Metabolomics (2007) 3:231–241
DOI 10.1007/s11306-007-0081-3

ORIGINAL ARTICLE

Proposed minimum reporting standards for data analysis in metabolomics

Royston Goodacre · David Broadhurst · Age K. Smilde · Bruce S. Kristal · J. David Baker · Richard Beger · Conrad Bessant · Susan Connor · Giorgio Capuani · Andrew Craig · Tim Ebbels · Douglas B. Kell · Cesare Manetti · Jack Newton · Giovanni Paternostro · Ray Somorjai · Michael Sjöström · Johan Trygg · Florian Wulfert

Received: 20 July 2007 / Accepted: 27 July 2007 / Published online: 25 August 2007

Abstract The goal of this group is to define the reporting requirements associated with the statistical analysis (including univariate, multivariate, informatics, machine learning etc.) of metabolite data with respect to other measured/collected experimental data (often called meta-data). These definitions will embrace as many aspects of a complete metabolomics study as possible at this time. In chronological order this will include: *Experimental Design*, both in terms of sample collection/matching, and data acquisition scheduling of samples through whichever spectroscopic technology used; *Deconvolution* (if required); *Pre-processing*, for example, data cleaning, outlier detection, row/column scaling, or other transformations; Definition and parameterization of subsequent

R. Goodacre · D. Broadhurst (✉) · D. B. Kell
School of Chemistry and Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7ND, UK
e-mail: david.broadhurst@manchester.ac.uk

A. K. Smilde
Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam 1018 WV, Netherlands

A. K. Smilde
TNO Quality of Life, Utrechtseweg 48, P.O. Box 360, Zeist 3700 AJ, Netherlands

B. S. Kristal
Department of Neurosurgery, Brigham and Women's Hospital, 221 Longwood Ave, Boston, MA 02115, USA

J. D. Baker
Pfizer, Inc, Ann Arbor, MI, USA

R. Beger
Division of Systems Toxicology, National Center for Toxicological Research, 3900 NCTR Road, Jefferson, AR 72079, USA

C. Bessant · C. Manetti
Cranfield University, Silsoe, Bedfordshire MK45 4DT, UK

S. Connor
Safety Assessment, GlaxoSmithKline, Park Road, Ware, Herts SG12 0DP, UK

G. Capuani
Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, Rome 00185, Italy

A. Craig
BlueGnome Ltd, Breaks House, Mill Court, Great Shelford, Cambridge CB2 5LD, UK

T. Ebbels
Department of Biomolecular Medicine, Imperial College London, London SW7 2AZ, UK

J. Newton
Chenomx Inc,
Suite 800, 10050 112 St, T5K 2J1
Edmonton, AB, Canada

G. Paternostro
Burnham Institute for Medical Research, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

R. Somorjai
Institute for Biodiagnostics, NRCC, 435 Ellice Ave, R3B 1Y6 Winnipeg, MB, Canada

M. Sjöström · J. Trygg
Research Group for Chemometrics, Organic Chemistry, Department of Chemistry, Umeå University, Umea 901 87, Sweden

F. Wulfert
Division of Food Sciences, University of Nottingham, Sutton Bonington Campus, Loughborough LE12 5RD, UK

visualizations and *Statistical/Machine learning Methods* applied to the dataset; If required, a clear definition of the *Model Validation Scheme* used (including how data are split into training/validation/test sets); Formal indication on whether the data analysis has been *Independently Tested* (either by experimental reproduction, or blind hold out test set). Finally, data interpretation and the visual representations and hypotheses obtained from the data analyses.

Keywords Chemometrics · Multivariate · Megavariate · Unsupervised learning · Supervised learning · Informatics · Bioinformatics · Statistics · Biostatistics · Machine learning · Statistical learning

1 Introduction

It is clear that algorithms do not drive metabolomics investigations; however the question(s) one seeks to answer with metabolomics are clearly likely to dominate any subsequent data analysis strategy. In many metabolomics experiments, the number of samples collected is much smaller than the number of metabolites or variables (features) measured, and simple visual inspection of all the data is not likely to be sufficient to complete the analysis. Therefore, there is a need for some method to extract information from the flood of data (Goodacre et al. 2004; Hall 2006; Weckwerth and Morgenthal 2005). Thus, statistical (univariate, multivariate) or machine learning analysis (for the difference see Breiman 2001) of metabolite data, as well as various attendant informatics analyses, is necessary to produce knowledge that can be tested and lead to new hypotheses and biological understanding.

Figure 1 shows the flow of information (pipeline) in a metabolomics experiment (Brown et al. 2005). The first step is judicious design of the experiment (DoE), followed by data capture, followed by storage of the data and associated metadata, followed by pre-processing, followed by data analysis and followed by data interpretation. In some cases, there may be iterative loops at one or more of these stages (Paolucci et al. 2004a, b). This pipeline ends in one or more report(s) that attempt(s) to chronicle this process (or workflow), summarize the data analysis, then derive some conclusions from the data analysis and test those results. At present, this reporting process is performed in a very ad-hoc manner, with different authors reporting in very different ways (even within the same institution, or same lab, across multiple publications), and sometimes barely explaining at all exactly how the statistical analyses were performed.

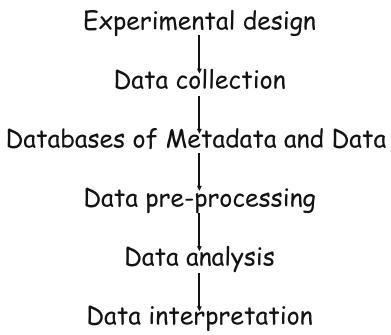

Fig. 1 The initial linear flow diagram identifying the clear flow of information (pipeline) in a typical metabolomics experiment. This is necessarily simplified as the real situation is much more complicated: However, in all cases it starts with a biological question and it ends with data interpretation

This document aims to formalize the reporting of metabolomics data analysis in two ways: (i) to define a 'reporting' scheme (detailed in Tables 1–6) so as to avoid confusion about terminology, and; (ii) to present the first publication of what is considered by this committee to comprise the minimal reporting requirements for each stage, from the pre-processing of the data to the validation of the hypotheses obtained through initial data analyses. In conjunction with this formal part of the document, at several points the authors will also recommend some basic strategies for performing the various stages of the data analysis workflow being defined. These will be used to illustrate the reporting ideas and to help guide the less experienced members of our community. This committee strongly seek and encourage feedback from the community about these guidelines and asks for potential changes to improve upon the guidelines.

2 Design of experiments (DoE)

Any good scientific study starts with rigorous experimental design (Montgomery 2001). Formalizing this process is both useful to the practitioner and to the reader of associated publications. There are generally two stages involved in the study design. First, the biological study itself, this can be as simple as collecting a set of biological replicates for a steady state system or be as complex as a major epidemiological project (Bland 2000; Rothman and Greenland 1998) involving many thousands of subjects at various levels of exposure to a factor of interest, multiple ages, different genders and co-morbidities, and/or at all stages of a disease process. Realistically, all currently defined statistical and machine learning methods are only capable of interpolation, that is to say they give answers within their knowledge realm and can not accurately extrapolate beyond this. Therefore, as much metadata as possible should be made openly available, and all metadata that has been used to discover information in the metabolomics dataset must be made available. This will allow

Table 1 Level 1

Term	Explanation	Remarks
Pre-processing	Generic term for methods to go from raw instrumental data to clean data for data processing	
Pre-treatment	Transforming the clean data to make them ready for data processing (scaling, centering, etc)	Bro and Smilde (2003)
Processing	The actual data analysis (PCA, PLS, ASCA, GP etc.)	
Post-processing	Transforming the results from the processing for interpretation and visualization (e.g., antilog etc)	Nicholson et al. (1999)
Validation	All activities aimed at assuring the quality of the conclusions drawn from the data analysis	
Interpretation	Hypothesis generated, pathways affected, or visualization of the data.	

Scheme for reporting metabolomics data analysis: The reporting scheme has several levels. Level 1 is the subdivision in several steps of the metabolomics pipeline. Level 2 is a subdivision of the level 1 items

Notation: $\mathbf{X}$ has $i = 1,...,I$ rows (the samples) and $j = 1,...J$ columns (the metabolites, chemical shift regions, or m/z.rt channels). In case of hyphenated methods (e.g., GC-MS, LC-MS, LC-NMR etc), the rows of $\mathbf{X}$ are strung out (or vectorized) GC-MS (or LC-MS) profiles

Table 2 Level 2: pre-processing

Term	Explanation	Remarks
Deconvolution	Resolving overlapping peaks in an NMR spectrum or GC or LC chromatogram using the second dimension (usually MS). In the case of GC or LC this generates a peak table where each metabolite is represented by one variable	Jonsson et al. (2004); Kvalheim and Liang (1992); Tauler et al. (1992); Veldhuis et al. (1987); Weljie et al. (2006)
Peak-picking	Peaks in an NMR or GC or LC-MS chromatogram are selected that may represent signals. This results in a table with either ppm or m/z.rt channels and corresponding intensities.	see CODA (Windig et al. 1996), and (Katajamaa and Oresic 2005)
Target analysis	Peaks in an NMR spectrum or GC or LC chromatogram at a specific δ or m/z channel are integrated and used in a peak table	Andreev et al. (2003)
Alignment	Synchronizing the chromatograms or NMR spectra such that each metabolite signal has the same retention time or chemical shift in each sample.	see Warping, COW, DTW, PLF (Forshed et al. 2005; Skov et al. 2007; Tomasi et al. 2004; Vogels et al. 1996)
Apodization function and weighting factors	Function and parameters used to multiply free induction decays (FIDs) before Fourier transform to NMR spectra	
Phasing	Method used to phase-correct peaks in Fourier transformed NMR spectra	Phasing can be done manually or automatically by NMR software
Base-line Correction	Method used to address baseline tilts and drifts in Fourier transformed NMR spectra.	Methods used to correct baseline features in NMR. Usually done automatically or semi-automatically
Bucketing	Method used to define chemical shift bins sizes and integrate the bin intensities	Holmes et al. (1992); Nicholson et al. (1999)

the scientific community to assess adequately the validity of the study at large. This is particularly the case in metabolomics when the numbers of variables can exceed the numbers of samples, and the statistical powering used; consequently, the software used to calculate the statistical power (under assumptions of normality or otherwise) must be given. In larger studies, and longer-term longitudinal investigations it may prove to be difficult to submit all of the metabolomics and metadata at the time of publication, so summaries of salient parts of such data should be reported in a table and if possible full meta-data should be available upon request; or via the author's website, or better (Kell 2007) by accompanying the manuscript as supplementary information if that mechanism exists with the journal publishing the work. However, where release of full meta-data is not possible immediately, for example where interim results are published, it would be accepted that there may be a delay prior to the release of such data and publication can proceed without it. An example of essential data that does always need to accompany a publication is the patient characteristics table (Table 1) in Sabatine et al. (2005). In 'case'/'control' clinical studies, examples of such important metadata may be the way in which 'controls' are matched to 'cases,' descriptive statistics about the distribution of ages in the sample cohort, genetic single nucleotide polymorphisms (SNPs), or count

Table 3 Level 2: pre-treatment

Term	Explanation	Remarks
Normalization	Operation performed within or across rows to make the row profiles comparable in size	E.g., for correcting dilution differences in urine analysis by NMR or LC-MS
Centering	Operation across the rows to translate the center of gravity of the dataset	Also called reference subtraction. SMART also centers, but not with a mean (Keun et al. 2004)
Mean-centering	Commonly used method for centering in which each column is expressed in deviations from its mean (across the rows)	Subtracts the mean of the column, thereby translating the center of gravity of the data to the origin
Scaling	Operation performed within a column to make the column profiles more comparable	
Autoscaling	A form of scaling which mean-centers each value of the column followed by dividing row entries of a column by the standard deviation within that column	Also called UV (unit variance) or Z-scaling
Range scaling	Mean-centering followed by dividing row entries of a column through the range within that column	van den Berg et al. (2006)
Pareto scaling	Mean-centering followed by dividing row entries of a column through the square root of the standard deviation within that column	
Transformations (Log, Square Root, Box- Cox)	Transformations to linearize or otherwise change the scale of the data, e.g., to remove heteroscedastic noise	
Missing values	Data in the table which are not available for the analysis	Rubbin and Little (1987)
Outliers	Data points (samples, variables or a specific combination of both) which deviate from the distribution of the majority of the data	

Table 4 Level 2: processing

Term	Explanation	Remarks
Model	The model selected for analyzing the data (PCA, PLS, LDA, QDA etc.)	
Method	The method selected for analyzing the data (e.g. GP, etc.)	
Parameter estimation	Parameters in models/methods that have to be fitted to the data	
Meta-parameter estimation	A parameter that helps define the structure and optimization of the model (e.g., number of LVs in PLS, ridge parameter etc.)	

Table 5 Level 2: post-processing

Term	Explanation	Remarks
Back-transformation	Transforming the data back to the original domain (if a transformation was performed prior to the analysis)	Cloarec et al. (2005b)
Visualization	Plots that represent the original data or the results from the data analysis in a such a way that facilitates interpretation	

Table 6 Level 2: validation

Term	Explanation	Remarks
Training set	Subset of samples used to estimate the parameters	
Monitoring set	Subset of samples used to estimate the metaparameters	
Test set	Subset of samples used to establish the generalizability of the model/method	

ratios for known sources of bias such as male/female or smokers/non-smokers/ex-smokers etc are exhibited. An important piece of meta-data is the method used to identify the classes of the sample, including the expected accuracy of that method. Many diagnosis methods (e.g., grade of cancer Campbell et al. 2001) are far from 100% accurate, and this is obviously going to affect the later data analysis. In addition a particularly important issue that can cause confounding or bias (Broadhurst and Kell 2006; Ioannidis 2005; Ransohoff 2005) is the likelihood that 'cases' will be often be taking considerably more pharmaceutical drugs than are the 'matched' 'controls.' Summation data may also be necessary in some studies due to legal requirements, such as US HIPAA regulations, which restricts use/release of some medical information.

The second part of the DoE is the analytical design. Once samples are collected (or grown) the analysis procedure should be reported with enough detail so that the

experiment can be replicated by others. This will include the number of analytical replicates that were processed, the protocol for loading samples into the respective analytical instrument (e.g., randomized or designed), the instrumental parameters and the duration of the experiment. At first glance this may appear beyond the scope of the data analyst; however this information may prove crucial in the interpretation of the data. For example, a particular study might compare disease ('case' and 'control') samples where all the 'cases' are measured on a Monday and all the 'controls' on a Tuesday; it would be impossible to know whether any clustering was due to disease state or day of data acquisition, until the hypothesis or biomarkers were further validated with subsequent testing. Best of course is to randomize the time of analysis equally between the two cohorts (if sample numbers are small stratified random sampling may be more reliable (Bland 2000)).

3 Data reduction/deconvolution

In the context of metabolomics, the term deconvolution (the separation of overlapping signals into individual chemical peaks) is mainly used in the realm of hyphenated chromatography/mass spectrometry (MS) where raw 3D matrices (time vs. mass vs. intensity) are particularly complex. In the realm of nuclear magnetic resonance (NMR), deconvolution may describe the separation of overlapping peaks either into individual resonances (e.g., lipoproteins from lactate in serum/plasma spectra (see for example Serrai et al. 1998) or to metabolite lists using reference spectral libraries to aid deconvolution (for example Provencher 1993; Weljie et al. 2006). In order to avoid confusion and to expand the process to encompass other methods of converting raw signal (from any measurement technology) into a list of quantitative metabolite concentrations, the term *data transformation* is preferred**.** Thus, for a single study, the starting point for pre-processing and then data analysis is a single matrix $N \times D$, where, N = the number of samples and D = the number of variables. These variables may, for example, represent actual metabolites, or conversely may correspond to binned regions of a continuous property of the data (e.g., binned chemical shifts in NMR, or wavenumbers in Fourier transform infrared spectroscopy (FT-IR)).

There are examples of data analysis methods that can be applied to more complex representations of metabolite information; however, they are more the exception than the rule and therefore will be treated as such. Practitioners of these methods are encouraged to follow the reporting standards described below as closely as possible. Minimum reporting standards for 'deconvolution' are very much instrument- (and even manufacturer-) dependent and are discussed by the Chemical Analysis working group for MS (with and without chromatography; viz. gas chromatography (GC)-MS and liquid chromatography (LC)-MS), NMR spectroscopy, and FT-IR spectroscopy in Sumner et al. (2007).

4 Pre-processing and pre-treatment

We outline a proposed scheme to describe each step of the metabolomics workflow in Table 1. In Table 1 we make a distinction between pre-processing and pre-treatment. For NMR data, raw free induction decay (FID) weighting, phasing, baseline correction and referencing to an internal standard, normalizing to spectral area, and conversion to magnitude spectra are considered pre-processing procedures. The process of defining the sizes of chemical shift bins and integrating the intensities in the chemical shift bins is pre-treatment. NMR peak alignment to account for chemical shift variability, whether global over the whole spectrum, or localized to specific regions is regarded as pre-processing.

There are many methods for pre-processing data matrices of the sort produced by metabolomics studies, many of which are order-independent. In this light of this it is necessary to report what pre-processing has been done, in what order, and what software was used. The easiest way to report this is in the form of a workflow. This can be expressed in the text or preferably as a flow diagram (especially if the workflow is unusual or complex).

Table 2 defines a list of the popular order-independent pre-processing methodologies and their definitions, and Table 3 lists of popular pre-treatment algorithms. These definitions include whether the methods have any meta-parameters, and whether the method operates per sample, per variable, or on the total matrix. Any 'new' or unusual method not listed here should be clearly explained (rather than referencing this table) and should be reported in a similar fashion to that shown here.

Order-dependent pre-processing methods are more difficult to report as they may need some level of interpretation or assessment by the reader before the analysis moves onto the next stage (such as outlier detection, e.g., by using principal components analysis (PCA) or other techniques, missing value imputation etc.). More than one workflow may need to be reported. For example, the author could present one workflow for the outlier detection and another for subsequent cluster/discriminant analysis. The order of execution and contents of these sub-workflows should be made very clear (this could include reference to previously published workflows). Any assumptions made about the structure of the data should

also be made clear, as should any decisions made at any stage (e.g., on what basis were outliers removed).

5 Data analysis and algorithm selection

The sort of question that one wants to answer generally drives the workflow, including the selection of the appropriate algorithm (or set of algorithms). See Fig. 2 for a diagrammatic representation of the most popular ones. However, algorithms are also likely to be specified based on previous experiences and local expertise; it is not the role of this article to suggest 'preferred' algorithms (and anyway the best methods depend on the problem domain Wolpert 1997). In addition, it is not feasible to discuss the pros and cons of each method as these are often subjective, but we can define a reporting structure based on the biological application. We can also suggest that a key step in the validation of any statistical result of metabolomics data is the visualization of the proposed result against the raw data to confirm that it is not an artefact of data acquisition, pre-processing, pre-treatment or noise magnified by the statistical scaling method applied.

6 Univariate analysis

Although metabolomics experiments do generate multivariate data (see below); one can employ univariate methods to test individually for metabolites that are increased or decreased significantly between different groups (note: consideration of the issues of multiple parallel hypothesis are needed when applying univariate tests to multivariate data). These tests include parametric methods for data that are normally distributed, the most common being ANOVA (analysis of variance), *t*-tests, and *z*-tests (for a review of univariate methods and a discussion on the issues of *multiple hypothesis testing* see Broadhurst and Kell 2006). When a normal distribution of the data cannot be assumed, then non-parametric methods can be used, e.g., the Kruskal–Wallis test. These tests produce a test statistic from which statistical significance and confidence can be calculated. Usage and reporting of these univariate procedures is haphazard in publications across the whole of science. It is suggested that the metabolomics community take section IV.A.6.c. of the *Vancouver Guidelines* (http://www.icmje.org/) as the starting point:

> "Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results. When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information."

In their article entitled 'Statistics with Confidence' Altman and colleagues (Altman et al. 2000) also suggest that:

> "For the major finding(s) of a study we recommend that full statistical information should be given, including sample estimates, confidence intervals, tests statistics, and P values—assuming basic details such as sample sizes and standard deviations have been reported earlier in the paper."

In addition, graphics/visualizations should be accompanied by sufficient metadata such that a knowledgeable reader could reproduce them given access to the data and to appropriate software.

An example: Suppose that in a study comparing samples from 100 diabetic and 100 non-diabetic men of a certain age, a difference of 6.0 mmHg was found between their mean systolic blood pressures. This could be reported as either: (i) a Student's *t* test was performed after normality was assessed using the 'XYZ' test. The test statistic was 2.4, with 198 degrees of freedom and had an associated *P*-value of 0.02. The 95% confidence interval for this difference was calculated to be from 1.1 to 10.9 mmHg, or (ii) these groups differed at $P < 0.05$ (Student's *t*-test, after normality was assessed using the XYZ test) mean ± SD difference was 6.0 ± xx. Combinations of these may also be used, either alone or conjunction with other measures such as standard error of the mean.

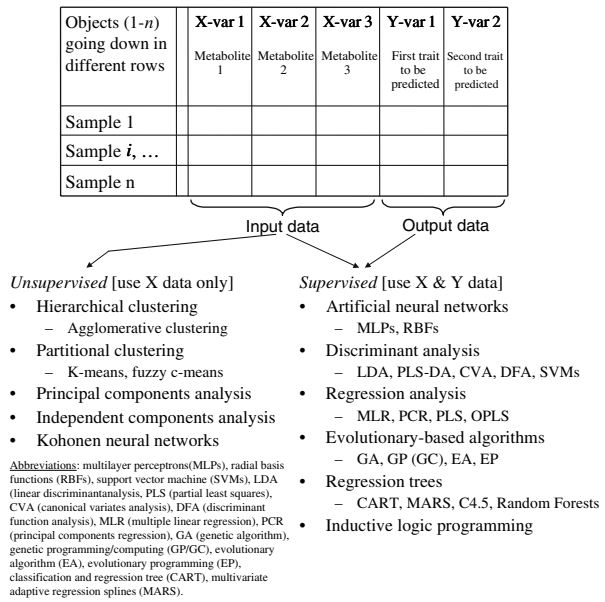

Fig. 2 Typical multivariate algorithms used for unsupervised and supervised analysis

Methods of univariate data analysis are constantly being developed and/or refined. Over the past 30 years accepted ideologies have been repeatedly questioned and advances in computational power have produced new ways of interpreting established tests (for example, bootstrapping of test statistics Efron and Gong 1983; Efron and Tibshirani 1993). It is strongly recommended that the author(s) to provide sufficient information for novel method(s) to be independently reproducible and verifiable.

7 Multivariate analysis

As mentioned earlier, metabolomics data are essentially multivariate. For a single study, the starting point for data analysis is a single matrix N × D where, N is the number of samples (objects) and D is the number of variables (metabolites/binned chemical shifts/wavenumbers/masses/retention times etc.). Each independent variable may be thought of as a single geometric dimension, such that each sample may be considered to exist as a single point in an abstract entity referred to as D-dimensional hyperspace. The role of the analyst is to understand/interpret/query the underlying properties of the data points distributed in this hyperspace. This may be done in the form of unsupervised dimensionality reduction (i.e., usually projection into a space of significantly smaller dimensionality whilst trying to minimize information loss), or clustering (identifying groups of points which are more similar to each other than to the rest of the data), or by correlation analysis, or by supervised pattern recognition/machine learning (including multivariate regression and discriminant analysis). It is beyond the scope of this document to describe (or criticize) all the available methods (good starting points are Duda et al. 2001; Hastie et al. 2001). This would be an impossible task given the current number and the speed in which new methods are being developed. The central point of this report is not to prescribe methodologies, but rather to ensure that the methods used are reported consistently and in detail. With this in mind, we have tried to distil the salient features of *unsupervised* and *supervised* multivariate analysis in order to unify the reporting process. There are several excellent reviews on current multivariate techniques (Beebe et al. 1998; Chatfield and Collins 1980; Duda et al. 2001; Everitt 1993; Krzanowski 1988; Lavine 1998; Manly 1994; Martens and Næs 1989; Massart et al. 1997).

8 Unsupervised methods

Unsupervised methods can be split into two basic forms. First, there is *dimension reduction* into a lower dimensional space typified by principal component analysis (PCA; Jolliffe 1986). This is fundamentally a multivariate linear *transformation* and is often used as a pre-processing step prior to application of a supervised method. One important reporting method needed for such methods is a reference to the software used or the theoretical source of the mathematical algorithm (and possibly the related computer code) should be included and a statement about any assumptions made about the characteristics of the dataset before transformation.

A second type of unsupervised method is *Cluster Analysis* (see e.g., Everitt 1993). Here the algorithm attempts to 'find' clusters of similarly characterized samples (i.e., points clustering together in the multi-dimensional feature space). Once found they may be used to classify each sample and similarities between clusters may be assessed (for example using hierarchical cluster analysis). Again, in terms of reporting, details of the algorithm or software used must be given, including the similarity measure used. It is also the case that pretty well any clustering algorithm will perform a clustering, and it is important to assess the validity of such clusters. A variety of methods exist for this and if readers are to be persuaded that the clustering has meaning then these should to be used and reported. A recent reviewed describes the different approaches available (Handl et al. 2005). Reporting of other novel and standard unsupervised techniques, such as self-organizing maps (Kohonen 1989), should follow similar guidelines.

In both cases, the method for optimizing meta-parameters such as the number of components in PCA or architecture of the self-organizing map should be reported.

It is also possible to determine correlations between the metabolites or variables measured by conducting *Correlation Analysis*. In this case as well as generating correlations usually depicted as nodes linked by edges (Broadhurst and Kell 2006; Ebbels et al. 2006), often with the correlation coefficient given, it is possible to construct correlation matrix pseudocolor maps which can be use as first step to explore the data (Cloarec et al. 2005a; Miccheli et al. 2006) and integrate data from different sources (Crockford et al. 2006). In terms of reporting, details of the correlation algorithm or software used must be given, along with any cut of point for what is considered a significant correlation.

9 Supervised learning

When one knows the classes or values of the responses that one is trying to predict (also known as the *Y*-data) associated with each of the sample inputs (*X*-data), then supervised methods may be used. Ideally, the goal here

is to find a model (mathematical/rule based/decision tree) that will correctly associate all of the *X*-data with the target *Y*-Data. The desired responses may be categorical (e.g., disease vs. healthy) or quantitative/continuous (e.g., blood glucose, body mass index, age, etc.). Usually, supervised methods require some sort of meta-parameterization (Table 4). Meta parameters are parameters that help define the structure and optimization of the model. For example, in a PLS model (Eriksson et al. 2001; Martens and Næs 1989; Trygg and Wold 2002), the actual weights (loadings) are the parameters of the model. The number of latent variables is a single meta parameter. In neural networks (e.g., Bishop 1995; Ripley 1996), the learning rate and criteria for halting the learning process are examples of meta parameters. In genetic programming (Kell 2002; Koza 1992; Langdon 1998)—mutation rate, crossover type, operator lists and maximum tree depth are examples of meta parameters. In methods employing variable selection, the information specifying which variables have been selected for a particular model is also considered a meta-parameter, since this is often used to optimize model performance. All of these attributes must be reported alongside the final model itself in order for other scientists to be able to have a chance of reproducing the model.

Supervised methods often need to be reported in considerably more detail than unsupervised ones. The more meta-parameters involved, the more detailed the reporting required (see example below). As with the unsupervised methods, details of the algorithm or software used must be given. In addition, all static meta-parameters should be reported, and the method of optimizing the dynamic meta-parameters described in detail—including internal model validation (*vide infra*).

An example: Suppose that one was optimizing the meta-parameters for a feed forward neural network using the gradient descent back propagation algorithm (Wasserman 1989) which was being trained to differentiate between diseased versus healthy patients. Although the following list is not exhaustive, the meta-parameters that one would have to optimize include: number of layers the neural network contained, number of nodes in the hidden layers, whether a bias node (set to +1) was used or not, what the learning rate and momentum were, what squashing function was used (e.g., sigmoidal, *tanh*, linear, etc.), when the error was propagated (after each training pair was presented to the network, or after all presentations), and how many iterations/epochs the neural network was trained for. Likewise, and of equal importance, one should note the approaches tried and discarded. Ideally, researchers could also report a synopsis of the results of these discarded approaches.

10 Model generality and model validation

If the results of a metabolomics study indicate some sort of general model, or general biomarker discovery, for a particular biological system/study then reports of this assertion must be backed up by some sort of model validation.

Model validation is needed in both supervised and unsupervised analysis. Model validation is often misinterpreted as simply model optimization through internal validation. Internal validation is used in supervised methods to optimize meta-parameters (e.g., number of latent variables in PLS, or number of training epochs in neural networks). Even if internal validation appears to often improve predictive capability of the final model chosen, this model's generalization capability tends to be overoptimistic (overfitting), hence it is not a replacement for extrinsic model validation.

Model validation in its simplest form involves splitting the available data into 3 sets: training, monitoring, and test (Table 6). The training data are used to build one or more possible models, the monitoring data are then used to assess and optimize the quality of every 'trained' model (e.g., via meta-parameters) and the independent test data are used to measure the generalization/predictive expectation of the final published 'optimal' model. There are several validation methods in which the training and monitoring sets are initially combined and then subsequently dynamically partitioned into temporary training/monitoring datasets (e.g., bootstrapping and K-fold cross-validation). Approaches which make use of training and monitoring data can be said to be examples of empirical model selection, however other approaches exist which are theoretically justified. For example a Bayesian approach might use the training data to determine either a maximum *posteriori* point estimate of the model parameters or their joint posterior distribution. In this case no monitoring data is used as the 'optimal' model is selected automatically. The generality of such a model of course still needs to be assessed with the test data set. For a recent review of model selection literature see http://www.modelselection.org. It is important to note that in all cases the test set is 'blind' to the model building and selection process. The test procedure involves applying the test data to the 'optimal' model. The subsequent model predictions are compared to known (blind) responses and a test statistic calculated (e.g., Q^2 statistic Eriksson et al. 2001). Such models are known to be extremely overoptimistic, especially for K = 1 (Golbraikh and Tropsha 2002).

For unsupervised methods such a PCA, the monitoring data set is not needed. However, care has to be taken in how results are presented in publications. For example, if

when using PCA no test set is applied then one cannot make bold claims about data clustering when plotting principal component y against principal component x when x and y are anything other than 1 and 2 respectively. Searching all possible component axes is a form of multiple testing (supervised analysis) and therefore requires the application of proper corrections to the clustering statistic (e.g., Bonferroni).

Suggesting the best method of model validation is beyond the scope of this document. Here we are concerned with the reporting of methodologies and results. Thus, we simply encourage authors to provide external (i.e., blind) measures of model generality and/or avoid making false claims of model (or test) generality beyond the scope of the data presented.

Proposed *minimum* information reported for multivariate analysis

- Description of workflow.
- Details of the algorithm used must be explicitly given and include the software package and version number or date employed for computation included. For established algorithms, one should reference the published literature but provide the meta-parameters. *State-of-the-art*, *radical* or *unconventional* algorithms should in general be avoided if they have not been independently assessed and approved (by peer review) and published in a methodology paper.
- Details of how data are split into training and external validation sets, including, as far as possible, proof that both sets are equally representative of the sample space/data distribution (with respect to meta data).
- Details of how internal validation/meta-parameter optimization is performed.
- Details about the chosen metric for assessing the predictive ability of the model (supervised methods only).
- Prediction scores for both training and external test sets.
- Details on data interpretation and visualization.

Additional *recommended* information reported for Multivariate analysis:

- Descriptive statistics about model prediction to accompany the prediction scores. For example, in binary response models ('case'/'control') most univariate tests can be used for continuous prediction scores (produced by PLS and LDA etc). Alternatively, present the confusion matrices of binary outputs. For calibration models plots of predicted *vs*. expected are useful for readers.
- If many different model types are tested then a summary of all results is recommended.

11 Discussion and conclusions

This document proposes the minimum level of reporting required in order to account accurately for any chemometric/statistical methods supporting the conclusions of metabolomics studies. By using these standards the metabolomics community as a whole will benefit by the subsequent dissemination of clear concise facts. Included here is a scheme (proposed reporting vocabulary) of terms that will help remove some of the confusion currently noted when comparing/reproducing various studies. Whilst we cannot be prescriptive on the exact mechanism by which these are reported, most data analyses start with the production of the initial data matrix which will then be analyzed. It is worth considering that the reporting structure should be split into distinct sections. The first aggregates all the steps that were taken to turn the raw analytical data into the initial data matrix (each row being one sample and each column being one feature). The next part should then describe the levels of pre-processing used to clean and prepare the data for the main modeling process. The final stage is the analysis the clean (transformed) data matrix. The definition of workflows at each of these stages is key to allowing the reporting standard to be modular and flexible. Each stage may include several sub-workflows in series (or parallel).

In conclusion, The Data Analysis Working Group (DAWG) will continue to update this consensus document that describes a minimum core set of necessary data related to the data analyses associated with metabolomics experiments. Further, the DAWG will work cooperatively with other MSI groups to build an integrated standard. The primary motivation is to establish acceptable practices that will maximize the utility, validity, and understanding of metabolomics data. To achieve this objective we actively seek input from specialists as well as generalists from the metabolomics community. Only through active community involvement will a functional solution be achieved. To this end, we would strongly encourage all authors to make available electronically, and as a condition of publication, all data used in the making of a claim that the value of a metabolic or spectroscopic biomarker is significantly different between two classes. This is the easiest way to allow the community swiftly to validate such claims (or otherwise), and was instrumental in the rapid unmasking of artefacts in some well-publicized proteomic biomarker experiments (Baggerly et al., 2004).

Acknowledgements D.B. wishes to thank the UK BBSRC and UK MRC. R.G. wishes to thank the UK BBSRC and EC META-PHOR (Food-CT-2006-03622) for financial support. BSK acknowledges NIH support. JT acknowledges Swedish Research Council for financial support. Disclaimer: The views presented in this article do not necessarily reflect those of the U. S. Food and Drug Administration.

References

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (2000). *Statistics with confidence* (2nd ed.). Blackwell Publishers.

Andreev, V. P., Rejtar, T., Chen, H. S., Moskovets, E. V., Ivanov, A. R., & Karger, B. L. (2003). A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Analytical Chemistry, 75*, 6314–6326.

Baggerly, K. A., Morris, J. S., & Coombes, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: Comparing datasets from different experiments. *Bioinformatics, 20*, 777–785.

Beebe, K. R., Pell, R. J., & Seasholtz, M. B. (1998). *Chemometrics: A practical guide*. New York: Wiley.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Bland, M. (2000). *An introduction to medical statistics*. Oxford: Oxford University Press.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*, 199–215.

Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics, 17*, 16–33.

Broadhurst, D., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics, 2*, 171–196.

Brown, M., Dunn, W. B., Ellis, D. I., Goodacre, R., Handl, J., Knowles, J. D., O'Hagan, S., Spasic, I., & Kell, D. B. (2005). A metabolome pipeline: From concept to data to knowledge. *Metabolomics, 1*, 39–51.

Campbell, T., Blasko, J., Crawford, E. D., Forman, J., Hanks, G., Kuban, D., Montie, J., Moul, J., Pollack, A., Raghavan, D., Ray, P., Roach, M., Steinberg, G., Stone, N., Thompson, I., Vogelzang, N., & Vijayakumar, S. (2001). Clinical staging of prostate cancer: Reproducibility and clarification of issues. *International Journal of Cancer, 96*, 198–209.

Chatfield, C., & Collins, A. J. (1980). *Introduction to multivariate analysis*. London: Chapman & Hall.

Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Jane Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., Holmes, E., & Nicholson, J. K. (2005a). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Analytical Chemistry, 77*, 1282–1289.

Cloarec, O., Dumas, M. E., Trygg, J., Craig, A., Barton, R. H., Lindon, J. C., Nicholson, J. K., & Holmes, E. (2005b). Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^{1}H NMR spectroscopic metabonomic studies. *Analytical Chemistry, 77*, 517–526.

Crockford, D. J., Holmes, E., Lindon, J. C., Plumb, R. S., Zirah, S., Bruce, S. J., Rainville, P., Stumpf, C. L., & Nicholson, J. K. (2006). Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabonomic toxicology studies. *Analytical Chemistry, 78*, 363–371.

Duda, R. O., Hart, P. E., & Stork, D. E. (2001). *Pattern classification* (2nd ed.) London: John Wiley.

Ebbels, T. M. D., Buxton, B. F., & Jones, D. T. (2006). *SpringScape*: Visualisation of microarray and contextual bioinformatic data using spring embedding an 'information landscape'. *Bioinformatics, 22*, e99–e108.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37*, 36–48.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., & Wold, S. (2001). *Multi- and megavariate data analysis*. Umea, Sweden: Umetrics AB.

Everitt, B. S. (1993). *Cluster analysis*. London: Edward Arnold.

Forshed, J., Torgrip, R. J., Aberg, K. M., Karlberg, B., Lindberg, J., & Jacobsson, S. P. (2005). A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *Journal of Pharmaceutical and Biomedical Analysis, 38*, 824–832.

Golbraikh, A., & Tropsha, A. (2002). Beware of q2!. *Journal of Molecular Graphics & Modelling, 20*, 269–276.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers—acquiring and understanding global metabolite data. *Trends in Biotechnology, 22*, 245–252.

Hall, R. D. (2006). Plant metabolomics: From holistic hope, to hype, to hot topic. *New Phytologist, 169*, 453–468.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21*, 3201–3212.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. Berlin: Springer-Verlag.

Holmes, E., Bonner, F. W., Sweatman, B. C., Lindon, J. C., Beddell, C. R., Rahr, E., & Nicholson, J. K. (1992). Nuclear magnetic resonance spectroscopy and pattern recognition analysis of the biochemical processes associated with the progression of and recovery from nephrotoxic lesions in the rat induced by mercury(II) chloride and 2-bromoethanamine. *Molecular Pharmacology, 42*, 922–930.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med* 2, e124.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Jonsson, P., Gullberg, J., Nordström, A., Kusano, M., Kowalczyk, M., Sjöström, M., & Moritz, T. (2004). A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Analytical Chemistry, 76*, 1738–1745.

Katajamaa, M., & Oresic, M. (2005). Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics, 6*, Art. No. 179.

Kell, D. B. (2002). Genotype-phenotype mapping: Genes as computer programs. *Trends in Genetics, 18*, 555–559.

Kell, D. B. (2007). Metabolomic biomarkers: Search, discovery and validation. *Expert Review in Molecular Diagnostics, 7*, 329–333.

Keun, H. C., Ebbels, T. M., Bollard, M. E., Beckonert, O., Antti, H., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2004). Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chemical Research in Toxicology, 17*, 579–587.

Kohonen, T. (1989). *Self-organization and associative memory*. Berlin: Springer-Verlag.

Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection*. Cambridge, MA: MIT Press.

Krzanowski, W. J. (1988). *Principles of multivariate analysis: A user's perspective*. Oxford: Oxford Univeristy Press.

Kvalheim, O. M., & Liang, Y. Z. (1992). Heuristic evolving latent projections: Resolving two-way multicomponent data. 1. Selectivity, latent-projective graph, datascope, local rank, and unique resolution. *Analytical Chemistry, 64*, 936–946.

Langdon, W. B. (1998). *Genetic programming and data structures: Genetic programming + data structures = automatic programming!* Boston: Kluwer.

Lavine, B. K. (1998). Chemometrics. *Analytical Chemistry, 70*, R209–R228.

Manly, B. F. J. (1994). *Multivariate statistical methods: A primer*. London: Chapman & Hall.

Martens, H., & Næs, T. (1989). *Multivariate calibration*. Chichester: John Wiley.

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., DeJong, S., Lewi, P. J., & Smeyers-Verbeke, J. (1997). *Handbook of chemometrics and qualimetrics: Part A*. Amsterdam: Elsevier.

Miccheli, A. T., Miccheli, A., Di Clemente, R., Valerio, M., Coluccia, P., Bizzarri, M., & Conti, F. (2006). NMR-based metabolic profiling of human hepatoma cells in relation to cell growth by culture media analysis. *Biochimica et Biophysica Acta, 1760*, 1723–1731.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.) Chichester: Wiley.

Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica, 29*, 1181–1189.

Paolucci, U., Vigneau Callahan, K. E., Shi, H., Matson, W. R., & Kristal, B. S. (2004a). Development of biomarkers based on diet-dependent metabolic serotypes: Characteristics of component-based models of metabolic serotype. *Omics, 8*, 221–238.

Paolucci, U., Vigneau Callahan, K. E., Shi, H., Matson, W. R., & Kristal, B. S. (2004b). Development of biomarkers based on diet-dependent metabolic serotypes: Concerns and approaches for cohort and gender issues in serum metabolome studies. *Omics, 8*, 209–220.

Provencher, S. W. (1993). Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magnetic Resonance in Medicine, 30*, 672–679.

Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews Cancer, 5*, 142–149.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.) Philadelphia: Lippincott, Williams & Wilkins.

Rubbin, D. B., & Little, R. J. A. (1987). *Statistical analysis with missing data*. New York: Wiley.

Sabatine, M. S., Liu, E., Morrow, D. A., Heller, E., McCarroll, R., Wiegand, R., Berriz, G. F., Roth, F. P., & Gerszten, R. E. (2005). Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation, 112*, 3868–3875.

Serrai, H., Nadal, L., Leray, G., Leroy, B., Delplanque, B., & de Certaines, J. D. (1998). Quantification of plasma lipoprotein fractions by wavelet transform time-domain data processing of the proton nuclear magnetic resonance methylene spectral region. *NMR in Biomedicine, 11*, 273–280.

Skov, T., van den Berg, F., Tomasi, G., & Bro, R. (2007). Automated alignment of chromatographic data. *Journal of Chemometrics, 20*, 484–497.

Sumner, L.W., Amberg, A., Barrett, B., Beger, R., Beale, M.H., Daykin, C., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hardy, N., Higashi, R., Kopka, J., Lindon, J. C., Lane, A. N., Marriott, P., Nicholls, A. W., Reily, M. D., & Viant, M. (2007). Proposed minimum reporting standards for Chemical analysis. *Metabolomics, 3*, in this issue.

Tauler, R., Durand, G., & Barcelo, D. (1992). Deconvolution and quantitation of unresolved mixtures in liquid-chromatographic—diode-array detection using evolving factor-analysis. *Chromatographia, 33*, 244–254.

Tomasi, G., van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as pre-processing methods for chromatographic data. *Journal of Chemometrics, 18*, 231–241.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures, O-PLS. *Journal of Chemometrics, 16*, 119–128.

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics, 7*, 142.

Veldhuis, J. D., Carlson, M. L., & Johnson, M. L. (1987). The pituitary-gland secretes in bursts—appraising the nature of glandular secretory impulses by simultaneous multiple-parameter deconvolution of plasma-hormone concentrations. *PNAS, 84*, 7686–7690.

Vogels, J. T. W. E., Tas, A. C., Venekamp, J., & Van Der Greef, J. (1996). Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. *Journal of Chemometrics, 10*, 425–438.

Wasserman, P. D. (1989). *Neural computing: Theory and practice*. New York: Van Nostrand Reinhold.

Weckwerth, W., & Morgenthal, K. (2005). Metabolomics: From pattern recognition to biological interpretation. *Drug Discovery Today, 10*, 1551–1558.

Weljie, A. M., Newton, J., Mercier, P. M., Carlson, E., & Slupsky, C. M. (2006). Targeted profiling: quantitative analysis of ^{1}H-NMR metabolomics data. *Analytical Chemistry, 78*(13), 4430–4442.

Windig, W., Phalp, J. M., & Payne, A. W. (1996). A noise and background reduction method for component detection in liquid chromatography mass spectrometry. *Analytical Chemistry, 68*, 3602–3606.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*, 67–82.