Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2006

Impact Factor: 5.64 · DOI: 10.1021/ac051495j · Source: PubMed

Sabina Bijlsma
TNO
28 PUBLICATIONS 1,219 CITATIONS

Elwin Verheij
TNO
66 PUBLICATIONS 2,440 CITATIONS

Sunil Kochhar
Nestlé S.A.
119 PUBLICATIONS 4,706 CITATIONS

Ian Macdonald
University of Nottingham
392 PUBLICATIONS 10,247 CITATIONS

Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation

Sabina Bijlsma,*,† Ivana Bobeldijk,† Elwin R. Verheij,† Raymond Ramaker,† Sunil Kochhar,‡ Ian A. Macdonald,§ Ben van Ommen,∥ and Age K. Smilde†

Business Unit Analytical Sciences and Business Unit Physiological Sciences, TNO Quality of Life, P.O. Box 360, 3700 AJ Zeist, The Netherlands, BioAnalytical Science Department, Nestlé Research Center, P.O. Box 44, CH-1000 Lausanne 26, Switzerland, and School of Biomedical Sciences, University of Nottingham Medical School, Queen's Medical Centre, Clifton Boulevard, Nottingham, NG7 2UH, United Kingdom

A large metabolomics study was performed on 600 plasma samples taken at four time points before and after a single intake of a high fat test meal by obese and lean subjects. All samples were analyzed by a liquid chromatography–mass spectrometry (LC–MS) lipidomic method for metabolic profiling. A pragmatic approach combining several well-established statistical methods was developed for processing this large data set in order to detect small differences in metabolic profiles in combination with a large biological variation. Such metabolomics studies require a careful analytical and statistical protocol. The strategy included data preprocessing, data analysis, and validation of statistical models. After several data preprocessing steps, partial least-squares discriminant analysis (PLS-DA) was used for finding biomarkers. To validate the found biomarkers statistically, the PLS-DA models were validated by means of a permutation test, biomarker models, and noninformative models. Univariate plots of potential biomarkers were used to obtain insight in up- or downregulation. The strategy proposed proved to be applicable for dealing with large-scale human metabolomics studies.

Metabolites are the end product of cellular regulatory and metabolic processes, and their levels can be regarded as the response of biological systems to environmental changes. Metabolomics is a nontargeted analysis of low molecular mass endogenous and exogenous (e.g., nutrients) metabolites. In the paper of Fiehn,[1] a summary is given of all the different terms that are used in the literature for measuring metabolites in a system. Most of the metabolomics work described in the literature was done using nuclear magnetic resonance (NMR). However, in the past few years MS-based techniques have been used more frequently.[2–11] Metabolomics data obtained by NMR or liquid chromatography–mass spectrometry (LC–MS) are processed using multivariate statistical methods, mostly principal component analysis,[12–14] to identify the differences between the studied test groups. This new area of research has found applications in pharmaceutical research, where it is used for discovery of biomarkers for different diseases, safety markers, or mechanistic research in drug development.

Most examples described in the metabolomics literature are with animal systems such as rat or mouse. In these examples, there is relatively little biological variation (inbred animals with standardized environment), and in combination with the area of research, such as disease, study of toxicity, or drug treatment, drastic differences are typically observed between the groups. The studies can therefore be limited to a relatively low number of subjects (typically 5–10) per group. If the number of samples is limited and if high-throughput or semi-high-throughput methods are used, all samples can be analyzed within one batch. Normalization, calibration, and scaling of the data are less pronounced.

* To whom correspondence should be addressed. Phone: +31 30 694 4156. Fax: +31 30 694 4894. E-mail: bijlsma@voeding.tno.nl.

† Business Unit Analytical Sciences, TNO Quality of Life.

‡ Nestlé Research Center.

§ University of Nottingham Medical School.

∥ Business Unit Physiological Sciences, TNO Quality of Life.

[1] Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155–171.

[2] Lenz, E. M.; Bright, J.; Knight, R.; Wilson, I. D.; Major, H. *J. Pharm. Biomed. Anal.* **2004**, *35*, 599–608.

[3] Lenz, E. M.; Bright, J.; Knight, R.; Wilson, I. D.; Major, H. *Analyst* **2004**, *129*, 535–541.

[4] Idborg-Björkman, H.; Edlund, P. O.; Kvalheim, O. M.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2003**, *75*, 4784–4792.

[5] Lafaye, A.; Junot, C.; Ramounet-Le Gall, B.; Fritsch, P.; Tabet, J. C.; Ezan, E. *Rapid Commun. Anal. Chem.* **2003**, *17*, 2541–2549.

[6] Plumb, R.; Granger, J.; Stumpf, C.; Wilson, I. D.; Evans, J. A.; Lenz, E. M. *Analyst* **2003**, *128*, 819–823.

[7] Plumb, R. S.; Stumpf, C. L.; Gorenstein, M. V.; Castro-Perez, J. M.; Dear, G. J.; Anthony, M.; Sweatman, B. C.; Connor, S. C.; Haselden, J. N. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1991–1996.

[8] Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *11*, 1181–1189.

[9] Clish, C. B.; Davidov, E.; Oresic, M.; Plasterer, T. N.; Lavine, G.; Londo, T.; Meys, M.; Snell, P.; Stochaj, W.; Adourian, A.; Zhang, X.; Morel, N.; Neumann, E.; Verheij, E.; Vogels, J. T. W. E.; Havekes, L. M.; Afeyan, N.; Regnier, F.; van der Greef, J.; Naylor, S. *OMICS* **2004**, *8*, 3–13.

[10] van der Greef, J.; Davidov, E.; Verheij, E.; Vogels, J.; van der Heijden, R.; Adourian, A. S.; Oresic, M.; Marple, E. W.; Naylor, S. In *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*; Harrigan, G. G., Goodacre, R., Eds.; Springer: New York, 2003; Chapter 10.

[11] Wilson, I. D.; Plumb, R.; Granger, J.; Major, H.; Williams, R.; Lenz, E. M. *J. Chromatogr., B* **2005**, *817*, 67–76.

[12] Martens H.; Naes, T. *Multivariate Calibration*; Wiley: Chichester, 1989.

[13] Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; De Jong, S.; Lewi P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B*; Elsevier: Amsterdam, 1998.

[14] Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997.

10.1021/ac051495j CCC: $33.50 © 2006 American Chemical Society
Published on Web 12/14/2005

Moreover, little literature is available where models based on metabolomic data were statistically validated.[15,16] However, in the paper by Idborg et al.,[17] a valid strategy is presented for data preprocessing, data analysis, and validation of metabolomics studies. This strategy is valid only for small-scale metabolomics studies and not generally applicable. In the review by Wilson et al.,[11] they discussed that care should be taken when identifying markers from data analysis. The following problems may occur: (i) Samples cannot be analyzed within one batch. This can lead to "batch-to-batch" differences in the data. (ii) Identification of "markers" that are not associated with the toxic condition, disease, or effect studied (drug metabolites or system peaks). (iii) Statistical overfitting of the models resulting in chance correlations (false positives).

In human nutrition studies, high biological variation and less pronounced effect between the studied groups with or without intervention are expected. Studies usually involve many subjects, in most cases combined with a number of time points for each subject. Processing of data from hundreds of samples requires a different approach and more extensive validation than described in most metabolomic papers published to date.

In cooperation with Nestlé Research Center (Lausanne, Switzerland) and the EU NUGENOB project TNO performed a large metabolomics study involving plasma from 150 human subjects at 4 time points before and after single intake of a high-fat test meal.[18] (NUGENOB is the acronym of the project "Nutrient–Gene interactions in human obesity—implications for dietary guidelines" supported by the European Community (contract QLK1-CT-2000-00618; see the website www.nugenob.org.) These subjects have been examined by a scrutiny of dietary lifestyles, a 1-day clinical investigation program including a single intake of a high-fat test meal followed by a 10-week hypocaloric intervention with either high- or low-fat content. The main objective of this multicenter human obesity project was to elucidate the role of interactions between macronutrient composition of the diet with particular emphasis on fat intake and specific genetic variants. One of the aims of the metabolomics study was to find differences in the plasma metabolome (metabolites in plasma) between obese and lean subjects at fasting conditions and after the intake of a high-fat test meal. The goal was to analyze as many metabolites as possible. Therefore four analytical methods were used: LC–MS lipids, LC–MS polar compounds, gas chromatography–mass spectrometry (GC/MS), and NMR.

This paper describes a pragmatic approach that was developed for dealing with these types of large and complex data sets. Attention was paid to preprocessing and calculation of the multivariate models as well as to the "statistical" validation of the models and biomarkers. The developed strategy is demonstrated using LC–MS lipidomic data. An extensive external validation of all individual steps of the strategy has not been performed. Most steps are well-established in the literature. The application of biomarker and noninformative models is new. We have done some validation on 80% rule and imputation of missings, for example, but these results are not shown in this paper. Even after applying the strategy, the results need to be biologically evaluated. The validation of potential biomarkers should take place in a separate study, focused on targeted analysis of a few potential biomarkers with specific methods. Biological interpretation and validation of the potential biomarkers as well as a comparison of all results from different analytical methods (LC–MS lipids, LC–MS polar compounds, GC/MS, NMR) were outside the scope of this paper and will be described elsewhere.

MATERIALS AND METHODS

Overview of the Study. The study involved a multicenter investigation of gene nutrient interactions in eight different centers around Europe in which more than 700 obese subjects were studied before and after a 10-week weight loss program, with baseline values being compared to a group of more than 100 lean subjects.[18] Before the weight loss program, the obese subjects and the lean subjects consumed a high-fat test meal. Blood samples were taken into ethylenediaminetetraacetic acid before and at hourly intervals for 3 h after the test meal. Four plasma samples from each of the 150 selected subjects (100 obese and 50 lean subjects, 600 samples in total) were used in the metabolomics analysis. The eight clinical centers were similarly represented in the obese and lean groups.

Quality Control Samples. Quality control samples were prepared by pooling plasma samples from this study for the purpose of assessing the "batch-to-batch" (see also Sample Sequence and Sample Batches) variability of the analytical methods. Two samples were prepared from plasma of obese subjects (mixture of time points), and one sample was prepared from plasma of lean subjects. The samples were divided into aliquots and stored at −20 °C until analysis.

Sample Preparation. A 10-μL sample of blood plasma was placed in a vial and stored at −20 °C until analysis. The samples were allowed to thaw prior to analysis, and 200 μL of IPA containing three internal standards (C17:0 lysophophatidylcholine (LPC) 1 μg/mL, C24:0 phosphatidylcholine (PC) 1 μg/mL, and C51:0 triglyceride (TG) 2 μg/mL). The samples were centrifuged for 3–5 min at 10 000 rpm in order to remove the precipitated proteins, and the clear extract was transferred into an autosampler vial.

LC–MS. The analysis was performed on a LC–MS system consisting of a Waters (Etten-Leur, The Netherlands) HPLC 600 MS pump equipped with a 717 autosampler and a 600S system controller. A 10-μL sample of extract was injected, and the compounds from different lipid classes were separated based on their polarity (very lipophilic compounds elute last) ionized with electrospray ionization in the positive mode. The detection of the compounds was performed in the full scan mode with a ThermoFinnigan (Breda, The Netherlands) TSQ 700 LX triple quadrupole mass spectrometer, equipped with the API 2 interface. For data acquisition, Excalibur V1.4 was used. Different classes of lipids are analyzed: PCs, LPCs, diglycerides, TGs, sphingomyelins

[15] Beckonert, O.; Bollard, M. E.; Ebbels, T. M. D.; Keun, H. C.; Antti, H.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chim. Acta* **2003**, *490*, 3–15.

[16] Jonsson, P.; Bruce, S. J.; Moritz, T.; Trygg, J.; Sjöström, M.; Plumb, R.; Granger, J.; Maibaum, E.; Nicholson, J. K.; Holmes, E.; Antti, H. *Analyst* **2005**, *130*, 701–707.

[17] Idborg, H.; Zamani, L.; Edlund, P.-O.; Schuppe-Koistinen, I.; Jacobsson, S. P. *J. Chromatogr., B*, in press.

[18] Petersen, M.; Taylor, M. A.; Saris, W. H. M.; Verdich, C.; Toubro, S.; Macdonald, I.; Rössner, S.; Stich, V.; Guy-Grand, B.; Langin, D.; Martinez, A. J.; Pedersen, O.; Holst, C.; Sørensen, T. I. A.; Astrup, A. The NUGENOB Consortium. *Int. J. Obesity* **2006**, in press.

[19] Verhoeckx, K. C.; Bijlsma, S.; Jespersen, S.; Ramaker, R.; Verheij, E. R.; Witkamp, R. F.; van der Greef, J.; Rodenburg, R. J. *Int. Immunopharmacol.* **2004**, *4*, 1499–1514.

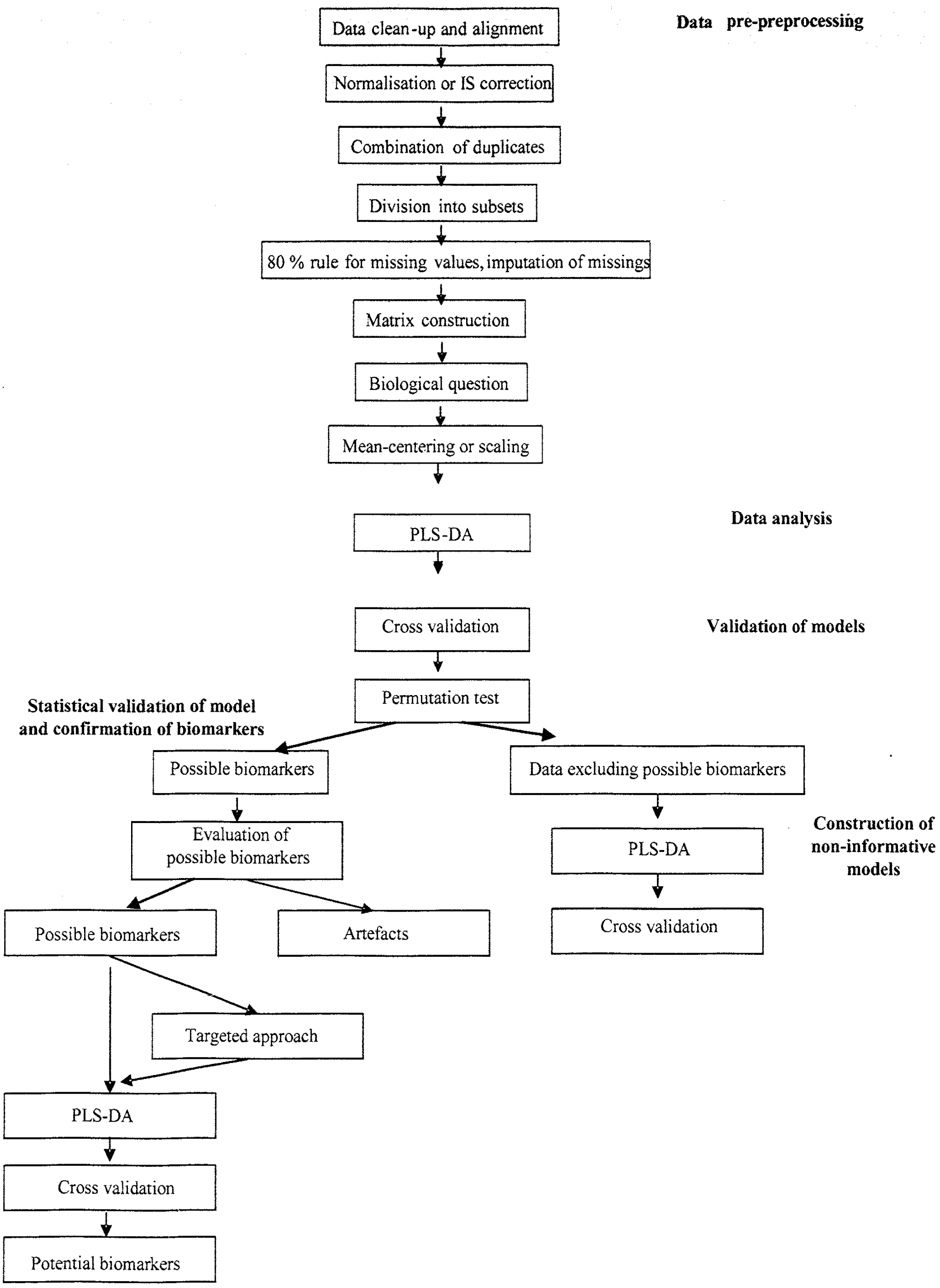
Data clean-up and alignment
Data pre-preprocessing
Normalisation or IS correction
Combination of duplicates
Division into subsets
80 % rule for missing values, imputation of missings
Matrix construction
Biological question
Mean-centering or scaling
PLS-DA
Data analysis
Cross validation
Validation of models
Permutation test
Statistical validation of model and confirmation of biomarkers
Possible biomarkers
Data excluding possible biomarkers
Evaluation of possible biomarkers
PLS-DA
Construction of non-informative models
Possible biomarkers
Artefacts
Cross validation
Targeted approach
PLS-DA
Cross validation
Potential biomarkers

Figure 1. Summary of the strategy for data preprocessing, data analysis, and model validation presented in this paper.

(SPMs), and cholesterol esters (ChEs). Free fatty acids and free cholesterol are not detected using this method. The LC−MS method used has been described elsewhere.[19]

Tentative identification of relevant compounds was performed using the data acquired as described above and results from other published and unpublished related metabolomics studies. For the identification, several chromatograms were evaluated. The chromatograms were selected in such a way that both groups of subjects and all time points were represented. Identification of lipid class is based on the retention time and *m*/*z* (mass to charge) ratio.

Identification within the lipid classes (in a specific retention time area) is based on the observed *m*/*z* and relative retention times of specific *m*/*z* peaks within one lipid class. The retention time of lipids within one class is dependent on the number of carbon atoms and the number of double bonds. Each double bond reduces the retention time if the same number of carbons is involved. Within the LC−MS lipid platform, it is not possible to distinguish between the different isomers of, for example, C54:7 triglyceride. This can only be achieved with additional techniques and methods.

Sample Sequence and Sample Batches. The 600 samples were divided into 8 batches, of which 7 batches contained 72 samples and 1 batch contained 96 samples. In each analysis batch, both groups (lean and obese) were represented proportionally. If mean-centering per person is applied, both the large biological variation between the subjects and a possible "batch offset" will be removed. In addition to the samples of interest, in each batch, three quality control (QC) samples were analyzed. Within a batch, the samples were randomized. Each sequence was measured in duplicate. After the first injection of the complete batch, a second injection was performed in the same order. Each analysis (injection) takes 35 min; all 1248 injections (1200 sample injections and 48 QC injections) were performed without cleaning the instrument or changing the HPLC column.

Software. Data (pre-) processing, data analysis, and statistical model validation were performed in the Matlab environment (Matlab version 6.5.1, Release 13, The Mathworks, 2003 and the PLS Toolbox, version 3.0.2, 2003). Impress V 1.10 was used for peak picking.[20] Winlin V 2.3 was used for fine alignment and cleanup of the data.[21] For the targeted method explained in the next section, LC-Quan V 1.4 (ThermoFinnigan) was used.

STRATEGY FOR DATA TREATMENT

The strategy applied for data preprocessing, data analysis, and validation is explained in detail in this section. The power of this strategy is demonstrated with the LC–MS lipidomic data set. Some results obtained are shown in the next section. A flow scheme of the strategy including all individual steps is shown in Figure 1.

Terminology Used in This Paper. *Possible biomarkers:* identified or unidentified compounds represented by variables and brought forward by the first data analysis that have not been statistically validated. *Potential biomarkers:* identified or unidentified compounds represented by variables and brought forward by the first data analysis that have been statistically validated by the described strategy. *Biomarker profiles:* set of possible or potential biomarkers as brought forward by multivariate data analysis. *Kinetic profiles:* profiles of all analyzed metabolites for each subject in time.

Data preprocessing. (a) Assessment of the Data Quality. The quality of the data was assessed by selecting a number of representative compounds for each lipid class analyzed with this method as target compounds and calculating their concentrations (relative to the appropriate internal standard; see further in text) in the QC samples. The relative standard deviations (RSD) were calculated for each compound. The RSDs are <6% for all LPCs, <15% for all PCs, <25% for all TGs, <15% for all SPMs, and < 23% for ChEs. However, the higher RSDs for ChE need to be taken into consideration when interpreting the results of the data anlysis. Based on plots (not shown) of the relative concentrations of representative compounds in the QC samples plotted against time, it could be concluded that the remaining variation after correction for the internal standards was random and not a gradual decrease or increase in time (no trends observed). Based on these results, it was concluded that the quality of the data was good.

(b) Peak Picking, Integration, Alignment, and Data Cleanup. The peak picking with Impress software results in a list of peaks (variables) characterized by m/z and retention time in minutes or scan number ($(m/z)\cdot rt$). For each chromatogram, ~4000 variables can be detected, representing $[M + H]^+$, isotopes, adducts, and fragments of individual compounds. All peaks below the arbitrary area threshold of 70 000 were rejected. Deisotoping was not performed. For alignment, the following compounds (peaks) were used as retention time standards: 404.282, 279.335, 496.397 (internal standard (IS)), 524.445, 622.540 (IS), 701.658, 758.711, 812.758, 874.950 (IS), and 930.996. The compounds marked as IS are internal standards added to each plasma sample. The other compounds are lipid compounds naturally present in human plasma or contaminants (plasticizers) present in each chromatogram. The retention times of these peaks were set to a fixed value as observed in batch 4. The retention times of all other compounds eluting between these peaks were corrected for the deviation of the retention time standards from the fixed values by linear interpolation. After alignment, the data were transferred to WinLin. This software was used to check the alignment and for data cleanup. A typical procedure is the removal of all peaks with m/z <300 and with scan numbers <200. These peaks do not represent lipid compounds. The data set was then exported to Matlab for further preprocessing and processing.

(c) Normalization or IS Correction. In Matlab, the data were normalized by correcting for the response of internal standards. The normalization was performed per retention time area corresponding to different lipid classes. All variables with scan numbers below 500 were normalized using the first internal standard (mass 510, scan 421, C17:0 LPC). Scan numbers between 500 and 900 were normalized using the second internal standard (mass 622, scan 541, C24:0 PC). Scan numbers higher than 900 were normalized using the third internal standard (mass 866, scan 971, C51:0 TG).

(d) Missing Values. The missing values in the data set were caused for the following reasons: (i) The peak was present in the sample/chromatogram, but missed by peak picking. (ii) The peak was not present in sample/chromatogram. (iii) The peak was present in sample/chromatogram, but the intensity was below the threshold.

Missing values because of missed peaks during peak picking were reduced by combining duplicate measurements. For each variable, three cases were considered: (i) If both measurement values were zero the combined value was zero. (ii) If both measurements values were both nonzero, the combined value was equal to the average of the two measurements values. In this case. the duplicates were comparable. (iii) If one of the two measurement values was zero and the other measurement value was nonzero, the combined value was equal to the nonzero measurement value.

Next, the data were divided into eight subsets, based on four time points and lean and obese subjects. For the subsets created consider, for example, $t = 0$ measurements for lean and obese subjects only (two subsets). If a variable had a nonzero measurement value in at least 80% of the variables within one of the two subsets, the variable was included in the data set; otherwise the variable was removed. This procedure will be referred to as the "80% rule". Using this rule, missing values caused by peaks that were not present in the sample/chromatogram were reduced.

Despite these preselections, there were still 11% of "zeros" remaining in the combined data set, because of peak intensities below the threshold. The remaining missing values within a subset were imputed using regularized expectation maximization.[22]

The three variables corresponding to the main peaks of the internal standards (used for correction, value 1 in all the samples)

[20] van der Greef, J.; Vogels, J. T. W. E.; Wulfert, F.; Tas, A. C. Method and System for Identifying and Quantifying Chemical Components of a Mixture. U.S. Patent 2004267459, 2004.

[21] Vogels, J. T. W. E.; Tas, A. C.; Venekamp, J.; van der Greef, J. *J. Chemom.* **1996**, *10*, 425–438.

[22] Schneider, T. *J. Climate* **2001**, *14*, 853–871.

were removed from the data set. Variables with a certain value for one sample only and zero for the remaining samples were removed also (705 variables).

(e) Data Sets and Biological Questions. The different subsets were combined into the following two data sets.

$t = 0$ (fasting conditions), obese (100 subjects) versus lean (50 subjects) resulting in a 150 subjects × 947 variables data matrix. The biological question is, Is there a difference between obese and lean subjects lipid metabolites present in fasting blood plasma (blood plasma taken at $t = 0$)?

Kinetics (all four time points) obese (99 subjects) versus lean (50 subjects) resulting in a 149 subjects × 3792 variables data matrix. The biological question is, Is there a difference in plasma lipid metabolites between obese and lean subjects in kinetic profiles?

(f) Centering and Scaling. The $t = 0$ data set was mean-centered or autoscaled[12] after all data preprocessing steps described so far. Mean-centering or autoscaling was applied in order to remove the overall offset.[23] Moreover, autoscaling avoids the possibility that a few high-intensity variables dominate the final solution. The *t*0 data set ensures that only differences between subjects are used for separating obese/lean.

In human studies, a large biological variation is expected in the metabolic profiles. To remove intersubject variation in time profiles of the kinetics data set, each variable was mean-centered for each subject (the mean of all time points was subtracted from each time point value).[24] After this centering step, the kinetics data matrix was autoscaled or not scaled.

Finally, the kinetic data set was rearranged in such a way that the rows of each data matrix represent the subjects and the columns represent mass × retention time × sample collection time point (each variable is a certain mass with corresponding retention time at a certain sample collection time point).

For example, subject A, m/z 412 at scan 1126 and time points 1 and 2:

A   412.1126.1   412.1126.2   .....

This arrangement of the data ensures that only within-subject differences are used for separating obese/lean kinetics.

Data Analysis. An extension of partial least squares (PLS),[12–14,25] partial least-squares discriminant analysis (PLS-DA),[26,27] was used for data analysis. PLS-DA was chosen, because the model has a predictive nature. In PLS-DA, scores × loadings pairs, also called latent variables (LVs), are not calculated to maximize the explained variance in the predicting data set (*X* block with LC–MS metabolomic profiles) only, but also to maximize the covariance with the data to be predicted (*Y*-block with class assignments). The *Y*-block contained "0" and "1" only, corresponding to lean and obese class assignment, respectively.

It is desirable to reduce the number of variables. Variable selection is a difficult problem in the case of megavariate data sets. Classical variable selection methods in regression analysis, e.g., forward selection and stepwise regression,[28] fail due to the high number of variables. In the PLS literature, several alternatives have been suggested, e.g., using the regression coefficients or the VIP-statistic.[29,30] In this study, partial least-squares uninformative variable elimination (PLS-UVE)[31] was used as variable selection method. This method did not give better results compared to PLS-DA, probably caused by the fact that PLS-UVE is only suitable in case of homoscedastic data, which is not the case here. Because of the length of the paper, these results are not reported but are available on request.

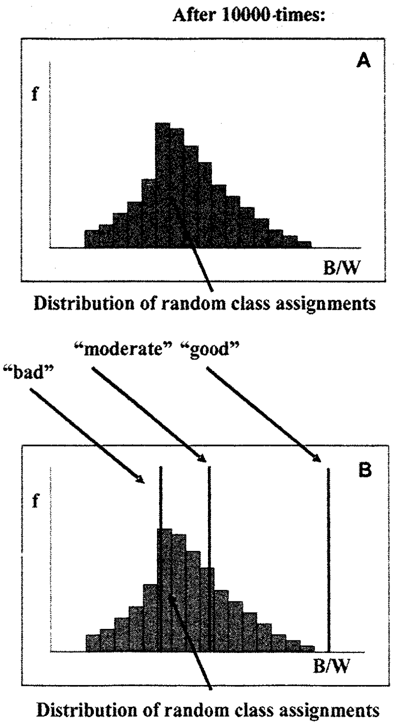

Figure 2. Histogram of B/W distribution after 10 000 permutations (A), "real class" as part of the distribution of random class assigments and "real class" far from the distribution of random class assignments (B).

In this study, from the regression vector plot of PLS-DA, a set of possible biomarkers was defined using the variables with the highest absolute regression values by means of an arbitrary cutoff value. Possible biomarkers found should always be checked with the original data and validated. It should be kept in mind that in the case of so-called "megavariate" data sets, which means that the number of variables is large compared to the number of objects, this approach suffers from chance correlations (false positives).

Statistical Model Validation. (a) Cross-Validation. To choose the optimal number of LVs for the PLS-DA models, 10-fold cross-validation (CV) was applied.[12] Using CV, the predictive ability of the model is tested. In the first step of CV, samples were left out and the remaining samples were used to build a PLS model. The model was used to predict the class assignment of the "left out" samples. This was repeated until all samples were

[23] Bro, R.; Smilde, A. K. *J. Chemom.* **2003**, *17*, 16–33.

[24] Jansen, J. J.; Hoefsloot, H. C. J.; van der Greef, J.; Timmerman, M. E.; Smilde, A. K. *Anal. Chim. Acta* **2005**, *530*, 173–183.

[25] Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1–17.

[26] Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166–173.

[27] Vong, R.; Geladi, P.; Wold, S.; Esbensen, K. *J. Chemom.* **1988**, *2*, 281–296.

[28] Draper, N. R.; Smith, H. *Applied regression analysis*; John Wiley & Sons: New York, 1998.

[29] Perez-Enciso, M.; Tenenhaus, M. *Hum. Genet.* **2003**, *112*, 581–592.

[30] Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

[31] Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M.; Sterna, C. *Anal. Chem.* **1996**, *68*, 3851–3858.

Table 1. Overview of the PLS-DA Results for Fasting Conditions (t = 0) and Kinetics[a]

centering	scaling	outliers	error rate	LVs	perm	conclusion
			$t = 0$			
–	AUTO	0	9	5	good	lean/obese difference
MNCN	–	0	11	13	good	lean/obese difference
			Kinetics			
MNCN-PP	AUTO	2	35	9	good	possible lean/obese difference
MNCN-PP	–	2	32	12	good	possible lean/obese difference

[a] AUTO, autoscaling; error rate, expressed as percentage number of misclassifications; LVs, number of latent variables chosen; MNCN(-PP), mean-centering (per person); outliers, number of outliers removed; perm, permutation test.

left out once. The number of LVs yielding the lowest percentage of missclassifications (error rate) was chosen as the optimal model. In this study, an error rate lower than 30% was considered to be a significant difference between obese and lean subjects. Error rates between 30 and 40% were classified as possible differences. A misclassifications percentage of 50% means a flip of the coin in case of a two-class problem.

(b) Permutation Test. A permutation test[32,33] can be performed in order to test whether differences found between groups are significant. In the permutation test, the Y-block (class assignment) was permuted 10 000 times. For every permutation, a PLS-DA model was build between X-block and permuted Y-block using the same (optimal) number of LVs as determined previously. For every PLS-DA model built, a sum of squares between/sum of squares within (B/W) ratio was calculated for the class assignment predictions. These ratios can be plotted in a histogram. This is called "the distribution of random class assignments". For the real class assignment, the corresponding B/W ratio is calculated and plotted in the distribution of random class assignments as shown in Figure 2. In case this B/W ratio is part of the distribution of random class assignments, it can be concluded that the differences between the two clusters are not significant, which is illustrated in Figure 2.

(c) Evaluation of Possible Biomarkers. Data analysis resulted in a list of possible biomarkers (variables). The cutoff of the regression values for possible biomarkers is arbitrary and differs from model to model. The list consisted of biomarkers from PLS-DA models with a good permutation test and error rates lower than 30%. Using chemical knowledge, artifacts such as gradient peaks were removed. Tentative identification of the possible biomarkers was performed as described in Materials and Methods. For the models, where possible biomarkers were found, new models were calculated using the possible biomarkers only from the original model after the removal of artifacts.

(d) Noninformative Models. As an extra validation of the statistical results, PLS-DA models with "uninformative" variables were calculated. The goal of these models was to confirm the separation of the groups based on the possible biomarkers evaluated. For these models, all variables considered as possible biomarkers were excluded from the data set. The error rate of the cross-validation of these models should be higher than the error rate obtained for the models that contained possible biomarkers. This is not always the case and is caused by the arbitrary cutoff of the regression vector from the PLS-DA models, when evaluating the possible biomarkers.

(e) Biomarker Models, Targeted Data Analysis. As an additional statistical evaluation, possible biomarkers from all models, identified as compounds (both known and unknown), were compiled into one list. For this list of compounds, a targeted method was set up, using LC-Quan software. The software identified and "relatively" quantified all the compounds in this list (with respect to an internal standard). In this way, a new data set was created. The relative quantification in this data set is more reliable than the untargeted approach as described earlier. New biomarker models were calculated using this data set. It is expected that the error rate of these models is better than the error rate obtained for the first models.

RESULTS AND DISCUSSION

Data Analysis, t = 0. The PLS-DA results for plasma taken at fasting conditions (t = 0) are listed in Table 1. There is a significant difference between the plasma lipid profiles of obese and lean subjects, because of the low error rates (<30%) and good permutation test results. Mean-centering and autoscaling of the data set give comparable results. Using autoscaling, the model is less complex (lower number of LVs) if compared to mean-

Table 2. Overview of Possible Biomarkers Selected for Targeted "Validation"

compound	compound
C18:2 LPC[a,b]	C48:1 TG[b]
C16:0 LPC[a,b]	C56:7 TG[b]
C18:1 LPC[a]	C54:3 TG[a]
C18:0 LPC[a,b]	C54:5 TG[b]
C32:1 PC[a]	C50:2 TG[a]
C38:6 PC[a,b]	C58:8 TG[b]
C34:2 PC[b]	C56:6 TG[b]
C38:5 PC[a]	C52:3 TG[b]
C34:1 PC[a,b]	C14:1 SPM[a]
C40:6 PC[a]	C16:1 SPM[a]
C38:4 PC[a,b]	C16:0 SPM[a]
C40:4 PC[a]	C20:0 SPM[a]
C48:3 TG[b]	C18:0SPM[a,b]
C44:0 TG[b]	C18:1 SPM[a]
C52:5 TG[a]	C22:0 SPM[a]
C58:8 TG[b]	C20:4 ChE[a,b]
C54:6 TG[b]	C18:2 ChE[b]
C52:4 TG[b]	C18:1 ChE[b]

[a] Possible biomarker for t = 0. [b] Possible biomarker in the kinetic profiles.

[32] Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: New York, 1993; Chapter 15.

[33] Good, P. I. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*; Springer-Verlag: New York, 2000.

Table 3. Overview of "Noninformative" Models for $t = 0$ and Kinetics[a]

SET	centering	outliers	error rate$_{non-inf}$	error rate$_{orig}$	LVs$_{non-inf}$	LVs$_{orig}$	regression values evaluated
$t = 0$	MNCN	0	20	11	14	13	<−0.15 >0.15
$t = 0$	MNCN	0	28	11	5	13	<−0.05 >0.05
$t = 0$	MNCN	0	31	11	4	13	<−0.01 >0.01
kinetics	MNCN−PP	2	43	32	4	12	<−0.1 >0.1

[a] Error rate, expressed as percentage number of misclassifications; LVs, number of latent variables chosen; MNCN(-PP), mean-centering (per person); noninf, noninformative; orig, original; outliers, number of outliers removed.

Table 4. Overview of the PLS-DA Results for the Targeted Method for $t = 0$ and Kinetics[a]

SET	centering	scaling	error rate	LVs	perm	conclusion
*t*0	MNCN	0	7	15	good	lean/obese difference
kinetics	MNCN-PP	0	26	11	good	lean/obese difference

[a] Error rate, expressed as percentage number of misclassifications; LVs, number of latent variables chosen; MNCN(-PP), mean-centering (per person).

centering. However, autoscaling is sensitive to artifacts. Variables of low abundance and a relatively high contribution of analytical error are emphasized by autoscaling, thereby blurring the results. Next, possible biomarkers were evaluated for the $t = 0$ PLS-DA models for both autoscaling and mean-centering.

Data Analysis, Kinetics. The PLS-DA results for kinetics are listed in Table 1. From the table, it can be concluded that there are possible differences between the plasma metabolic profiles of obese and lean subjects for both scaling methods (high error rate (>30%) despite good results for the permutation test). Next, variables with the highest absolute regression values were considered as possible biomarkers. A common possible biomarker for all time points seems to be C18:2 LPC.

Noninformative Models. A possible biomarker list was compiled from all models ($t = 0$ and kinetics) showing significant or possible difference; see Table 2. Noninformative models were calculated using all variables of the original data set, without the possible biomarkers. The calculations were limited to models where mean-centering was used only. The results are shown in Table 3. The error rate increases in all cases. For some of the noninformative models, the error rate is still good. This can be explained by the arbitrary cutoff of the regression values for the PLS-DA models when the possible biomarkersare evaluated. It is possible that several variables with regression values just below the evaluated levels significantly contribute to the separation between the compared groups. If the limits for the regression values are lowered, the error rate increases further. Several noninformative models were calculated.

Targeted Approach for Possible Biomarkers. The compounds from Table 2 were considered as targeted compounds and were reintegrated. New PLS-DA models were calculated for this limited data set with mean-centering in case of $t = 0$ and with mean-centering per person in case of kinetics. An overview of the results from the new models is given in Table 4. It can be concluded that the error rate of the differences in kinetics between lean and obese has been improved. The observed difference is now significant. For $t = 0$, the error rate is comparable to the error rate of the original model. This is an extra statistical confirmation of compounds of the difference between lean and obese subjects. These compounds are considered to be potential biomarkers and will be evaluated biologically.

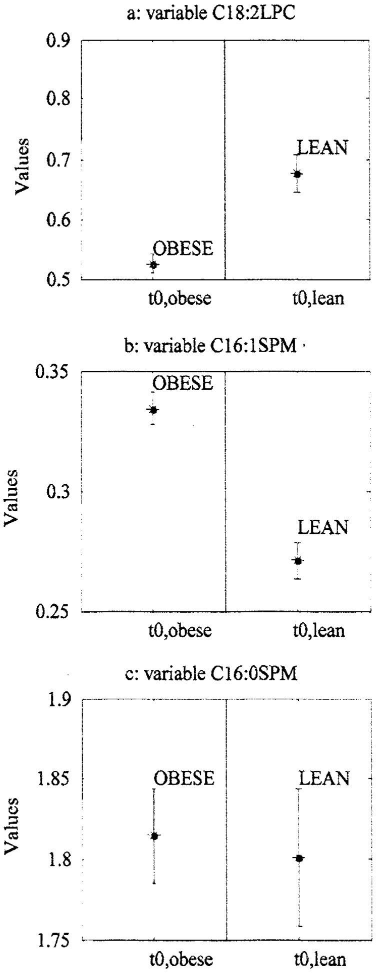

Figure 3. Univariate up- or downregulation plots for $t = 0$ using targeted data for potential biomarker: C18:2 LPC (a), C16:1 SPM (b), and C16:0 SPM (c). An asterisk (*) represents the mean value, and the error bars indicate the standard deviation of the mean.

To obtain insight in the results of multivariate data analysis, univariate plots of the "raw" data were made without the application of univariate statistical tests. For these plots, data after IS correction, combination of duplo measurements, and imputation of missings were used. In these plots, the mean and standard deviation of the mean were plotted for each group. Although there is no simple relationship between univariate and multivariate analysis, such plots can help to interpret the results, namely, indicate up- or downregulation of individual compounds between groups. For the potential biomarkers found using PLS-DA, univariate plots were made. Figure 3a shows a plot for potential biomarker C18:2 LPC for $t = 0$. In this figure, the mean concentration seems to be higher for lean subjects. In Figure 3b, the plot is shown for biomarker C16:1 SPM for $t = 0$, where the mean concentration seems to be higher for obese subjects. Figure 3c represents a plot for potential biomarker C16:0 SPM for $t = 0$, where it is concluded that the mean concentration for obese and lean subjects is comparable.

CONCLUSIONS

In this paper, a strategy for preprocessing, analysis, and validation of large human metabolomic studies is described. The strategy proved to be applicable in the case of a large biological variation between samples and is therefore very powerful. The pragmatic approach is demonstrated on LC−MS lipidomic metabolomics data with small differences, but can also be applied to other data from large human studies. Data preprocessing included combining duplicate measurements instead of averaging in order to correct for errors (missed peaks) during peak picking. After combining the duplicate measurements, the data contained less missing values. Remaining missing values were imputed. Data analysis resulted in lists of possible biomarkers using the obtained regression vectors from PLS-DA models. It is difficult to choose a good cutoff value for the regression vectors using PLS-DA and megavariate data sets. Using chemical knowledge, artifacts such as gradient peaks were removed from the possible biomarker lists. Statistical validation of the possible biomarkers was performed using cross-validation, permutation tests, biomarker models, and noninformative models. Finally, this resulted in a list of potential biomarkers. Using univariate plots, up- or downregulation of a potential biomarker was determined. From the results of the lipidomic data set it can be concluded that despite the large biological variation (partially caused by the multicenter approach) a significant difference was observed between the plasma lipid metabolic profiles of obese and lean subjects at fasting ($t = 0$) conditions. For the kinetic set, a possible difference was found between lean and obese subjects. It should be kept in mind that further biological interpretation and evaluation of potential biomarkers is required. Without biological validation, the reliability of potential biomarkers is not guaranteed. The comparison of results of the application of this strategy to the data of LC−MS lipidomic and three other analytical methods (LC−MS polar, GC/MS, NMR) will be published in a separate paper.

Received for review August 19, 2005. Accepted November 16, 2005.

AC051495J