Metabolomics (2012) 8:S161–S174
DOI 10.1007/s11306-011-0366-4

ORIGINAL ARTICLE

Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline

Olga Hrydziuszko · Mark R. Viant

Received: 23 April 2011 / Accepted: 22 September 2011 / Published online: 8 October 2011

Abstract Missing values in mass spectrometry metabolomic datasets occur widely and can originate from a number of sources, including for both technical and biological reasons. Currently, little is known about these data, i.e. about their distributions across datasets, the need (or not) to consider them in the data processing pipeline, and most importantly, the optimal way of assigning them values prior to univariate or multivariate data analysis. Here, we address all of these issues using direct infusion Fourier transform ion cyclotron resonance mass spectrometry data. We have shown that missing data are widespread, accounting for ca. 20% of data and affecting up to 80% of all variables, and that they do not occur randomly but rather as a function of signal intensity and mass-to-charge ratio. We have demonstrated that missing data estimation algorithms have a major effect on the outcome of data analysis when comparing the differences between biological sample groups, including by t test, ANOVA and principal component analysis. Furthermore, results varied significantly across the eight algorithms that we assessed for their ability to impute known, but labelled as missing, entries. Based on all of our findings we identified the k-nearest neighbour imputation method (*KNN*) as the optimal missing value estimation approach for our direct infusion mass spectrometry datasets. However, we believe the wider significance of this study is that it highlights the importance of missing metabolite levels in the data processing pipeline and offers an approach to identify optimal ways of treating missing data in metabolomics experiments.

Keywords FT-ICR · Metabolic profiling · Missing data · Missing entries · Signal processing

1 Introduction

Many questions addressed using metabolomic approaches are similar to those being asked in transcriptomic and/or proteomic investigations, e.g. which metabolites, genes and/or proteins differ significantly between biological groups under considerations such as healthy versus diseased, or control versus drug-treated samples (Defamie 2008). Moreover, for many ‘omics experiments the final data formats (after instrument specific processing) are alike, with a rectangular matrix containing gene expression values or metabolite relative abundances, and organised with each variable measured in a unique column and each sample analysed in a unique row. This consistency of data format has facilitated the use of the same or similar univariate and multivariate statistical methods (including computational data analysis or pattern recognition methods) in metabolomics as are used in other ‘omics approaches (Goodacre et al. 2004). However, while in other ‘omics fields there has and continues to be considerable interest in understanding and developing appropriate techniques to handle missing data (prior to statistical analysis), it has received minimal attention in

Electronic supplementary material The online version of this article (doi:10.1007/s11306-011-0366-4) contains supplementary material, which is available to authorized users.

O. Hrydziuszko · M. R. Viant
Centre for Systems Biology, University of Birmingham,
Edgbaston, Birmingham B15 2TT, UK
e-mail: oxh732@bham.ac.uk

M. R. Viant (✉)
School of Biosciences, University of Birmingham, Edgbaston,
Birmingham B15 2TT, UK
e-mail: M.Viant@bham.ac.uk

metabolomics. Missing values (also referred to as missing data or missing entries) may arise in metabolomics experiments for a number of reasons. In the case of direct infusion Fourier transform ion cyclotron resonance (DI FT-ICR) mass spectrometry (MS) based metabolomics, they could have a biological and/or technical origin. A metabolite abundance value for a specific sample may not be available when several samples are analysed and then all the measurements are compiled into a data matrix for further comparison or analysis. For some samples a specific peak may not be present for genuine biological reasons, e.g. due to heterogeneity between samples, or in other cases its abundance is below the detection limit of the mass spectrometer, or alternatively it was not measured properly owing to a technical problem such as a temporary reduction in electrospray performance due to particulate material in the spray nozzle (Payne et al. 2009).

Metabolomics researches often face such problems with missing data. These problems can be (and have been) addressed by (a) simply disregarding all the variables for which missing data are present (Xia et al. 2009), (b) using data analysis methods (including univariate and multivariate) that have been shown to be able to handle some proportion of missing data (Kenny et al. 2010; Blanchet et al. 2011), or (c) estimating missing data with various imputation algorithms (Kenny et al. 2010; Xia et al. 2009). Focusing on the part of the data for which all the measurements are present could be an optimal solution when only a small proportion of variables are affected by missing data, however, this is typically not the case in most metabolomics experiments. Some statistical software packages allow univariate statistical testing (e.g. t test or ANOVA in R or Matlab) on samples with missing data by simply disregarding the missing entries, and some multivariate exploratory or predictive approaches have been developed to handle some amount of missing data [e.g. missing data Principal Component Analysis or Partial Least Squares Discriminant Analysis (Walczak and Massart 2001; Andersson and Bro 1998)]. While this strategy may be appropriate when dealing with large sample size studies that contain few missing data, it may be problematic for metabolomics studies in which the sample size is often limited and thus ignoring missing data could diminish the power of the statistical tests. Furthermore, it is not uncommon that missing data occurs predominantly in one biological group (at least for the case of DI FT-ICR MS metabolomics) and when combined with a small sample size this can lead to an insufficient number of metabolites measured for this approach (e.g. at least two detected measurements per biological group are needed to perform a t test in the R or Matlab environments). Hence, imputing missing data prior to data analysis represents a practical solution in applied metabolomics since (i) it yields a simple, consistent, rapid and automated data processing pipeline, (ii) the resulting data matrix is compatible with a very wide array of univariate or multivariate analyses, (iii) this approach facilitates a comparison of univariate and multivariate statistical results for specific metabolites of interest (e.g. biomarkers), and (iv) it provides a complete profile of metabolite concentrations that can be used in a consistent manner in other types of data analysis (e.g. integration with other 'omics datasets). The importance of appropriate handling of missing data has been recognised in the analysis of DNA microarray (Troyanskaya et al. 2001) and gel-based proteomics data (Albrecht et al. 2010; Pedreschi et al. 2008). For example, studies have been reported on how missing values affect statistical parameter estimations (Troyanskaya et al. 2001), how they influence the results of univariate (Scheel et al. 2005; de Brevern et al. 2004) and multivariate data analysis (Pedreschi et al. 2008), what is the optimal method of their imputation (Jornsten et al. 2005; Kim et al. 2004; Scheel et al. 2005; Tuikkala et al. 2008), and how to develop robust data analysis algorithms for datasets with significant amounts of missing entries (Kim et al. 2007). In metabolomics, none of the above questions has yet been thoroughly addressed. This is particularly surprisingly given that mass spectrometry based metabolomic analyses (e.g., DI FT-ICR MS) typically generate datasets with considerable amounts of missing data (Southam et al. 2007; Taylor et al. 2009). Furthermore, it has been suggested that missing values do not affect the data analysis outcome, but that their treatment (i.e. deletion or estimation) is carried out only for computational convenience (Steuer et al. 2007). Current methods for handling missing data in metabolomics involve simple methods such as replacing a missing value by the mean or median of the available measurements for that variable, replacing with some small arbitrary number, or k-nearest neighbour imputation [Steuer et al. 2007; GeneSpring MS software (Alignment Technologies)]. A quite different approach, reported by Sangster et al. (2007), estimates missing values by returning to the raw spectral data and integrating the areas of the missing peaks which are below the applied signal-to-noise ratio (SNR) threshold, but in close proximity to the peaks' known *m/z* value.

Here, we analyse missing data in the context of DI FT-ICR MS based metabolomics measurements, but with the findings of our analyses potentially transferable and of importance for other metabolomics studies. We investigate not only the nature of the missing data but also their effects on data analysis, both univariate and multivariate. Specifically we addressed the following questions: what are the potential origins of missing data in metabolomic datasets? Do they appear at random, or as a function of peak intensity and/or *m/z* value? Do they affect the outcome of commonly used univariate and multivariate data analyses? And if so, what is the optimal method of replacing their values as part of a consistent and automated data processing pipeline that will provide the metabolomics researcher with a complete data matrix that is compatible with many univariate and multivariate statistical analyses or other data mining algorithms? With more than a dozen imputation methods available and published (not limited to ‘omics studies), we focus our investigations around the eight commonly used and reported methods in applied ‘omics studies that are readily implementable (by other researchers) in the R computing environment, ultimately providing a (potentially expandable) benchmark for the questions above. Finally, to maximise the generality of our findings, we have investigated three widely differing biological datasets (including cellular, tissue and whole organism extracts from in vitro and in vivo experimentation) that were measured in positive and/or negative ion mode FT-ICR MS.

2 Materials and methods

2.1 Mass spectrometry datasets

The three FT-ICR MS datasets used here comprise of (1) CCL—cancer cell line, specifically an acute myeloid leukaemia cell line (K562) cultured and treated under hypoxic conditions, comprising of six control samples, six samples exposed to indomethacin (non-steroidal anti-inflammatory drug) and six samples treated with medroxyprogesterone acetate (component of hormonal contraceptives), all measured in positive (CCL_p) and negative ion mode (CCL_n); (2) DM—*Daphnia magna* (a freshwater invertebrate) exposed for 24 h to 1.5 mg/l of 2,4-dinitrophenol, a cellular metabolic toxicant which obstructs oxidative phosphorylation, comprising of 10 control and 10 exposed samples measured in negative ion mode only (Taylor et al. 2010); (3) HL—human liver biopsies taken throughout orthotopic liver transplantation, comprising seven biopsies taken soon after organ retrieval and seven further biopsies taken post-reperfusion of blood circulation in the recipient patient, measured in positive ion mode only (Hrydziuszko et al. 2010). All cell, tissue or whole organism samples were extracted using a methanol/chloroform/water method (Wu et al. 2008), and the polar metabolites were analysed using a hybrid 7-T direct infusion nanoelectrospray FT-ICR mass spectrometer (Thermo Fisher Scientific LTQ FT) over the range *m/z* 70–500. All CCL and DM samples were analysed in triplicate, and the HL samples in duplicate; these represent technical replicates of each sample for use in the subsequent noise filtering algorithm. Spectra were processed as described previously (Taylor et al. 2009), including a 3-step filtering algorithm to eliminate noise peaks (Payne et al. 2009). Specifically, the first filtering step comprised of a hard SNR threshold, below which peaks were rejected (2.5 for CCL datasets and 3.5 for DM and HL datasets). In the second step, only peaks present in two out of three technical replicates (for CCL and DM datasets) and two out of two replicates (HL dataset) were retained, and the intensities averaged to create a single spectrum per biological sample. In the third step, only peaks present in at least 50% of the samples were retained (specifically across all samples for each of the DM and HL datasets, and across samples within each biological group for the CCL datasets; Table 1). Probabilistic quotient normalisation was then performed on all of the datasets (prior to univariate and multivariate analyses) (Dieterle et al. 2006) followed by the generalised log transformation (prior to multivariate analysis only, in order to stabilise the variance across the peaks and avoid the highest abundance peaks dominating in the multivariate analyses) (Parsons et al. 2007). Individual peak intensities were confirmed to follow normal distributions as tested with the Shapiro–Wilk normality test (for >99% of the peaks that have no missing data and that have been measured in at least three samples; this is in agreement with our previous unpublished observations for other (including larger sample size) metabolomics datasets obtained via DI FT-ICR MS). At this stage of analysis, each dataset contained *m* peaks and *n* samples with multiple missing values (Table 1). The median of the coefficients of variation of the peak intensities, reflecting biological diversity within each dataset (Parsons et al. 2009), was 17.21, 20.24, 25.59 and 60.99% for CCL_n, CCL_p, DM and HL, respectively (excluding peaks with missing data) confirming, as expected, that the metabolic heterogeneity increased from cell line extracts to laboratory cultured organisms to clinical samples.

2.2 Occurrence and distribution patterns of missing data

The properties of the missing values in the FT-ICR MS datasets were examined using two methods. First, the distribution of the missing values across each dataset was determined to be ‘missing completely at random’ (MCAR) or not. This employed Little’s test of MCAR for multivariate data with missing values (Little 1998). Second, their occurrence patterns were assessed using Pearson’s correlation between missing data properties and the dataset features, specifically the amount of missing data versus both the abundances and *m/z* values of the non-missing data peaks.

Table 1 List of the DI FT-ICR MS based metabolomic datasets analysed together with some of their basic properties

Dataset	Brief description	Median of coefficient of variation [%]	No. of samples	No. of groups	No. of peaks	Missing values [%]	Peaks with missing values [%]
CCL_n	Human cancer cell line K562, negative ion mode	17.21	18	3	6770	22.01	51.67
CCL_p	Human cancer cell line K562, positive ion mode	20.24	18	3	4426	28.53	64.96
DM	*Daphnia magna* exposed for 24 h to dinitrophenol, negative ion mode	25.59	20	2	4196	14.63	55.22
HL	Human liver tissue prior and post liver transplantation, positive ion mode	60.99	14	2	1805	23.66	78.73

2.3 Impact of missing data imputation on statistical analyses

We then compared eight common and/or readily available missing data imputation (or estimation) methods in terms of their impact on univariate and multivariate data analysis as well as in terms of their performance for handling missing values (experimental design summarised in Fig. 1). This was performed on all three FT-ICR MS datasets, as described below. Specifically, the eight estimation methods comprised: (1) *S*—substitution of missing values with a small predefined value (e.g. 0.01) (as used in GeneSpring MS software; Alignment Technologies); (2) *HM*—substitution with half of the minimum value found in the non-missing data (Xia et al. 2009); (3) *M*—substitution with the mean of the non-missing values across all samples for that peak (Steuer et al. 2007); (4) *MED*—substitution with the median of the non-missing values across all samples (Steuer et al. 2007); (5) *KNN*—weighted k-nearest neighbour algorithm in which *k* (here $k = 5$; different *k* values did not significantly affect our analysis, data not shown) metabolites most similar in terms of their intensity profiles across all samples are identified based on the Euclidean distance similarity measure to the metabolite having a missing datum for a given sample; the missing datum is then estimated as the weighted average of the *k* metabolites for that sample with their contribution weighted by their similarity (Steuer et al. 2007; Troyanskaya et al. 2001); (6) *BPCA*—Bayesian PCA missing value estimation, a three stage algorithm based on principal component regression, Bayesian estimation and the expectation–maximisation repetitive algorithm; briefly during the principal component regression the missing data of a metabolite's intensity profile are estimated from the observed values using the PCA result, followed by a Bayesian estimation in which residual error and the projection of metabolites on the principal components are considered as normal independent variables with unknown parameters which are inferred in the final expectation–maximization algorithm step (Xia et al. 2009; Oba et al. 2003); (7) *MI*—multivariate imputation by chained equations (van Buuren and Groothuis-Oudsshoorn 2010); (8) *REP*—modified version of Sangster's method as used by us previously, for which a missing value is substituted with the average intensity of the nearest (in term of *m/z* value) peaks from the raw measurements of the technical replicates (Sangster et al. 2007). Methods *S* and *HM* substitute the missing values with a relatively small value and act on the assumption that missing data do not influence the outcome of the subsequent data analysis due to a low amount of missing data. Methods *M* and *MED* impute missing data using a row mean or median and assume that the metabolite's intensity is similar across all the experiments (i.e. samples). Furthermore, along with the *S* and *HM* methods, *M* and *MED* do not use the information contained in the structure of the data. *KNN* searches for the *k* metabolites (or more specifically peaks in the mass spectra) that have similar measured signal intensities across the biological samples as compared to the peak for which the missing entry is present. The missing value is then replaced with the weighted average of the corresponding non-missing values from the group of *k* peaks that were identified as most similar. *BPCA* and *MI* methods use the global structure of the dataset, in a way that all the metabolites are taken into consideration to obtain the imputed value. *BPCA* estimates the missing data in a three stage process starting with principal component regression, followed by Bayesian estimation and finishing with an expectation–maximisation like repetitive algorithm; this approach has been shown to outperform *KNN* for gene expression data (cDNA microarrays) when the number of samples was large (>30) and the missing data occurred randomly (Oba et al. 2003; Albrecht et al. 2010). *MI*, multivariate imputation by chained equations (available in the R environment in a MICE library) is a method of multiple imputation in which each variable is estimated using a regression model conditional on all the other variables iteratively looping through all the variables with missing data; here we used the predictive mean matching implementation (Little and Rubin 2002), similar to the

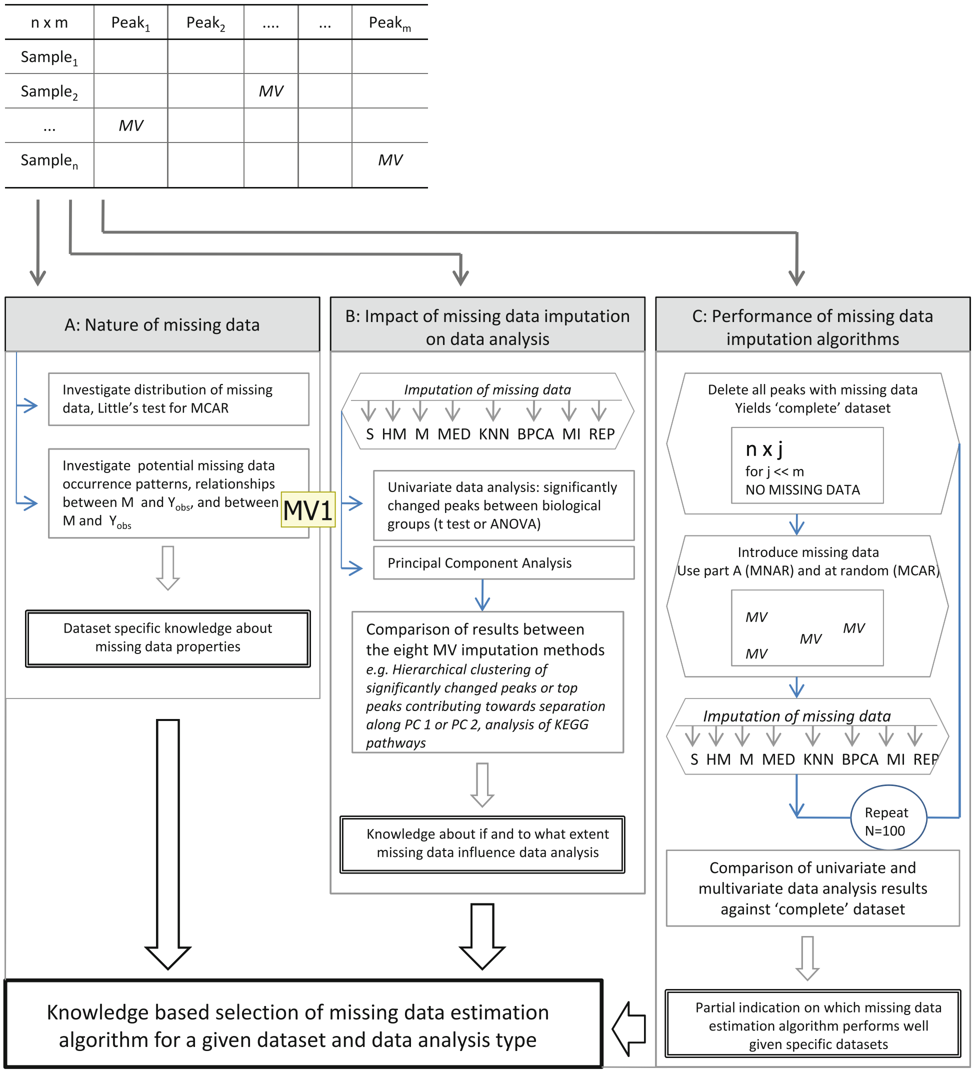
n x m
Peak1
Peak2
....
...
Peakm
Sample1
Sample2
MV
...
MV
Samplen
MV
A: Nature of missing data
Investigate distribution of missing data, Little's test for MCAR
Investigate potential missing data occurrence patterns, relationships between M and Yobs, and between M and Yobs
MV1
Dataset specific knowledge about missing data properties
B: Impact of missing data imputation on data analysis
Imputation of missing data
S HM M MED KNN BPCA MI REP
Univariate data analysis: significantly changed peaks between biological groups (t test or ANOVA)
Principal Component Analysis
Comparison of results between the eight MV imputation methods
e.g. Hierarchical clustering of significantly changed peaks or top peaks contributing towards separation along PC 1 or PC 2, analysis of KEGG pathways
Knowledge about if and to what extent missing data influence data analysis
C: Performance of missing data imputation algorithms
Delete all peaks with missing data Yields 'complete' dataset
n x j
for j << m
NO MISSING DATA
Introduce missing data
Use part A (MNAR) and at random (MCAR)
MV
MV
MV
MV
Imputation of missing data
S HM M MED KNN BPCA MI REP
Repeat N=100
Comparison of univariate and multivariate data analysis results against 'complete' dataset
Partial indication on which missing data estimation algorithm performs well given specific datasets
Knowledge based selection of missing data estimation algorithm for a given dataset and data analysis type

Fig. 1 Flow chart summarising the analyses of missing values (MV) performed in this study

regression method. Method *REP* attempts to utilise peak abundance information captured in the technical replicates that lies beneath the SNR threshold by estimating missing data via the average of the closest (in terms of *m/z*) peaks in each of the three (two for HL dataset) technical replicates.

After imputing the missing values for the three FT-ICR MS datasets, using all eight methods, statistical tests were employed to determine the metabolic differences between the sample classes (e.g. between the control and two drug-treated groups in the CCL study). This allowed us to examine the impact of each imputation method on finding significantly changed peaks via univariate testing [t-test or ANOVA between groups with Benjamini and Hochberg correction for multiple testing (Benjamini and Hochberg 1995)] as well as via multivariate principal component analysis (PCA) scores and loadings values. This approach was chosen as it is routinely used in metabolomics to provide an initial unsupervised explorative analysis and because it is appropriate for the sizes of the datasets investigated here (containing thousands of peaks and only up to ten samples per biological group); supervised methods such as partial least squares discriminate analysis would likely lead to overfitting in the modelling (Broadhurst and Kell 2006; Westerhuis et al. 2008). The eight missing value estimation methods were further compared in terms of the outcome of both univariate and multivariate analysis via hierarchical clustering, in which the Euclidean distance was calculated for the groups of peaks significantly changed between biological groups for the univariate analysis and for the top 5% of peaks contributing to the separation along the first and the second principal components (PC1 and PC2) based on their loading values. Since for the latter the order (i.e. ranking based on loading values) of these peaks holds important information about the PCA results, we developed a measure referred to as R_t that compares the number of shared peaks between any two imputation methods as well as their rank order; i.e. in addition to calculating the amount of overlap in the top 5% of peaks between two imputation methods, we assign a higher 'similarity' value for methods for which the top peaks are in the same order (see electronic supplementary material, 'Impact of missing data imputation on multivariate data analysis').

In addition, to assess the impact on the final biochemical interpretation of the data, we have compared the eight methods in terms of detecting significantly 'active' KEGG (Kyoto Encyclopedia of Genes and Genomes) human pathways (Kanehisa et al. 2008). Here we defined a significantly 'active' pathway as follows: for significantly changed peaks between groups (univariate analysis) or the top 5% of peaks contributing towards separation along PC1 or PC2, we assigned one (or more) putative metabolite names (Sumner et al. 2007) to each *m/z* value, based upon accurate mass measurements and the KEGG database, taking into account commonly detected ion forms ($[M - e]^+$, $[M + H]^+$, $[M + Na]^+$, $[M + {}^{39}K]^+$, $[M + 2Na - H]^+$, $[M + 2^{39}K - H]^+$ for positive ion mode and $[M + e]^-$, $[M - H]^-$, $[M + {}^{35}Cl]^-$, $[M + {}^{37}Cl]^-$, $[M + HAc - H]^-$ (HAc, acetic acid) for negative ion mode). Following that, for each putatively identified metabolite, we listed all KEGG pathways for which it is involved. The probability that a peak *i* belongs to a pathway *j* was calculated via conditional probability $P_i(\text{pathway}_j|\text{peak}_i)$ which equals to the number of putative metabolite assignments of peak_i that are involved in pathway_j divided by the number of putative assignments for peak_i. We marked pathways as significantly 'active' if for at least one of the peaks they were observed with a probability greater than or equal to 0.75.

2.4 Performance of missing data estimation algorithms

To assess the performance of the missing data imputation algorithms, we used 'complete' CCL_n, $CCL_{p,}$ DM and HL datasets that were created by excluding all peaks that contained missing values. Next, we deliberately introduced missing values, either completely at random (MCAR) or not at random (MNAR) (Little and Rubin 2002), generating datasets with missing data entries but for which we knew the real (original) values. These missing data were again imputed using each of the eight methods described above (for *REP* we estimated the values of the deliberately introduced missing entries using the information from the triplicate technical measurements, explained below). The imputed missing values (for all eight methods) were compared to the original deliberately excluded values in terms of normalised root mean square errors (NRMSE) (Troyanskaya et al. 2001) and also the outcome of univariate and multivariate (PCA) data analysis. The whole procedure (both for MCAR and MNAR) was repeated 100 times. Furthermore, the amount of missing data generated was ca. 20% for each 'complete' dataset, similar to the amount of real missing data in the original datasets. For MNAR, missing values were introduced to mirror the missing data properties that we discovered in the original datasets, mainly to capture their relationship with the signal intensity and *m/z* ratio (discussed further below).

3 Results and discussion

3.1 Occurrence and distribution patterns of missing data in DI FT-ICR MS metabolomics

A typical DI FT-ICR MS based metabolomics dataset measured and processed as described above contains ca. 20% of missing data (Table 1), which equates to up to 80% of peaks having at least one missing value across the analysed samples. Little's MCAR test revealed that this missing data does not occur completely at random ($p = 0.029$, 0.032, 0.021 and 0.045 for CCL_n, CCL_p, DM and HL datasets respectively), i.e. they do not follow a random distribution. If $\boldsymbol{Y}$ denotes a rectangular dataset (e.g., with each peak's abundance in a unique column and each sample in a unique row) containing missing data, $\boldsymbol{Y}$ can be split into two subsets $\boldsymbol{Y_o}$ (observed values) and $\boldsymbol{Y_m}$ (missing data). A dummy matrix $\boldsymbol{D}$ could then indicate the location of the missing values in $\boldsymbol{Y}$, such that $d_{ij} = 1$ if y_{ij} is missing and $d_{ij} = 0$ if y_{ij} is present. Following Rubin's established categories for the mechanisms of missing data (Rubin 1976) for MCAR (missing completely at random; intuitively perceived as random), by definition, there is no relationship Φ between $\boldsymbol{D}$ and either $\boldsymbol{Y_m}$ or $\boldsymbol{Y_o}$, meaning that a pattern (occurrence) of missing data is not dependent on the values that are missing (e.g. all low intensity values in the dataset are missing) nor on the observed data (e.g. missing data occur for those metabolites with measured low intensity). For non-randomness, missing values may occur as MAR (missing at random, with a relationship Φ between $\boldsymbol{D}$ and $\boldsymbol{Y_o}$; counter-intuitive since it denotes one type of non-randomness) or MNAR (missing not at random, with a relationship Φ between $\boldsymbol{D}$ and $\boldsymbol{Y_m}$ and possibly $\boldsymbol{Y_o}$; describing the other type of non-randomness) (Rubin 1976; Little and Rubin 2002). For both the random and non-random scenarios, missing data may arise due to technical errors, biological factors, or a mixture of the two. The relationship Φ between $\boldsymbol{D}$ and $\boldsymbol{Y_m}$ is theoretical and therefore cannot be assessed due to the lack of information about the missing data. However, the analysis of the relationship between $\boldsymbol{D}$ and $\boldsymbol{Y_o}$ showed that missing data in FT-ICR mass spectra are a function of both the abundances of observed peaks as well as their *m/z* values. For the former, the lower the (mean) peak abundance the greater the amount of missing data that the peak contains (Pearson correlation coefficient of −0.80, −0.85, −0.89 and −0.90 for the CCL_n, CCL_p, DM and HL datasets respectively;

Fig. 2). For the latter, the analysis of the technical replicates showed that the probability of observing noisy peaks (i.e. missing in one or more of the technical replicates) is high for low *m/z* value signals, decreasing in probability for mid-range *m/z* values, and increasing again for high *m/z* signals (Fig. 2). This trend was quite apparent for both CCL and the DM datasets, with a small exception (i.e. no increase for the high *m/z* peaks) for the most biologically variable HL dataset. This effect was of unknown origin. We have confirmed that these relationships are not due to our data processing, i.e. the three stage noise filtering algorithm. Further investigation of the potential source(s) of this relationship is beyond the scope of our study, possibly arising from a technical peculiarity of the FT-ICR MS instrumentation.

Two relationships were discovered above, i.e. $\boldsymbol{D}$ being correlated significantly with non-missing value peak abundances ($\boldsymbol{Y_o}$) and also with *m/z* values. These relationships highlight important information about the occurrence and distribution of missing data that needs to be considered prior to their treatment. The large percentage of peaks containing missing data implies that simply removing them would dramatically reduce the size of the dataset (by ca. 80%). Furthermore the abundance relationship indicates that removing peaks with missing data, or inadequate estimation of missing entries, could result in a substantial bias since only peaks with the highest abundances would be kept, hindering subsequent biomarker discovery. We have demonstrated this latter effect while investigating the influence of the third step of the noise filtering algorithm (retaining only those peaks present in s% or more of the samples, and equivalent to discarding peaks with missing data) on the distribution of missing data across biological groups. Figure A1 (electronic supplementary material) shows the effects on the missing data distributions for s ≥ 0%, s ≥ 25%, s ≥ 50% and s ≥ 75% for the DM and HL datasets. The most interesting result was obtained for the HL dataset, for which there was a significant difference in the number of missing values between the cold-phase and post-reperfusion groups when all peaks were retained (s ≥ 0%), but this difference became non-significant for settings of s ≥ 25% and above. From our biochemical knowledge of the processes occurring during liver transplantation, we hypothesise that this is a case where peaks with missing data that genuinely carried important biological information were mistakenly removed (assumed to represent noise peaks), as it has been shown that during the cold phase liver metabolism ceases and it restarts upon reperfusion with an increased production of bile acids and urea (Hrydziuszko et al. 2010). Overall, our assessment of the missing data within DI FT-ICR MS datasets reveals that missing values do not occur completely at random but instead as a function of (at least) peak abundance and *m/z* value, and that peaks with missing values potentially carry importance biological information.

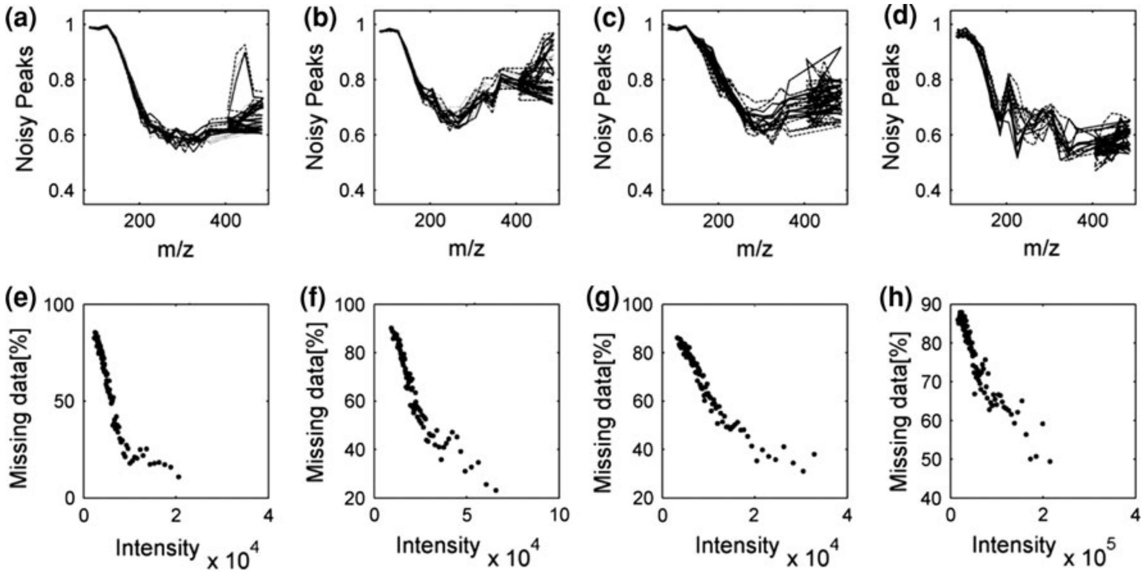

Fig. 2 Probability of the occurrence of noisy peaks as a function of *m/z* ratio for **a** $\mathrm{CCL_n}$, **b** $\mathrm{CCL_p}$, **c** DM and **d** HL datasets. Each sample in the CCL and DM datasets was measured as three technical replicates, and therefore noisy peaks are defined as occurring in one or two out of three measurements only. HL samples were measured as duplicates, with noisy peaks defined as occurring in one out of two measurements only. Percentage of missing data versus mean peak abundances (binned in 100 intervals with a sample filter ≥50%) for **e** $\mathrm{CCL_n}$, **f** $\mathrm{CCL_p}$, **g** DM and **h** HL datasets. The top 5% of peak abundances have been removed for plotting purposes

3.2 Impact of missing data imputation on univariate data analysis

The potential impacts of the eight missing data estimation techniques on the ability to discover significant differences in peak abundances between biological groups were evaluated using univariate statistical tests (either *t* tests or ANOVAs). Different imputation methods ultimately yielded quite diverse data analysis outcomes. Specifically, the number of peaks identified as significantly different between groups varied considerably between the eight estimation methods, from 2.65 to 14.70%, from 0.58 to 10.20%, from 7.44 to 14.20%, and from 1.72 to 14.24% for the CCL_n, CCL_p, DM and HL datasets, respectively (electronic supplementary material Table A1). As expected, based on the underlying mechanisms of the missing data estimation algorithms, some methods performed comparably, e.g. *S* and *HM*, and *M* and *MED* (electronic supplementary material Fig. A2, Table A2), while others were strongly dependent on the structure of the dataset, e.g. for the CCL datasets, *BPCA* and *REP* performed similar to *M* and *MED*, while for the DM and HL datasets, *REP* resembled *S* and *HM*.

Further investigation of the peaks detected as significantly different between biological groups showed that substantial proportions of these peaks were comprised of those which initially had missing data. Specifically, for the CCL_n, CCL_p, DM and HL datasets, the minimum percentage occurrence of this type of significant peak (from across all eight estimation methods) was 24.73% (for the *BPCA* method), 15.38% (*BPCA*), 23.72% (*M*) and 32.26% (*MED*), respectively (electronic supplementary material Table A1). The maximum percentages of significant peaks (which originally had missing values) were surprisingly high, at 72.62% (*KNN*), 83.66% (*HM*), 49.66% (*HM*) and 85.99% (*S*) for the same four datasets, respectively. In the worst case scenario of inadequate imputation of missing data entries, these minimum and maximum percentages provide an estimate of the false positive error rate associated with the arguably critical identification of significantly changing peaks. This error rate would most likely be in the upper range of percentages for methods such as *S* and *HM* if the missing data were to represent high abundance metabolites, and also for *M* and *MED* if the opposite were true with missing entries representing low abundance metabolites. This is further visualised in Figures A3 and A4 (electronic supplementary material) that show the distribution of the number of missing data (prior to imputation) that occur specifically within the significantly changing peaks only.

Having analysed the percentages of significantly changing peaks between biological groups that initially had missing data, we then investigated which of the samples originally had these missing values. Interestingly, the results showed that missing entries tend to be located in one of the biological groups, rather than being spread equally across all the groups (electronic supplementary material Table A3). This is a further important observation that helps to verify our earlier hypothesis that missing data may in fact represent true differences between biological groups, and therefore their accurate imputation is of considerable importance. Also, as above, it demonstrates a potential danger when using substitution (*S* and *HM*) or simple imputation (*M* and *MED*) methods. For the case of low peak abundances, it does not mean that a small arbitrarily chosen value would represent the missing data accurately; also, for high peak abundances, it should not be assumed that the mean or median of the non-missing values would represent an optimally imputed value. Rather, it is quite possible that when an inappropriate missing value estimation method is used we may not only lose the knowledge of which peaks are significant or not, but we may introduce further bias by identifying non-significant peaks as significantly different between groups.

Our results therefore allude to the potential bias in the biochemical interpretation of metabolomics data, if missing values are estimated incorrectly. To verify this we have compared what we refer to as ‘active’ human pathways observed following the estimation of missing data, across all eight algorithms. These again resulted in quite diverse outcomes of ‘active’ pathways with only 20.0, 14.4 and 0.0% (for CCL_n, CCL_p and HL datasets, respectively) of pathways observed across all of the missing data algorithms (Table 2; electronic supplementary material Tables A4–A5); note that this approach was not applied to the non-human DM dataset. The highest number of ‘active’ pathways was detected for the *S*, *HM* and *KNN* methods, while the lowest for *M*, *MED* and *BPCA*. Prior biochemical knowledge can also aid the interpretation of these findings. For example, for the HL dataset, arginine and proline metabolism and taurine and hypotaurine metabolism are known to play a substantial role in liver transplantation (Silva 2006; Kincius et al. 2007). Both these pathways were discovered to be ‘active’ following treatment of missing values by the *S, HM, KNN* and *REP* methods, while *BPCA* treatment did not lead to either being classed as active. Overall, the results presented here provide substantial evidence that the choice of missing value estimation method has a substantial effect on the outcome and interpretation of univariate statistical analysis.

3.3 Impact of missing data imputation on multivariate data analysis

Similar to the results from the univariate analyses, the eight missing data estimation techniques also led to diverse

Table 2 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the significantly changing peaks between cold phase and post reperfusion groups) in the human liver (HL) dataset, after estimating the missing values with eight different algorithms

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Purine metabolism	X*	X	X	–	X	X	X	X
ABC transporters	X	X	X	X	X	–	X	X
Neuroactive ligand-receptor interaction, Taurine and hypotaurine metabolism	X	X	–	X	X	–	X	X
Pyrimidine metabolism, arginine and proline metabolism	X	X	X	–	X	–	–	X
Nicotinate and nicotinamide metabolism, Tyrosine metabolism, Drug metabolism—cytochrome P45	X	X	–	–	X	–	X	X
Glycine, serine and threonine metabolism, Aminoacyl-tRNA biosynthesis, Cysteine and methionine metabolism, Alanine, aspartate and glutamate metabolism	X	X	–	–	–	X	–	X
Galactose metabolism	X	X	–	–	X	–	–	X
Lysine degradation, Histidine metabolism, Beta-alanine metabolism, Phenylalanine metabolism, Tryptophan metabolism	X	X	–	–	–	–	–	X

* X indicates that a pathway is 'active' for this particular method

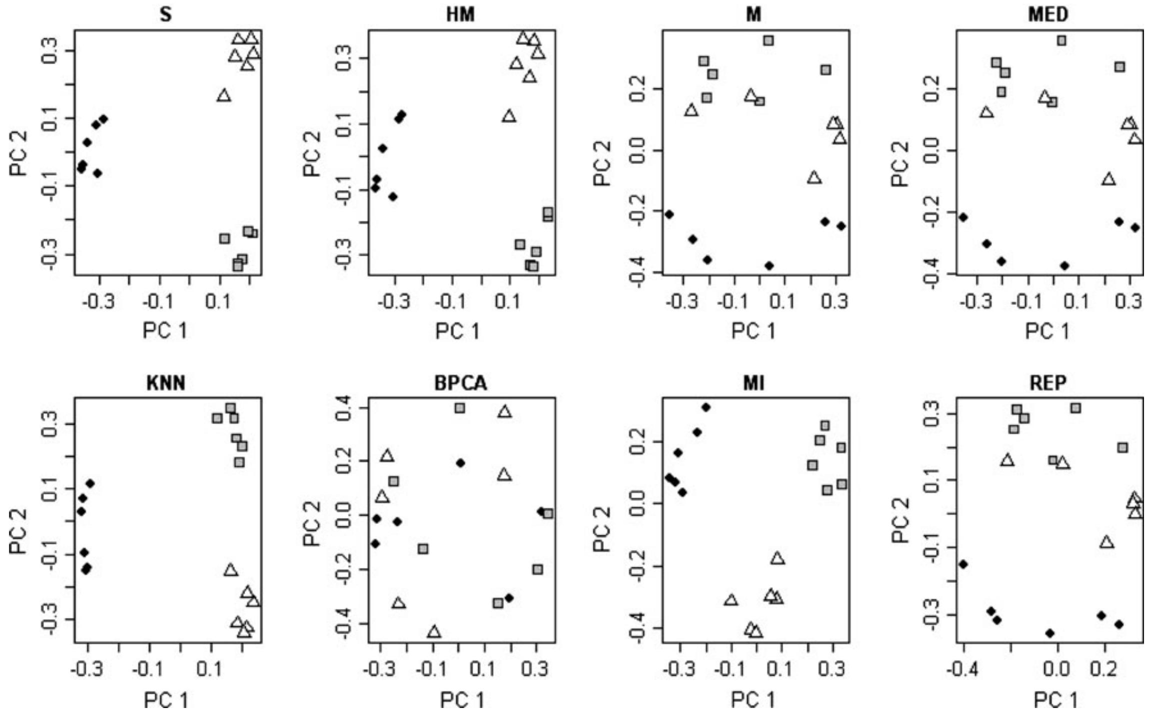

Fig. 3 Comparison of eight different missing value estimation methods based upon their effects on the PCA scores plots for the CCL_n dataset. Samples labelled as: control cancer cells (*diamonds*), indomethacin treated (*squares*) and medroxyprogesterone acetate treated (*triangles*)

outputs from the multivariate data analysis. Specifically, this was assessed from the clustering (or not) of samples from different biological groups on PCA scores plots (Fig. 3; electronic supplementary material Figs. A5–A7). The differences between the eight estimation techniques were most evident for the most biologically homogeneous dataset, the cell line extracts. For example, for CCL_n there were clear differences between the control, indomethacin treated and medroxyprogesterone acetate treated groups after estimating the missing entries with *S, HM, KNN* and *MI*, a separation between the control and two drug treated groups (but no separation between drug treatments) after *M, MED* and *REP*, and no separation between any of the groups after *BPCA*. The differential effects of the eight estimation methods on the PCA results were further demonstrated by the large spread of the variances captured by the first two principal components: the relative standard deviations of the variances were 36.10 and 16.58% for PC1 and PC2 respectively for CCL_n, 57.46 and 29.53% for CCL_p, 56.52 and 15.13% for DM, and 38.64 and 11.72% for the HL dataset (electronic supplementary material Table A6).

Comparison of the top 5% of peaks contributing towards the separation along PC1 and along PC2 showed, as for the univariate analyses, that some estimation methods performed quite similarly (electronic supplementary material Figs. A8–A9, Tables A7–A8). For example, the largest overlap in significant peaks for the univariate data analysis was between *S* and *HM* (97.51, 97.61, 99.83 and 97.28% overlap for CCL_n, CCL_p, DM and HL, respectively) followed by a slightly smaller overlap between *M* and *MED* (95.33, 92.86, 95.38 and 68.29%), while for the top 5% of peaks from the PCAs the similarities (expressed as R_t; see footnote in electronic supplementary material Table A7) were largest between *M* and *MED* (97.08, 84.82, 96.19 and 86.67% for PC1, and 94.46, 88.39, 92.38 and 75.56% for PC2) and followed by *S* and *HM* (76.68, 70.09, 82.38 and 61.11% for PC1 and 77.84, 64.29, 80.00 and 71.11% for PC2). For the PCA results, the smallest overlap was between the *S* and *BPCA* methods, whereas for the univariate data analysis the smallest overlap was between *S* and *M* or *MED*. An important observation from these findings is that the differences in the statistical results between the eight estimation methods were larger for the PCA than for the univariate analyses, indicating that the multivariate data analysis may be more sensitive to the missing data estimation technique used.

This observation of the higher sensitivity of multivariate analysis to missing data estimation was additionally verified by further examination of the top 5% of peaks contributing towards the separation of biological groups along the principal components (from each PCA). In general, these subsets of *m/z* values contained a larger proportion of peaks that initially contained missing data as well as a larger proportion of missing entries than their univariate equivalents (except for the *BPCA* method applied to the CCL_n, DM and HL datasets). For the *S* and *HM* methods, virtually all the peaks in this top 5% subset contained at least one missing value prior to their estimation (electronic supplementary material Table A6, Figs. A10–A11, Tables A9–A10). Furthermore, the considerable differences in the results of the PCAs between the eight estimation methods were further illustrated by substantial heterogeneity in the observed ‘active’ human pathways. Specifically, there were no common pathways following application of the eight tested estimation methods for CCL_n (for the top 5% of peaks contributing to PC1 and to PC2), CCL_p (for top 5% of peaks contributing to PC2) and HL (for top 5% of peaks contributing to PC1). Virtually none of the remaining dataset and principal component combinations exhibited overlap across all eight estimation methods, except for 2.11 and 3.85% of ‘active’ pathways for CCL_p (PC1) and HL (PC2), respectively (electronic supplementary material Tables A11–A15). Overall, these analyses provide definitive evidence that the choice of missing value estimation method has a substantial effect on the results of the multivariate statistical analysis used here.

3.4 Performance of missing data estimation algorithms

Assessing the effects of missing data estimation methods on the ‘complete’ datasets, which had missing values deliberately introduced either at random (MCAR, missing completely at random; mentioned here as a comparative benchmark only and with results reported in the electronic supplementary material) or in a way that mimics the missing data distribution and properties of actual FT-ICR MS metabolomics datasets (MNAR, missing not at random), revealed further interesting findings. Before examining these results, it is worth noting that the ‘complete’ datasets may not be a perfect representation of the ‘original’ datasets in terms of the metabolites’ intensities as a larger proportion of the peaks that were removed to create the ‘complete’ datasets were of relatively low intensity (i.e. the actual missing data did not occur at random, but occurred in part as a function of peak intensity). However, once missing data were intentionally introduced to the ‘complete’ datasets to mimic their distribution in the ‘original’ data, the ‘complete’ datasets regain similar properties to the ‘original’ datasets; in fact this represents the best possible approach to assess the performance of missing values estimation algorithms even when missing data are a function of intensity (Albrecht et al. 2010; Scheel et al. 2005) since the estimation methods are compared internally within the ‘complete’ matrix. Here, an obvious bias could occur for method *S,* with a small predefined value (0.01), while the seven other methods should capture the relationship of the ‘original’ data. This limitation should, however, be kept in mind and this component of the comparison of imputation algorithms should be combined with the findings from the missing data distributions and their influence on the data analysis (see above) to select the optimal missing data estimation method.

The performances of the eight methods were evaluated in terms of normalised root mean square errors (NRMSE), where the best approach results in the smallest NRMSE between the known, deliberately deleted, ‘missing’ data and the values that were subsequently estimated for them (averaged across N = 100 runs). Five of the eight estimation methods yielded similarly small average NRMSE values, specifically methods *MI*, *BPCA*, *KNN*, *MED* and *M*, while methods *REP, S* and *HM* performed poorly (Fig. 4;

electronic supplementary material Fig. A12, Table A16). This trend was observed both when the missing data was introduced completely at random (MCAR) as well as for the more realistic case of introduced values not at random (MNAR). For the case of MCAR, is it expected that *M* and *MED* yield good results since being a least squares estimator method they give the best approximation when no information on the missing data is available (when averaged over many runs). This hints at the challenge of deriving robust conclusions as to which imputation method is optimal for a given nature and distribution of missing values, discussed below.

Next the eight algorithms were assessed in terms of their impact on univariate data analysis, evaluated using the area under the receiver-operating characteristics (ROC) curve. Specifically, for each run (repeated 100 times), data were deliberately removed from the complete dataset (MCAR and MNAR) and then these 'missing values' were estimated via each of the eight methods, and significantly changed peaks at various statistical significance levels were identified and compared with the ones identified for the original complete datasets (with no missing values) using an ROC curve. The area under the ROC curve (AUC) was averaged across 100 runs, with the best method (i.e. closest to the complete dataset) having the highest AUC value (up to 1). The highest performance methods were similar to the best estimation algorithms from the NRMSE assessment, i.e. the largest AUC was observed for the *M, MED* and *KNN* methods, and the smallest for *REP, S* and *HM* methods (Fig. 4; electronic supplementary material Fig. A12, Table A17) with no differences being detected between introducing missing data as MCAR or MNAR.

To provide a framework for the interpretation of the multivariate data analysis, we first conducted a PCA on each of the four 'complete' datasets, and then tested the significance of any group separation by calculating p

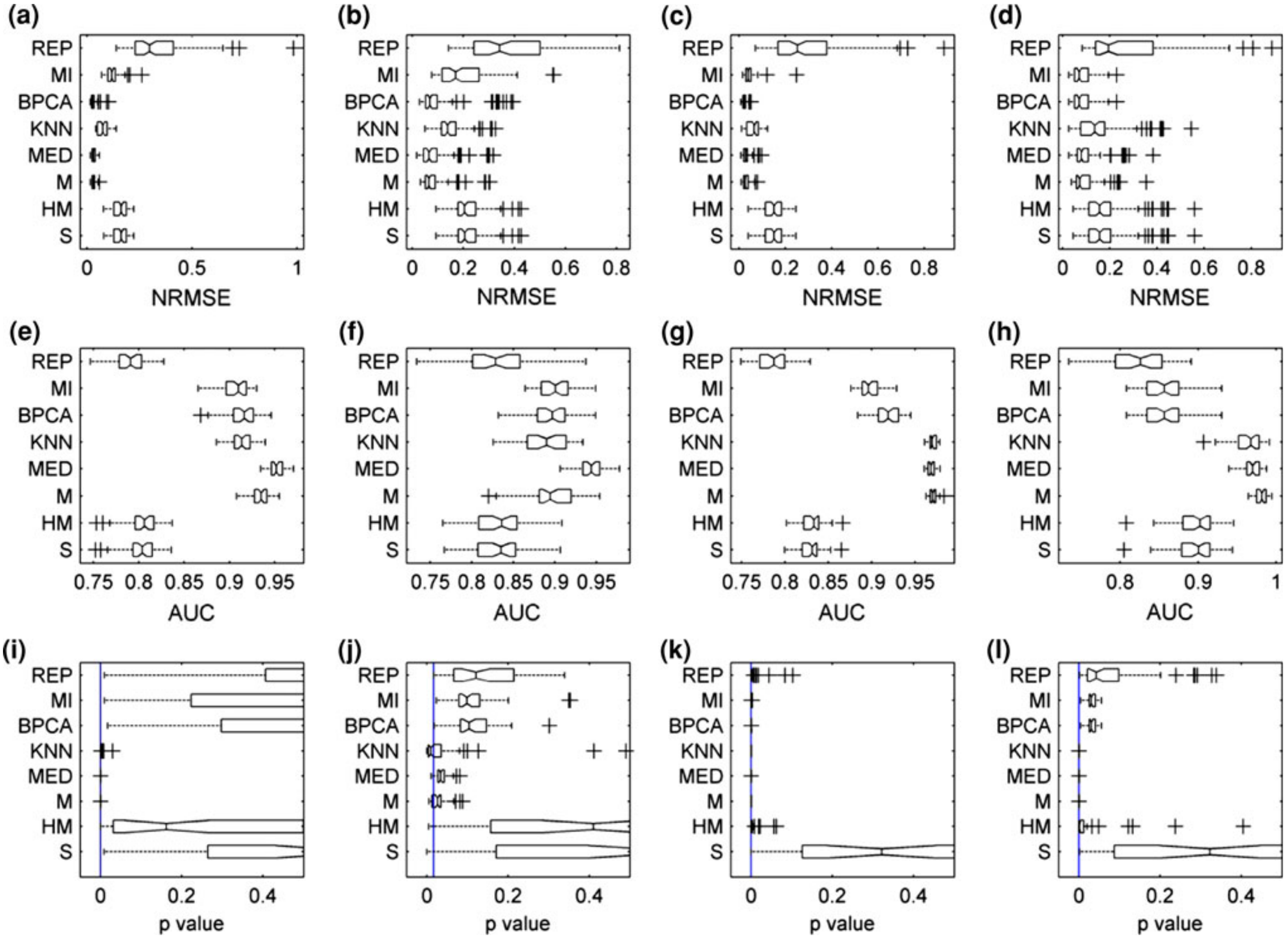

Fig. 4 Analyses of four DI FT-ICR MS datasets after first introducing and then estimating missing data in the 'complete' datasets as 'missing not at random', to best represent actual MS data (average of 100 runs). Boxplots of NRMSE values for the **a** CCL_n, **b** CCL_p, **c** DM and **d** HL datasets; boxplots of area under ROC curves (AUC) for **e** CCL_n, **f** CCL_p, **g** DM and **h** HL datasets; and distribution of p values (ANOVA or *t*-test on PC scores) for **i** CCL_n (PC2 axis), **j** CCL_p (PC2 axis), **k** DM (PC1 axis) and **l** HL (PC1 axis) datasets, where the vertical lines indicate the p values for the complete datasets and therefore represent the ideal result following missing value estimation

values (t test or ANOVA) for the PC scores along both PC1 and PC2 (electronic supplementary material Fig. A13). Similar to the ROC assessment, for missing data introduced more realistically as MNAR, *M, MED* and *KNN* outperformed the five other estimation methods by revealing the known significant separation between the biological groups in PCA scores plots. Specifically a significant separation along PC1 for the DM and HL datasets was revealed after imputing missing entries with the majority of the eight methods (except for method *S* for the DM dataset, and *S* and *REP* for the HL dataset), but there were few false positive errors along PC2 after *M* and *MED* imputation (DM dataset) and *BPCA* and *MI* imputation (HL dataset). Comparison to the MCAR shows that multivariate data analysis is more prone to errors and miscalculations than for the univariate data analysis. This is further supported by the similarities of the top 5% of peaks contributing to the separation of biological groups (along PC1 or PC2), for which the similarity values are quite different between MCAR and MNAR (electronic supplementary material Figs. A14–A15, Table A20). Overall this supports our earlier findings that multivariate data analysis is more sensitive than univariate analysis to the occurrence of missing values. Furthermore we have shown that the *M, MED* and *KNN* estimation methods appear to outperform the others, although see the discussion below.

3.5 Missing data estimation: is there an optimal method?

Based upon the studies to date on missing data, described here and conducted in other 'omics fields, there are currently no grounds to prescribe any one estimation method for dealing with missing entries in metabolomic datasets. However, based upon our findings, it is clearly of considerable importance to address the question of what is the optimal treatment of missing data. For example, simply deleting the variables that contain missing data or, as we have shown, estimating those values with an arbitrarily selected method will likely introduce a large bias to the dataset and significantly affect further data analysis and interpretation. The first step in selecting an appropriate estimation method should be focused on characterising the nature of the missing values within the given dataset. Typically a metabolomics study will have measured thousands of peaks and, as presented here, one should try to infer the relationships between the missing data and the non-missing data. In addition, one should try to establish whether missing data occur as a result of metabolite abundances being below the detection limit of the analytical platform (thought to be the primary case for our FT-ICR MS datasets) or instead if they represent non-detects (i.e. metabolites not measured due to a failure of the analytical platform.). With this information, as well as with the assessment of whether and to what extent missing data influence a particular data analysis method, an appropriate missing data estimation technique can be chosen, as discussed below; this three stage approach is outlined in Fig. 1.

In the case of the three DI FT-ICR MS based metabolomic datasets investigated here our initial findings suggest that the preferred methods of estimating missing values are *KNN*, *M* and *MED*. These three methods achieved a good balance between enabling the statistical analyses to reveal the expected metabolic differences between biological groups (Fig. 3; electronic supplementary material Figs. A5–A7) and yet did not identify too many potentially false positive biomarkers, i.e. the significantly changing peaks (univariate; electronic supplementary material Table A1) or those peaks contributing towards separation in PCA space (multivariate; electronic supplementary material Table A6) generally did not contain the highest number of missing data when compared to other estimation methods. The *KNN*, *M* and *MED* also performed the best when assessed using the 'complete' datasets with deliberately introduced missing values that were MNAR. These three methods yielded low NRMSE values, high AUC values associated with the univariate analyses, and impressively low p values associated with the separation of samples in multivariate space (Fig. 4).

We then sought to further compare these three methods, *M, MED* and *KNN*, to find the optimal approach. Although *KNN* was slightly outperformed by *M* and *MED* for the NRMSE and univariate analysis (using the 'complete' datasets with deliberately introduced missing values as MNAR) we hypothesised that the *M* and *MED* methods may introduce a larger bias into the datasets due to the way that they estimate missing values; i.e. it is likely that by calculating the mean or median of the non-missing measured values, estimated peak abundances would take on large values, possibly with the majority of peaks being estimated above the SNR threshold. We therefore evaluated the original datasets (from part B of Fig. 1) and confirmed that this is indeed the case. Specifically, for both the *M* and *MED* methods, almost all peaks (ca. 71–95%) were predicted to have intensities above the threshold (for the CCL_n, CCL_p, DM and HL datasets), as opposed to *KNN* for which there was an almost equal split of estimated intensities below and above the SNR threshold (ca. 38–66% of peaks above the threshold; electronic supplementary material Table A21). Considering the proven relationship in our FT-ICR MS datasets between peak abundances and the number of missing values, it is logical to expect that many, if not the majority, of missing values should lie on or below the SNR threshold. Therefore, based upon our analyses, we conclude that the *KNN* method imputes the most realistic values in these datasets and therefore is the preferred method over

M and *MED*. This conclusion is strongly backed by existing literature which states that imputing mean values is not a good approach (Little and Rubin 2002; Buuren and Groothuis-Oudsshoorn). Furthermore, both the *M* and *MED* methods are disadvantaged by the fact that they can cause an artificial reduction of variance. Overall, we therefore consider *KNN* to be the optimal missing value imputation method for the datasets examined here.

Generalising our findings and drawing upon previously published relevant literature, we recommend some pragmatic guidelines for the applied metabolomics researcher to decide upon which missing data estimation algorithms to use: (a) assess whether there is a need to impute missing data (especially if univariate analysis are to be conducted) as any imputation method will potentially bias the data analysis; (b) avoid replacing missing data with small arbitrarily chosen numbers (as in *S, HM* and *REP* methods) since this greatly affects the data analysis; only consider these methods when the number of missing data is low and the majority of missing entries are known to result from measurements below the limit of detection; (c) if the origin of the missing data is largely unknown or the majority of missing entries are non-detects rather than measurements below the limit of detection, select an estimation algorithm that is based on searching for local or global similarities such as *KNN* (that has been shown to be optimal in this study); (d) consider *MED* or *M* imputation only when the missing data represent true non-detects as opposed to measurements below the limit of detection; and (e) ideally evaluate the chosen method against alternative algorithms to avoid obviously biased missing data imputation. This last point need not be limited to the eight imputation approaches presented here, but could possibly include other Multiple Imputation methods or the Expectation–Maximisation algorithm as used in other fields (Schafer 1999; Little and Rubin 2002). Finally, the development of novel imputation methods remains an active field, for example the LinCmb (Jornsten et al. 2005) algorithm for microarray expression profiles that adapts to the structure of the data by changing emphasis on the local and global imputation methods, and hence the metabolomics researcher should be aware of on-going progress in this field to help guide their selection in the future.

4 Concluding remarks

We have shown that missing values play an important role in DI FT-ICR MS based metabolomics data, and that their estimation is very strongly reflected in the final data analysis outcome, for both univariate and multivariate approaches. Therefore, we conclude that the optimal treatment of missing data is a crucial step in the data processing pipeline to which special attention should be paid. Even though this study is based on three DI FT-ICR MS based metabolomic datasets, our analyses and findings are more generally applicable and of interest to all metabolomics studies. We propose a three step process in order to determine an optimal method for missing value estimation for a given dataset and analytical platform (summarised in Fig. 1), that includes: assessing the nature of the missing data, analysing the impact of missing data treatments on the final data analysis outcome, and analysing the performance of missing data algorithms on the 'complete' datasets if available. Using this three step approach, we conclude that the optimal missing data estimation technique for DI FT-ICR MS based metabolomics is the *KNN* method.

Acknowledgments We thank Drs. Alessia Lodi, Stefano Tiziani and Chris Bunce for provision of the FT-ICR MS datasets of the cancer cell extracts.

References

Albrecht, D., Kniemeyer, O., Brakhage, A. A., & Guthke, R. (2010). Missing values in gell based proteomics. *Proteomics, 10*(6), 1202–1211.

Andersson, C. A., & Bro, R. (1998). Improving the speed of multiway algorithms. Part I. Tucker 3. *Chemometrics and Intelligent Laboratory Systems, 42*, 93–103.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*, 289–300.

Blanchet, L., Smolinska, A., Attali, A., Stoop, M. P., Ampt, K. A. M., van Aken, H., et al. (2011). Fusion of metabolomics and proteomics data for biomarkers discovery: Case study on the experimental autoimmune encephalomyelitis. *BMC Bioinformatics, 12*, 254.

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

de Brevern, A. G., Hazout, S., & Malpertuy, A. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics, 5*, 114.

Defamie, V. (2008). Gene expression profiling of human liver transplants identifies an early transcriptional signature associated with initial poor graft function. *American Journal of Transplantation, 8*, 1221–1236.

Dieterle, F., Ross, A., Scholtterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabolomics. *Analytical Chemistry, 78*, 4281–4290.

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

Hrydziuszko, O., Silva, M. A., Perera, T. P. R., Richards, D. A., Murphy, N., Mirza, D., et al. (2010). Application of metabolomics to investigate the process of human orthotopic liver transplantation: A proof-of-principle study. *OMICS: A Journal of Integrative Biology, 14*, 143–150.

Jornsten, R., Wang, H., Welsh, W., & Ouyang, M. (2005). DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics, 21*, 4155–4161.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., & Itoh, M. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research, 36*, D480–D484.

Kenny, L. C., Broadhurst, D. I., Dunn, W., Brown, M., North, R. A., McCowan, L., et al. (2010). Robust early pregnancy prediction of later preeclampsia using metabolomics biomarkers. *Hypertension, 56*, 741–749.

Kim, K. Y., Kim, B. J., & Yi, G. S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics, 5*, 160.

Kim, D. W., Lee, K. Y., Lee, K. H., & Lee, D. (2007). Towards clustering of incomplete microarray data without the use imputation. *Bioinformatics, 23*, 107–113.

Kincius, M., Liang, A., Nickkholgh, K., Hoffmann, C., Flechtenmacher, C., Ryschich, E., et al. (2007). Taurine protects from liver injury after warm ischemia in rats: The role of Kupffer cells. *European Surgical Research, 39*, 275–283.

Little, R. J. A. (1998). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New Jersey: John Wiley & Sons.

Oba, S., Sato, M., Takemasa, I., Monden, M., et al. (2003). A Bayesian missing values estimation method for gene expression profile data. *Bioinformatics, 19*, 2088–2096.

Parsons, H. M., Ekman, D. R., Collette, T. W., & Viant, M. R. (2009). Spectral relative standard deviation: A practical benchmark in metabolomics. *Analyst, 134*, 478–484.

Parsons, H. M., Ludwig, C., Günther, U. L., & Viant, M. R. (2007). Improved classification accuracy in 1- and 2- dimensional NMR metabolomics data using the variance stabilizing generalised logarithm transformation. *BMC Bioinformatics, 8*, 234.

Payne, T. G., Southam, A. D., Arvanitis, T. N., & Viant, M. R. (2009). A signal filtering method for improved quantification and noise discrimination in Fourier transform ion cyclotron resonance mass spectrometry-based metabolomics data. *Journal of American Society for Mass Spectrometry, 20*, 1087–1095.

Pedreschi, R., Hertog, M. A. T. M., Carpentier, S., Lammertyn, J., Robben, J., et al. (2008). Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. *Proteomics, 8*, 1371–1383.

Rubin, D. R. (1976). Inference and missing data. *Biometrica, 63*, 581–592.

Sangster, T. P., Wingate, J. E., Burton, L., Teichert, F., & Wilson, I. D. (2007). Investigation of analytical variation in metabonomics analysis using liquid chromatography/mass spectrometry. *Rapid Commun. Mass Spectrometry, 21*, 2965–2970.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*, 3.

Scheel, I., Aldrin, M., Glad, I., Sorum, R., Lyng, H., & Frigessi, A. (2005). The influence of missing values imputation on detection of differentially expressed genes from microarray data. *Bioinformatics, 21*, 4272–4279.

Silva, M. A. (2006). Arginine and urea metabolism in the liver graft: A study using microdialysis in human orthotopic liver transplantation. *Transplantation, 82*, 1304–1311.

Southam, A. D., Payne, T. G., Cooper, H., Arvanitis, T. N., & Viant, M. R. (2007). Dynamic range and mass accuracy of widescan direct infusion nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. *Analytical Chemistry, 79*, 4595–4602.

Steuer, R., Morgenthal, K., Weckwerth, W., & Selbig, J. (2007) A gentle guide to the analysis of metabolomic data. In: *Metabolomics: Methods and protocols* (pp. 105–129). New Jersey: Humana Press.

Sumner, L. W., Amberg, A., Barret, D., Beale, M. H., Berger, R., et al. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). *Metabolomics, 3*, 211–221.

Taylor, N. S., Weber, R. J. M., Southam, A. D., Payne, T. G., Hrydziuszko, O., Arvanitis, T. N., et al. (2009). A new approach to toxicity testing in *Daphnia magna*: An application of high throughput FT-ICR mass spectrometry metabolomics. *Metabolomics, 5*, 44–58.

Taylor, N. S., Weber, R. J. M., White, T. A., & Viant, M. R. (2010). Discriminating between different acute chemical toxicities via changes in the daphnid metabolome. *Toxicological Sciences, 118*, 307–317.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*, 520–525.

Tuikkala, J., Elo, L. L., Nevalainen, O. S., & Aittokallio, T. (2008). Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics, 9*, 202.

van Buuren, S., & Groothuis-Oudsshoorn, K. (2010). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 1*, 68–74.

Walczak, B., & Massart, D. L. (2001). Dealing with missing data part I. *Chemometrics and Intelligent Laboratory Systems, 58*, 15–27.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*, 81–89.

Wu, H., Southam, A. D., Hines, A., & Viant, M. R. (2008). High throughput tissue extraction protocol for NMR- and MS-based metabolomics. *Analytical Biochemistry, 372*, 204–212.

Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: A web server for metabolomics data analysis and interpretation. *Nucleic Acids Research, 37*, W652–W660.