statistical power, and power is also affected when adjustment for confounders has to be performed. It is also not always clear how FWER and FDR correction methods handle the correlation in the data, and their consequences on power and FDR control. We argue that the greatest advantage of modeling correlations is the fact that now we can meaningfully benchmark and evaluate the power of multivariate methods, which are frequently used in metabolic phenotyping studies. Here, we presented a small example using a multivariate analysis method frequently used in metabolomic applications, Partial Least Squares.[37] We plan to study the issue of power analysis for commonly used multivariate techniques in a later paper.

In many situations, a high false positive rate can be more dangerous and costly than low power. Despite most of our examples focusing on power calculations or sample size determination to maximize power, other quantities and error types can be obtained equally easily with this approach. We performed the power analysis with the traditionally used NHST framework, but the methodology is not limited to classic hypothesis testing, and can be quickly adapted to perform power calculations in different ways (e.g., to derive how many samples would be necessary to estimate an effect to a desired level of precision). [38]

We envisage that statistical power in a metabolic phenotyping study depends on a combination of analytical method variability and selectivity with the underlying phenotypic variability of the population from which the study samples will be drawn. For effective power calculations, pilot studies should be designed to obtain good estimates of variable distributions and covariance structure.

The applications presented in this study are meant to illustrate the capacity of this approach to evaluate sample size and power analysis in metabolic phenotyping in different situations, from a specifically designed pilot study for a particular aim, to data reuse. The latter will be increasingly relevant, with the establishment of open data repositories like MetaboLights[39]. Figure 3 exemplifies how pilot study data variability and estimated effect sizes (based on the Cohen's d pooled standard deviation) can be used to inform study design and the difference in power and sample size requirements between variables. We provided two application examples from large human cohort studies, one profiled with NMR and the other with MS, with some approximate numbers for the amount of samples that might be needed to perform a study in a healthy free-living population.

Metabolic phenotyping has already demonstrated its potential for biomedical studies, being the cornerstone of entire research programs. Ethical and economic issues force scientists and physicians to provide reliable power analysis to secure funding. That is why suitable approaches had to be developed to address this issue. Here we provide a new, comprehensive and efficient approach to perform sample size determination and power analysis in metabolic phenotyping studies based on NMR or MS data that clearly exceeds the capacities of previously developed methods. It is clear that this approach is not restricted to metabolic phenotyping studies and is applicable for other types of data. We suggest that metabolic phenotyping study designs, particularly for grant applications, should from now on include sample size estimations based on available pilot data, to justify inclusion requirements and ensure meaningful experiments and results.

Acknowledgements

BB is supported by the Fédération pour la Recherche Médicale, the Société Française d'Anesthésie et Réanimation, the Académie Nationale de Médecine, the Association de Néonatologie de Port-Royal and the City of Lyon. GC is supported by the Imperial College Stratified Medicine Graduate Training Programme in Systems Medicine and Spectroscopic Profiling (STRATiGRAD). TE acknowledges support from the EU COSMOS project (grant agreement 312941) and the EU PhenoMeNal project (project reference: 654241). We thank Drs Bénédicte Elena-Herrmann, Jean Giacomotto, Laurent Ségalat, Pierre Toulhoat and Prof. Lyndon Emsley for providing the C. elegans data set. We would like to thank Bruker BioSpin, GmbH, Rheinstetten, Germany for metabolite quantification. This research has been conducted using the Airwave Study RTB Resource. The Airwave Health Monitoring Study is funded by the Home Office (grant number 780- TETRA) with additional support from the National Institute for Health Research (NIHR), Imperial College Healthcare NHS Trust (ICHNT) and Imperial College Biomedical Research Centre (BRC). We thank all participants in the Airwave Study for their contribution. We also thank the Institute for translational medicine and therapeutics (ITMAT).

Author contributions: BB, GC and TE designed the study, performed computational work and wrote the paper. JN and EH contributed expert advice to the project. AT and JH designed the ARIC cohort. PE and AV designed the AIRWAVE cohort. JP and ML performed the mass spectrometry analysis and preprocessing for the AIRWAVE samples.

Supporting information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Neyman, J.; Pearson, E. S. *Math. Proc. Camb. Philos. Soc.* 1933, *29* (04), 492–510.

(2) Eng, J. *Radiology* 2003, *227* (2), 309–313.

(3) Abdul Latif, L.; Daud Amadera, J. E.; Pimentel, D.; Pimentel, T.; Fregni, F. *Arch. Phys. Med. Rehabil.* 2011, *92* (2), 306–315.

(4) Ayeni, O.; Dickson, L.; Ignacy, T. A.; Thoma, A. *Plast. Reconstr. Surg.* 2012, *130* (1), 78e–86e.

(5) Van Teijlingen, E.; Hundley, V. *Nurs. Stand. R. Coll. Nurs. Gt. Br. 1987* 2002, *16* (40), 33–36.

(6) Van Teijlingen, E. R.; Rennie, A. M.; Hundley, V.; Graham, W. *J. Adv. Nurs.* 2001, *34* (3), 289–295.

(7) Lenth, R. V. *Am. Stat.* 2001, *55* (3), 187–193.

(8) Hajian-Tilaki, K. *Casp. J. Intern. Med.* 2011, *2* (4), 289–298.

(9) Wong, M. Y.; Day, N. E.; Wareham, N. J. *Stat. Med.* 1999, *18* (21), 2831–2845.

(10) Hoenig, J. M.; Heisey, D. M. *Am. Stat.* 2001, *55* (1), 19–24.

(11) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica Fate Foreign Compd. Biol. Syst.* 1999, *29* (11), 1181–1189.

(12) Nicholson, J. K.; Holmes, E.; Kinross, J. M.; Darzi, A. W.; Takats, Z.; Lindon, J. C. *Nature* 2012, *491* (7424), 384–392.

(13) Bellman, R. E. *Adaptive control processes: a guided tour*, Princeton University Press.; 1961.

(14) Cloarec, O.; Dumas, M.-E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* 2005, *77* (5), 1282–1289.

(15) Blaise, B. J.; Navratil, V.; Emsley, L.; Toulhoat, P. *J. Proteome Res.* 2011, *10* (9), 4342–4348.

(16) Abdi, H. *Encycl. Meas. Stat.* 2007.

(17) Benjamini, Y.; Hochberg, Y. *J. R. Stat. Soc. B* 1995, *57* (1), 289–300.

(18) Benjamini, Y.; Yekutieli, D. *Ann. Stat.* 2001, *29* (4), 1165–1188.

(19) Ferreira,, J. A.; Zwinderman, A. *Int. J. Biostat.* 2006, *5* (1).

(20) Blaise, B. J. *Anal. Chem.* 2013.

(21) Nyamundanda, G.; Gormley, I. C.; Fan, Y.; Gallagher, W. M.; Brennan, L. *BMC Bioinformatics* 2013, *14*, 338–345.

(22) Jung, S.-H.; Young, S. S. *J. Biopharm. Stat.* 2012, *22* (1), 30–42.

(23) Trutschel, D.; Schmidt, S.; Grosse, I.; Neumann, S. *Metabolomics* 2014, *11* (4), 851–860.

(24) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. *Nucleic Acids Res.* 2015, *43* (W1), W251–257.

(25) Van Iterson, M.; 't Hoen, P.; Pedotti, P.; Hooiveld, G.; den Dunnen, J.; van Ommen, G.; Boer, J.; Menezes, R. *BMC Genomics* 2009, *10* (1), 439.

(26) Blaise, B. J.; Giacomotto, J.; Elena, B.; Dumas, M.-E.; Toulhoat, P.; Segalat, L.; Emsley, L. *Proc. Natl. Acad. Sci.* 2007, *104* (50), 19808–19812.

(27) Blaise, B. J.; Giacomotto, J.; Triba, M. N.; Toulhoat, P.; Piotto, M.; Emsley, L.; Ségalat, L.; Dumas, M.-E.; Elena, B. *J. Proteome Res.* 2009, *8* (5), 2542–2550.

(28) *Am. J. Epidemiol.* 1989, *129* (4), 687–702.

(29) Dona, A. C.; Jiménez, B.; Schäfer, H.; Humpfer, E.; Spraul, M.; Lewis, M. R.; Pearce, J. T. M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* 2014, *86* (19), 9887–9894.

(30) Elliott, P.; Vergnaud, A.-C.; Singh, D.; Neasham, D.; Spear, J.; Heard, A. *Environ.*

Res. 2014, *134*, 280–285.

(31) Blaise, B. J.; Shintu, L.; Elena, B.; Emsley, L.; Dumas, M.-E.; Toulhoat, P. *Anal. Chem.* 2009, *81* (15), 6242–6251.

(32) Navratil, V.; Pontoizeau, C.; Billoir, E.; Blaise, B. J. *Bioinforma. Oxf. Engl.* 2013.

(33) Stehman, S. V. *Remote Sens. Environ.* 1997, *62* (1), 77–89.

(34) Trygg, J.; Wold, S. *J. Chemom.* 2002, *16* (3), 119–128.

(35) Efron, B. *Ann. Stat.* 1979, *7* (1), 1–26.

(36) Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition.; Routledge: Hillsdale, N.J, 1988.

(37) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* 2001, *58* (2), 109–130.

(38) Halsey, L. G.; Curran-Everett, D.; Vowler, S. L.; Drummond, G. B. *Nat. Methods* 2015, *12* (3), 179–185.

(39) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data http://nar.oxfordjournals.org/content/early/2012/10/28/nar.gks1004.full (accessed Jul 15, 2015).

Figure 1. Sample and effect size implementation. Principal component analysis score plot (A) trained on samples simulated from the pilot data using a multivariate normal without correlation (blue circles) with the 95% Hotelling's T^2 ellipse. Also shown is the projection of data simulated using a multivariate log-normal with correlation (green triangles) and the pilot data (red squares). (B) & (C) Principal component analysis score plots for a discrete effect (two groups, red and blue). Simulations are based on a random multivariate normal including correlations, using the mean and covariance of the pilot data set. Examples are given for an effect size of 0 (B) and 0.6 (C) implemented on all variables. The 95% Hotelling's T^2 ellipsoids for each group are represented in grey.

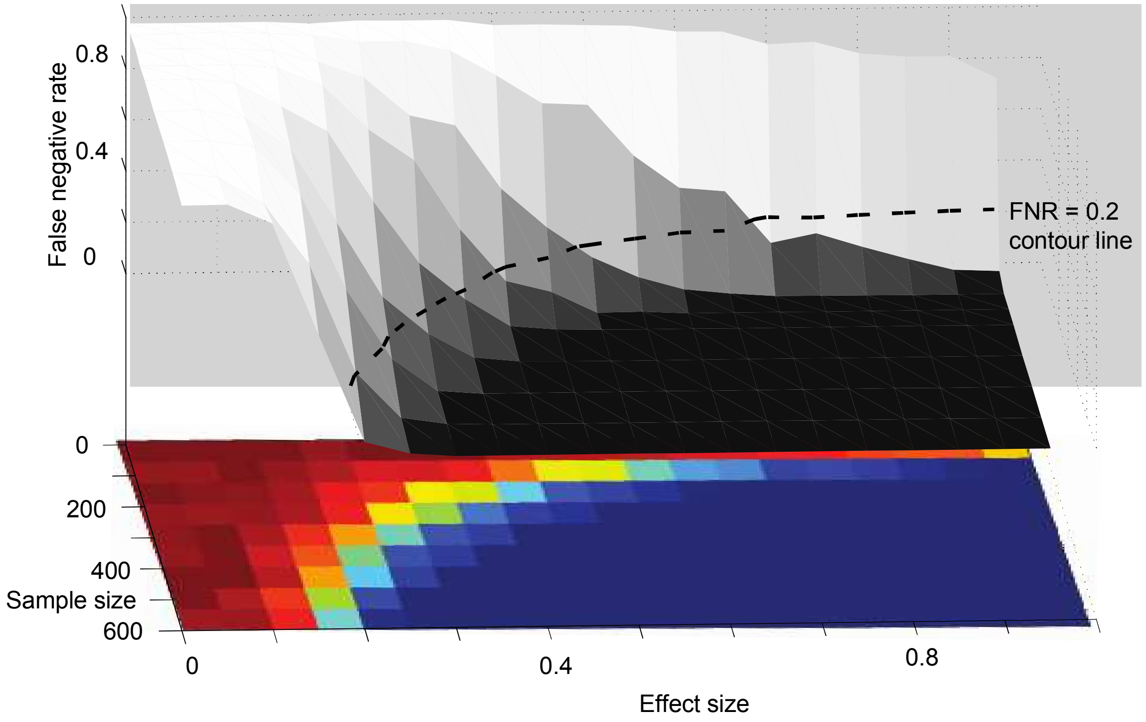
Figure 2. Investigation of effect and sample size for one variable for the regression case. Significance testing is performed with the Benjamini-Yekutieli correction. The greyscale surface represents the mean true positive (A) and false negative (B) rate for a valine signal (2.22-2.23ppm) as a function of sample and effect size. The heat map at the bottom indicates the extent of the 95% confidence interval: dark blue indicates a confidence interval intersecting zero; the further the confidence interval is from zero the warmer the colors, the reddest color representing the longest distance of the lower bound on the 95% confidence interval from zero. The dotted line in B represents the false negative rate contour at 0.2, which corresponds to a power of 0.8.

Figure 3. Sample size determination across the NMR metabolic profile for two group discrimination. Significance testing is performed with the Benjamini-Yekutieli correction. The color code shows the necessary sample sizes for each variable to achieve a false negative rate of 0.2 (corresponding to a power of 0.8) Lowest sample sizes are represented in warm colors and plotted on a typical NMR spectrum (A). Expansions of different areas of the spectrum (B: aromatic area, C: aliphatic and sugar area, D: aliphatic area). Expansions of the aliphatic areas with a power of 0.8 (E) or 0.95 (F). Note the change of color from orange to green of the doublet at 0.98ppm. Grey colors represent areas of the spectrum not selected by the SRV binning algorithm.

Figure 4. Applications in diverse settings. (A and B) O-PLS false negative rate (A) and Q^2_Y values (B) as a function of sample and effect sizes for a two-group classification on the pilot

dataset. (C) False negative rates for 200 samples (top) or an effect size of 0.5 (bottom) for the 24 quantified metabolic variables of the ARIC data, based on ^{1}H NMR analysis of human spot urine samples, in the regression case. (D) Overview of the effect of sample size on power for an effect size of 0.5, on the 7585 variables obtained by ultra performance liquid chromatography – mass spectrometry analysis of human serum samples from a healthy human cohort (AIRWAVE), in the regression case.

TOC Graphic


False negative rate
0.8
0.4
0
FNR = 0.2
contour line
0
200
400
Sample size
600
0
0.4
0.8
Effect size

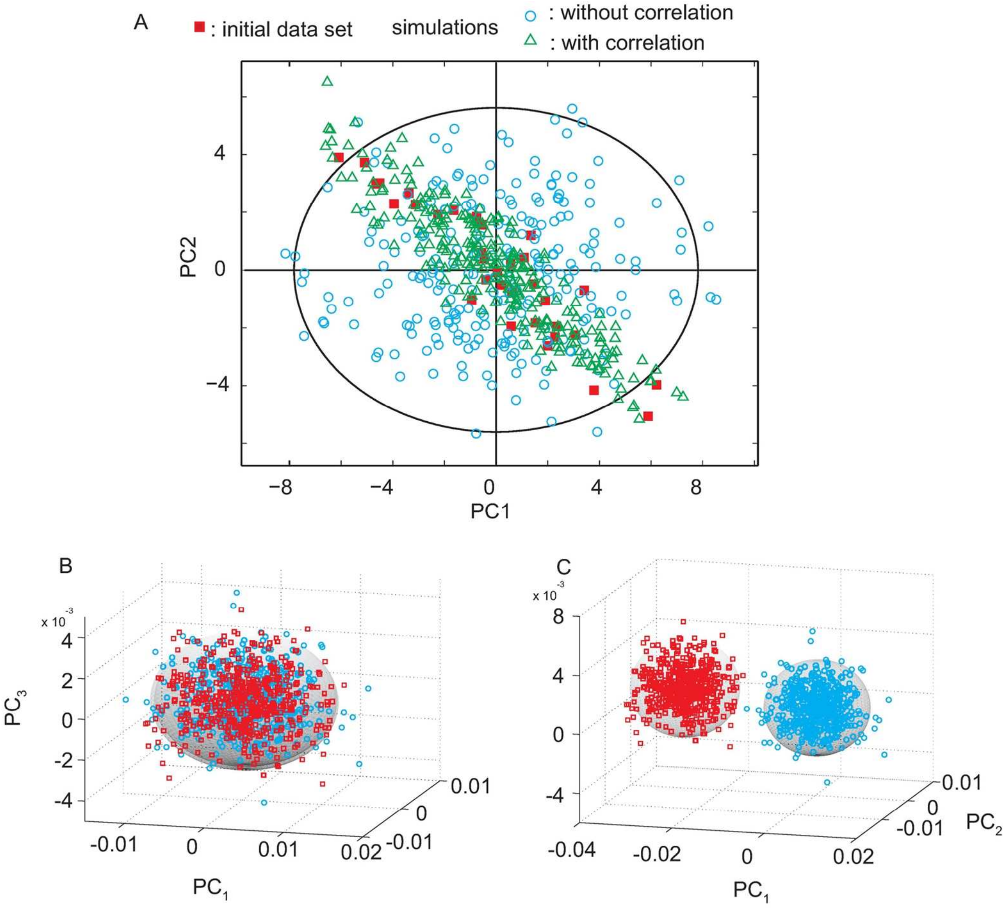


Figure 1
150x136mm (300 x 300 DPI)

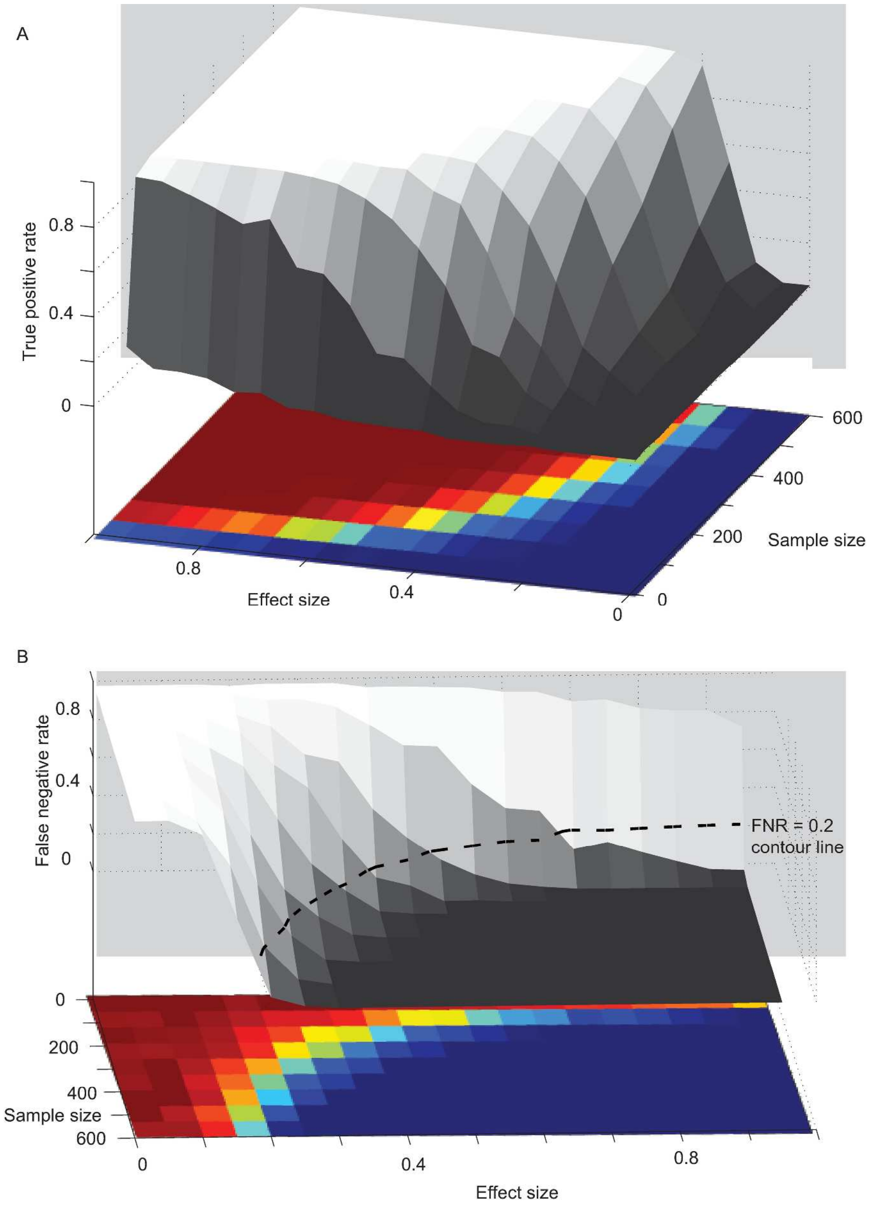


Figure 2
281x391mm (300 x 300 DPI)

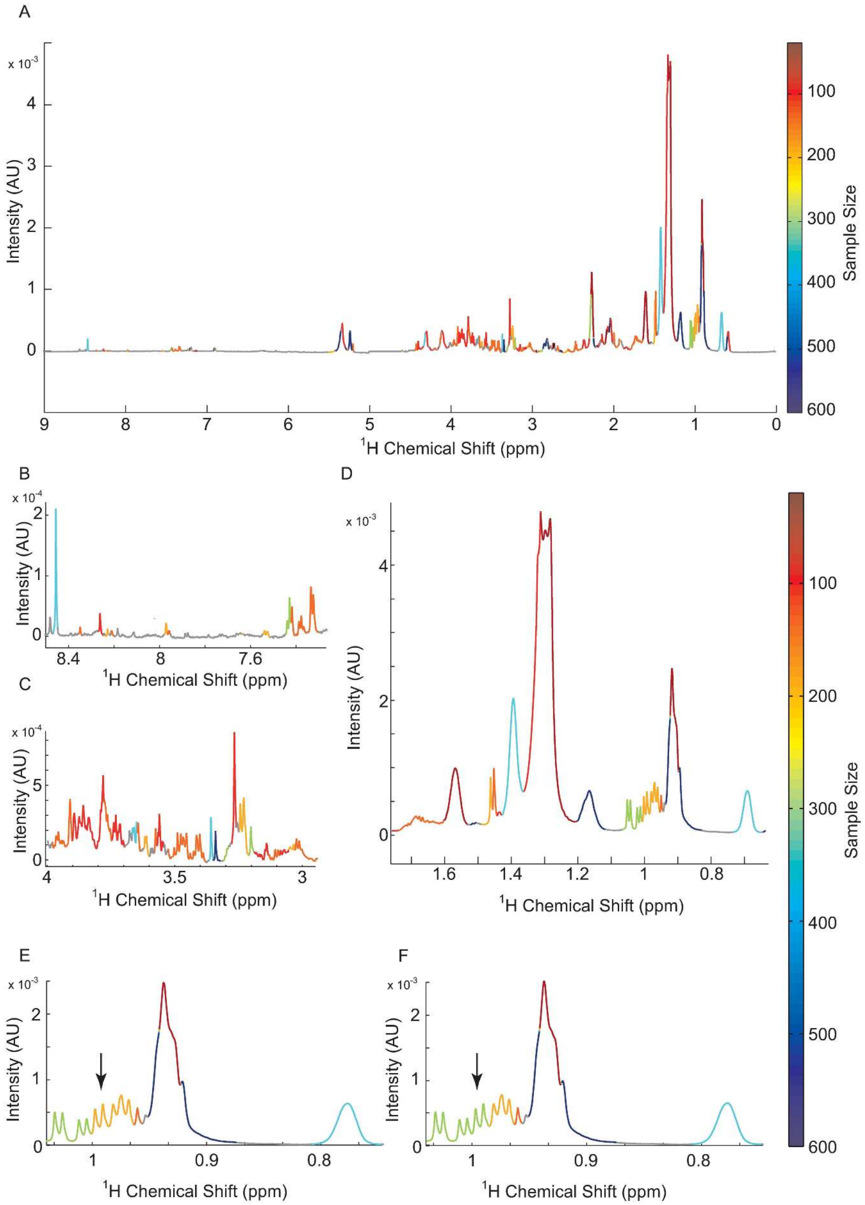


Figure 3
293x409mm (300 x 300 DPI)

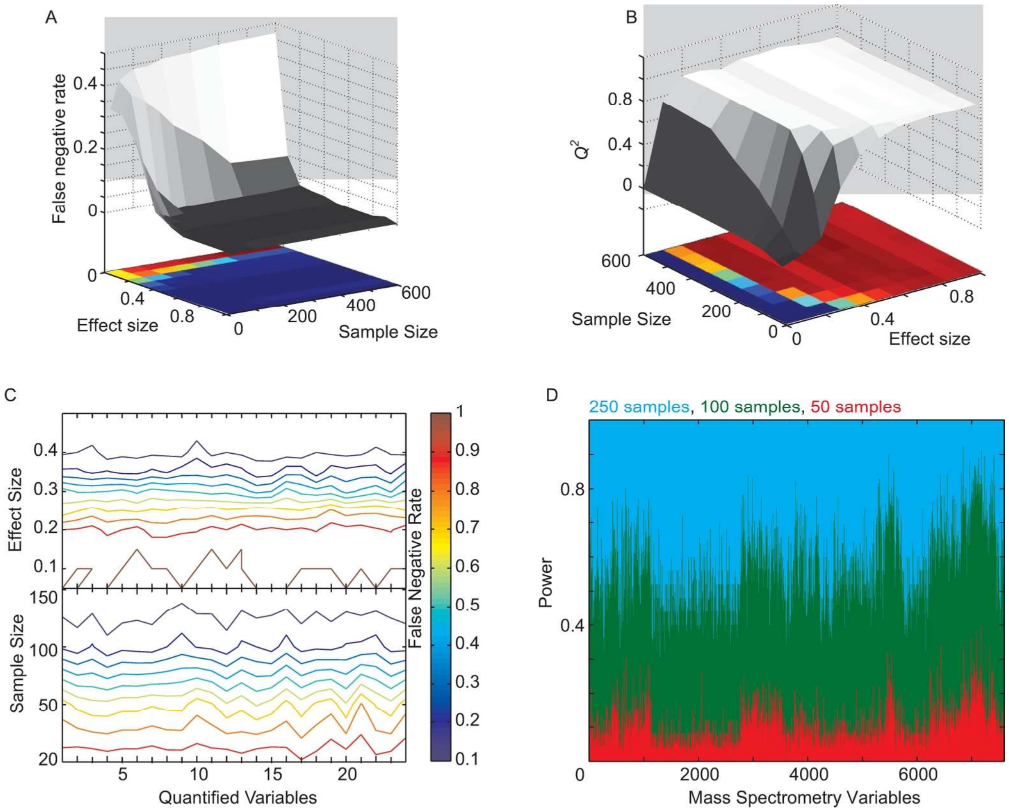
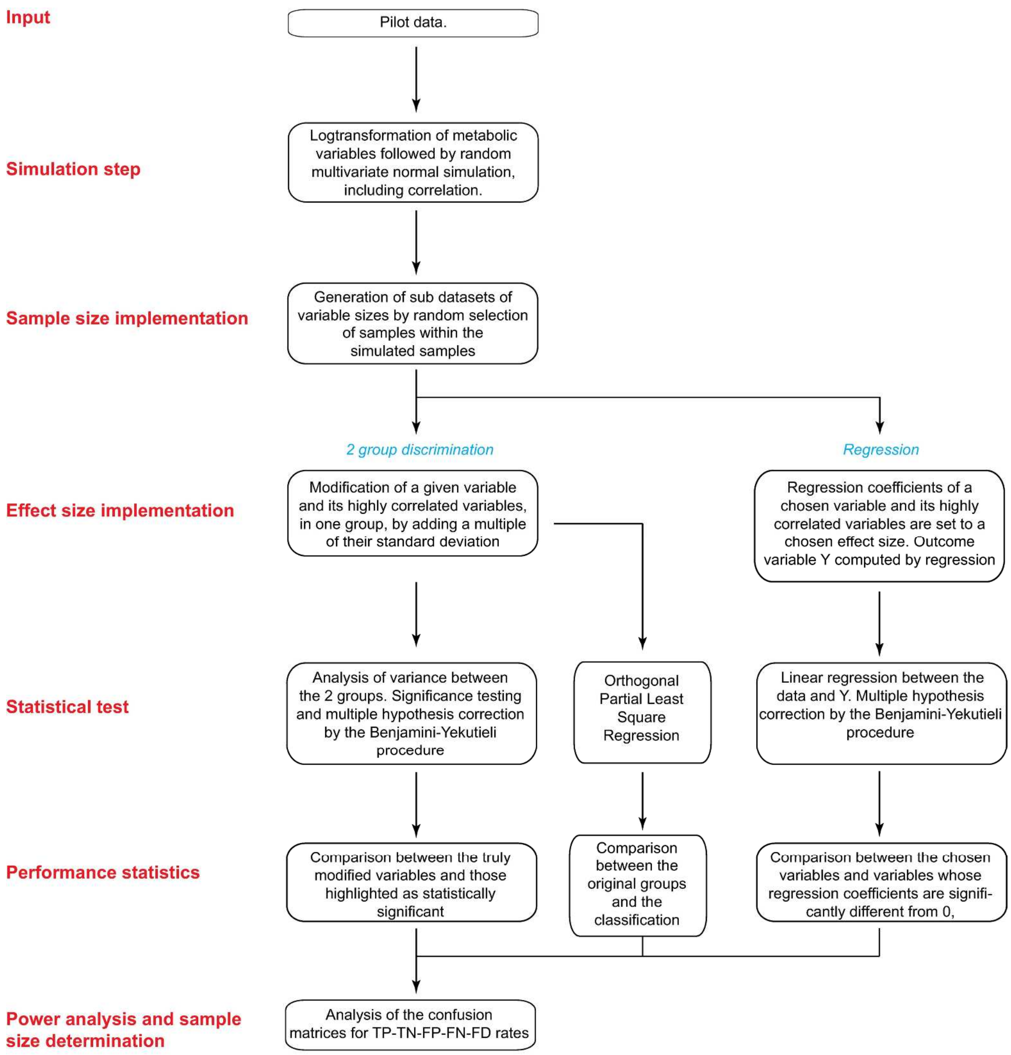
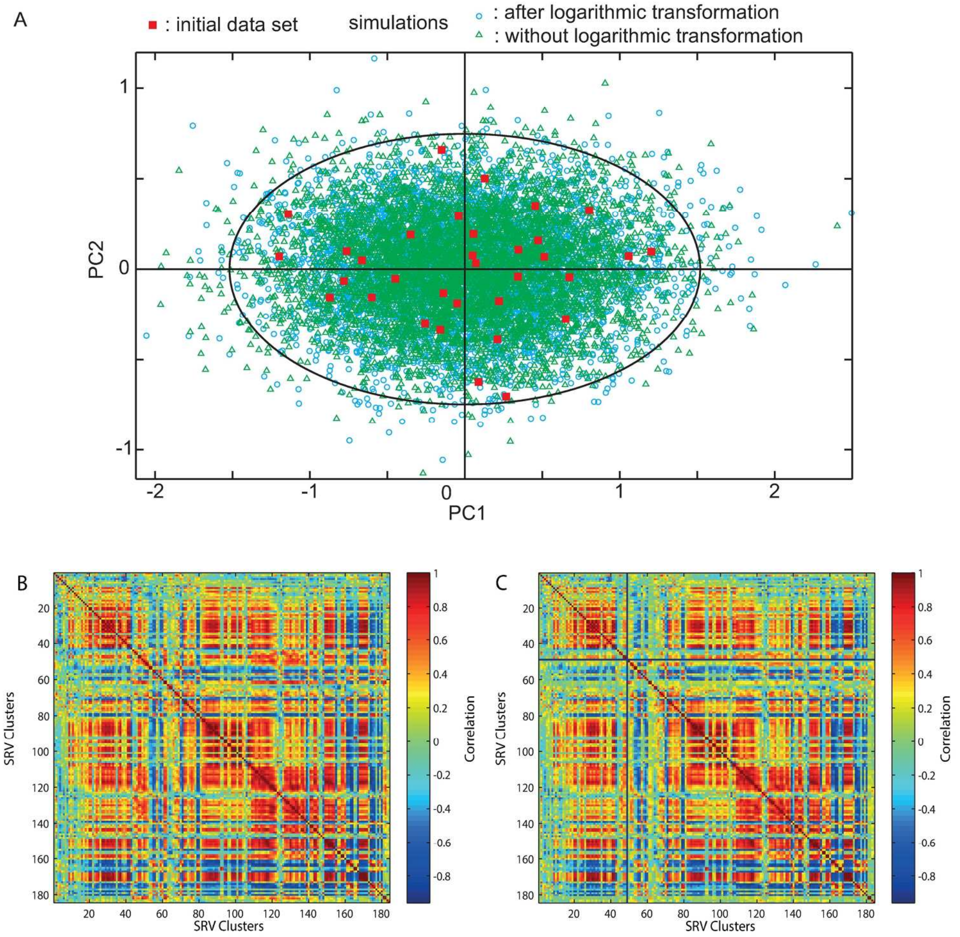


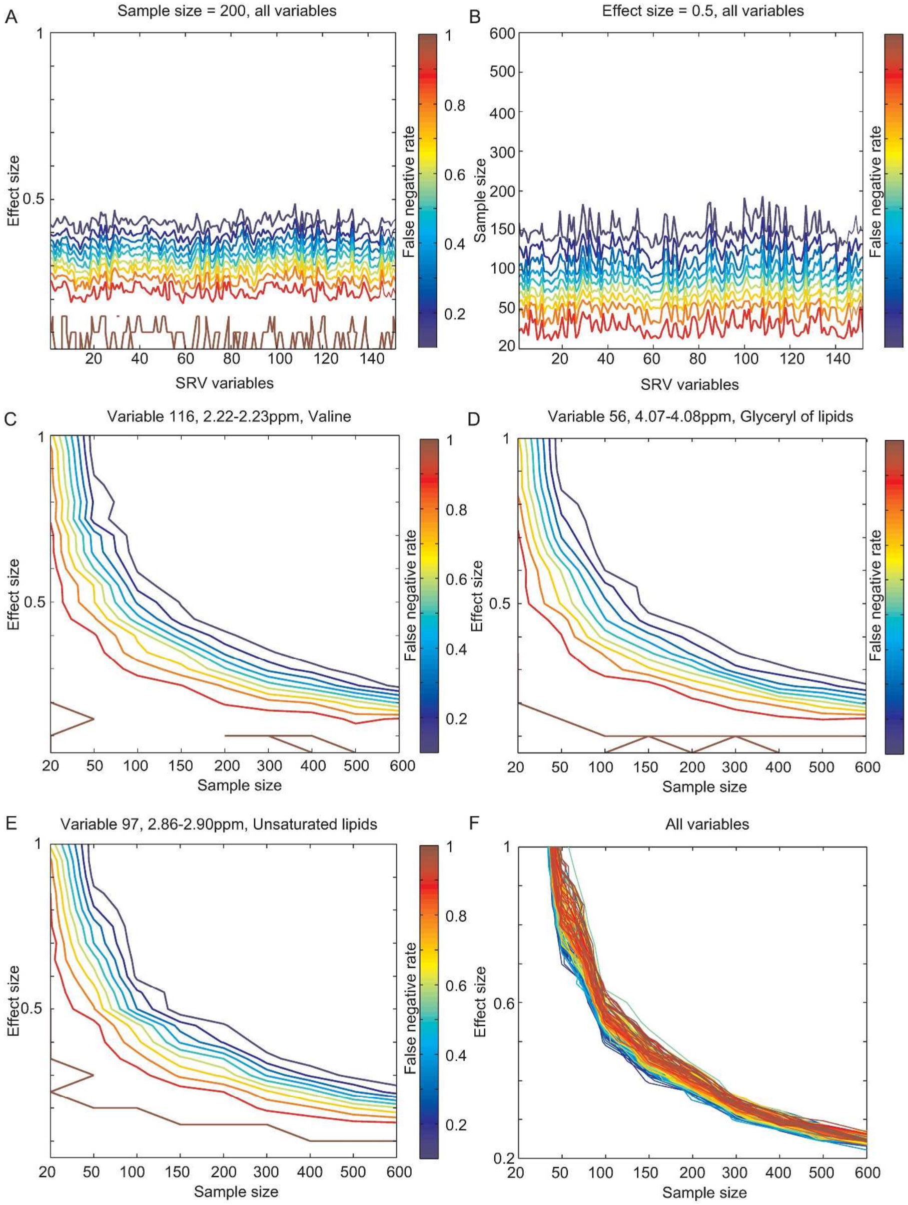
Figure 4
191x153mm (300 x 300 DPI)


Input
Pilot data.
Simulation step
Logtransformation of metabolic variables followed by random multivariate normal simulation, including correlation.
Sample size implementation
Generation of sub datasets of variable sizes by random selection of samples within the simulated samples
2 group discrimination
Regression
Effect size implementation
Modification of a given variable and its highly correlated variables, in one group, by adding a multiple of their standard deviation
Regression coefficients of a chosen variable and its highly correlated variables are set to a chosen effect size. Outcome variable Y computed by regression
Statistical test
Analysis of variance between the 2 groups. Significance testing and multiple hypothesis correction by the Benjamini-Yekutieli procedure
Orthogonal Partial Least Square Regression
Linear regression between the data and Y. Multiple hypothesis correction by the Benjamini-Yekutieli procedure
Performance statistics
Comparison between the truly modified variables and those highlighted as statistically significant
Comparison between the original groups and the classification
Comparison between the chosen variables and variables whose regression coefficients are significantly different from 0,
Power analysis and sample size determination
Analysis of the confusion matrices for TP-TN-FP-FN-FD rates

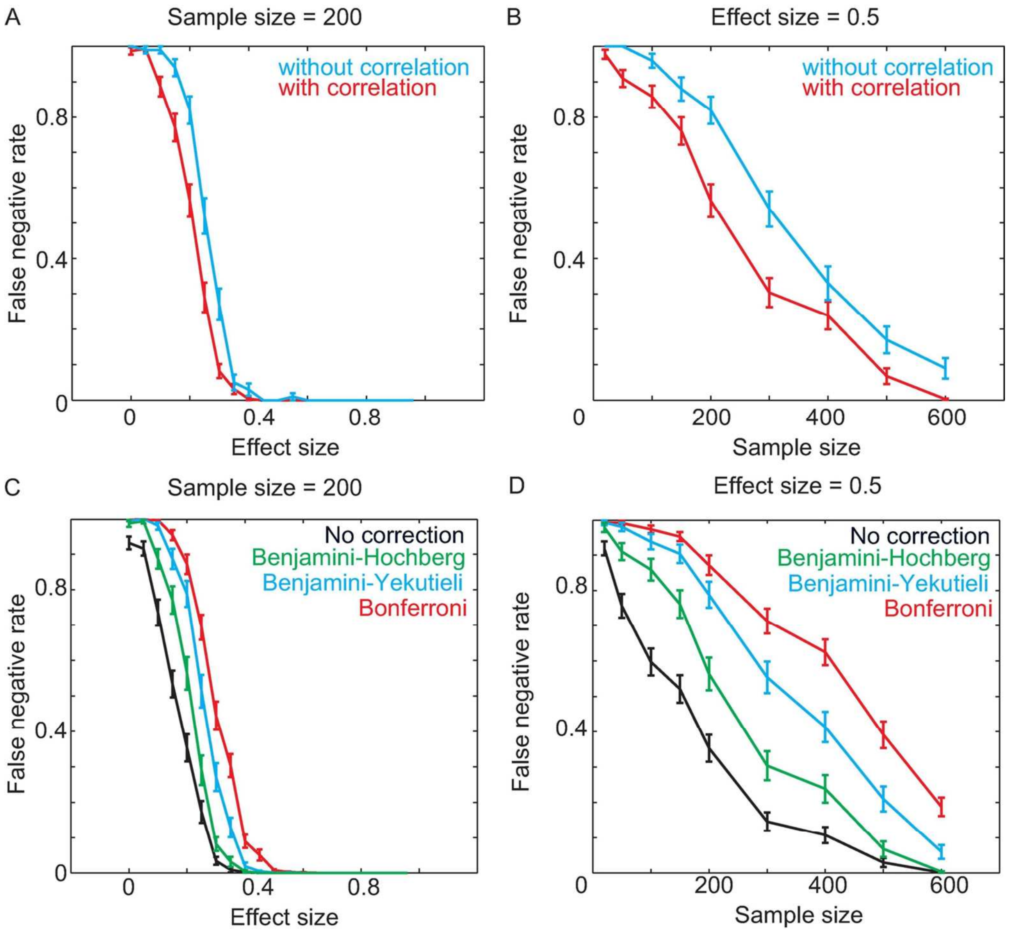
Supplementary figure 1
246x260mm (300 x 300 DPI)



Supplementary figure 2
200x190mm (300 x 300 DPI)



Supplementary figure 3
271x350mm (300 x 300 DPI)



Supplementary figure 4
138x128mm (300 x 300 DPI)