

UC Davis

UC Davis Previously Published Works

Title

Comparison of imputation and imputation-free methods for statistical analysis of mass spectrometry data with missing data

Permalink

<https://escholarship.org/uc/item/8jz6z1ds>

Journal

Briefings in Bioinformatics, 23(1)

ISSN

1467-5463

Authors

Taylor, Sandra
Ponzini, Matthew
Wilson, Machel
et al.

Publication Date


2022-01-17

DOI

10.1093/bib/bbab353

Peer reviewed

Comparison of imputation and imputation-free methods for statistical analysis of mass spectrometry data with missing data

Sandra Taylor , Matthew Ponzini, Mabelle Wilson and Kyoungmi Kim

Corresponding author. Sandra L. Taylor, Division of Biostatistics, Department of Public Health Sciences, School of Medicine, University of California, Davis, 2921 Stockton Boulevard, Suite 1400, Sacramento, CA 95817, USA. Tel.: +1-9167039171; Fax: +1-9167039124; E-mail: sltaylor@ucdavis.edu

Abstract

Missing values are common in high-throughput mass spectrometry data. Two strategies are available to address missing values: (i) eliminate or impute the missing values and apply statistical methods that require complete data and (ii) use statistical methods that specifically account for missing values without imputation (imputation-free methods). This study reviews the effect of sample size and percentage of missing values on statistical inference for multiple methods under these two strategies. With increasing missingness, the ability of imputation and imputation-free methods to identify differentially and non-differentially regulated compounds in a two-group comparison study declined. Random forest and *k*-nearest neighbor imputation combined with a Wilcoxon test performed well in statistical testing for up to 50% missingness with little bias in estimating the effect size. Quantile regression imputation accompanied with a Wilcoxon test also had good statistical testing outcomes but substantially distorted the difference in means between groups. None of the imputation-free methods performed consistently better for statistical testing than imputation methods.

Key words: metabolomics; mass spectrometry; missing data; imputation; sample size

INTRODUCTION

High-throughput mass spectrometry (MS) is commonly used to analyze small molecular compounds (e.g. metabolites, lipids, proteins) in biological samples. By profiling hundreds or thousands of compounds simultaneously, investigators aim to identify compounds suitable for diagnostic and prognostic tests, understand biological pathways of disease and identify potential therapeutic targets. A notable characteristic of data from MS studies is the extent of missing values [1–3]. Depending on the platform and processing, the amount of missing data can be

considerable, >20% overall [2] with many compounds having missing values of varying degrees individually [1, 4].

Missing values create challenges for statistical analyses as most statistical methods require complete data [5]. When missing values occur, two strategies are available: (i) eliminate or impute the missing values and apply statistical methods that require complete data and (ii) retain the missing observations and use statistical methods that specifically account for missing values without imputation. The first approach is the most common. Elimination of missing values is typically accomplished by dropping compounds with missing values in excess of a

Sandra Taylor is a principal biostatistician in the Division of Biostatistics, School of Medicine at the University of California, Davis. Her research focuses on analytical methods for mass spectrometry data and clinical data.

Matthew Ponzini is a biostatistician in the Division of Biostatistics, School of Medicine at the University of California, Davis. His research interests include analytical methods for metabolomics data.

Mabelle Wilson is a principal biostatistician in the Division of Biostatistics, School of Medicine at the University of California, Davis. Her research interests include analytical methods for handling missing data and mixed effects modeling.

Kyoungmi Kim is a professor of biostatistics in the Division of Biostatistics, School of Medicine at the University of California, Davis, with expertise in high-throughput omics research. Her research focuses on post-genomic approaches to disease biomarker discovery.

Submitted: 5 May 2021; Received (in revised form): 27 July 2021

pre-specified threshold followed by imputation of any remaining missing values. The second approach is to use imputation-free statistical methods such as accelerated failure time (AFT) models [6], two-part (TP) models [7] and mixture (MM) models [4, 8]. For these methods, missing values are not imputed but retained as missing and data analyzed as collected.

Considerable work has been done comparing performance of imputation methods applied to MS data [1, 3, 9–13]. These studies have differed in the missing data processes used in simulations, the degree of missingness, the metrics used to compare performance of the imputation methods and in the characteristics of the data sets evaluated in terms of sample size and features. Several studies have assessed performance based on how close imputed data values were to original values [10–12, 14], but of practical consideration, is the effect of imputation on subsequent statistical analysis procedures, including differential regulation analyses, principal component analysis (PCA) or partial least squares analysis. Hrydziusko and Viant [1] evaluated the effect of eight imputation methods on statistical testing of differential regulation and PCA hierarchical clustering for three data sets of small sample sizes (less than 20) but only considered missingness of about 20%. They recommended *k*-nearest neighbor (KNN) and reported poor performance by half-minimum (HM) imputation. In simulations based on eight proteomics data sets, Liu and Dongre [13] found imputing all missing values in a sample with the observed minimum from that sample to perform best with respect to identifying differentially expressed proteins when missing values reflected values below a detection limit while Bayesian PCA (BPCA) and singular value decomposition performed better when missing values were missing at random. This recent study is the only work on MS studies to have compared imputation performance with experimental manipulation of sample sizes. Of note, however, in their simulations, simulated compounds were independent from each other, which could have disadvantaged imputation methods such as BPCA that exploit the correlation pattern among compounds. Do et al. [9] evaluated power and type I error of correlation and regression analyses following 31 imputation methods for different combinations and levels of missingness using a single large data set ($N = 1750$) and identified KNN as the most robust method. Webb-Robertson et al. [3] and Wei et al. [12] focused on PCA results following imputation. Webb-Robertson et al. compared PCA results of 11 imputation methods applied to three small proteomics data sets with thousands of peaks, while Wei et al. simulated different types and levels of missingness using four data sets of varying sample sizes ($N = 37$ – 977) but with fewer compounds. Wei et al. recommended random forest (RF) where data were predominantly missing at random and quantile regression (QR) where missing values resulted from detection limit censoring. Webb-Robertson et al. noted that no method was universally the best.

While many studies have focused on imputation, little work has been conducted comparing imputation-free methods to each other and to imputation methods. Taylor and Pollard [7] found two-part statistics to outperform substitution of missing values with zero or applying two-sample tests to non-missing values. Taylor et al. [4] focused on comparing two imputation-free methods, AFT and MM models, but only considered one imputation method, KNN. Tekwe et al. [6] advocated the use of the AFT model based on a simulation study comparing imputation with row mean, KNN and probabilistic PCA, but missing data were generated entirely through censoring consistent with the AFT model assumptions.

No study has comprehensively evaluated application of imputation and imputation-free methods in MS studies over a broad range of conditions and specifically evaluated key practical considerations in MS studies, specifically sample size, overall level of missingness and level of missingness in individual compounds. Further, an analytical strategy that integrates both imputation and imputation-free methods could enhance MS studies by increasing the number of analyzable compounds if imputation-free methods can accommodate analysis of compounds with high levels of missingness, predominantly due to low concentrations, that are commonly dropped from analyses. The information necessary to guide statistical analysis approaches in practice, however, is currently lacking.

Despite extensive previous work, gaps remain in our understanding of imputation and imputation-free methods with practical implications for researchers. To address gaps of particular relevance for deciding on statistical analysis approaches, we compare the performance of selected imputation and imputation-free methods to detect differentially regulated compounds over a wide range of missingness and sample sizes. We emphasize results of statistical analyses for the detection of differentially regulated compounds in a two-arm study design and compare imputation and imputation-free methods. We also compare parametric versus non-parametric statistical testing procedures following imputation, which has not been previously considered. Using several simulation studies motivated by real data and conducted to retain the correlation structure among compounds, we develop empirical recommendations for analyzing MS data with missing values.

Methods

Imputation methods

We evaluated five imputation methods: HM [3], KNN [15], BPCA [16], RF [17] and QR [18, 19]. We selected KNN, RF and BPCA because these methods have been shown to be strong performers in previous studies. We included QR imputation because it is a relatively new method that has not been extensively studied. Finally, we included HM because it remains commonly used and it reflects a non-random mechanism for missing values in the concept of the lower limit of detection, which is only otherwise captured by QR imputation. These methods and software implementation are described in Supplementary Material available online at <http://bib.oxfordjournals.org/>.

Imputation-free methods

In contrast to imputation, imputation-free methods explicitly account for missing values in calculating an inferential test statistic. We evaluated four imputation-free methods: AFT, TP, MM models and the differential abundance analysis with Bayes shrinkage estimation of variance method (DASEV). Collectively, these methods capture the range of missing value mechanisms present in MS data. The AFT model is a survival analysis method that assumes missing values reflect compounds that are present but censored at concentrations below the detection limit [6]. In TP models [7, 20], a compound's distribution is represented by the proportion of missing observations and the distribution of observed values [7]. The model consists of two parts with one part testing for a difference in proportions of missing values and a second part evaluating the difference in means for observed values. These models are appropriate when missing values either represent the true absence of a compound or

reflect random technical measurement errors. Mixture models combine aspects of TP and AFT models to account for missing values resulting from censoring, true absence and technical limitations resulting in the failure to detect or quantify a compound [4]. The DASEV method attempts to improve on the MM model by employing an empirical Bayes shrinkage method to stabilize estimation of the variance under conditions of high levels of missingness [21]. These methods and software implementation are detailed in Supplementary Material available online at <http://bib.oxfordjournals.org/>.

Biological data sets

We used three biological data sets from previously published MS studies using different platforms to compare imputation and imputation-free methods: (i) polycystic kidney disease plasma metabolomics (PKD) [22], (ii) renal cell carcinoma urinary metabolomics (RCC) [23] and (iii) autosomal dominant polycystic kidney disease (ADPKD) plasma lipidomics (HALT) [24].

Polycystic kidney disease

The PKD data consist of non-targeted liquid chromatography/time of flight MS (LC/TOF-MS) analysis of plasma from 13 PKD patients and 13 healthy controls (HC) from a study evaluating the effects of meals, time of day and daily changes on the metabolome. A total of 873 metabolites were identified in at least one sample across all days and times with 40.6% missing values overall. For each sample, intensity values were total quantity normalized to the median total ion counts across all samples. We used 168 metabolites with no missing values in the dataset obtained from collected plasma from fasting patients on the first day of study.

Renal cell carcinoma

The RCC data set consists of metabolomics of urine from 29 RCC patients and 33 HC quantified using ultra-high performance liquid chromatography/tandem mass spectrometry optimized for basic species and acidic species. There were 298 known compounds identified with 12.7% missingness across all samples. Intensity values in each sample were divided by the sample's osmolality to adjust for differences in urine concentration among patients. We used 106 metabolites with no missing values.

Autosomal dominant polycystic kidney disease

The HALT data set is gas chromatography (GC-TOF/MS) lipidomics data for plasma of 544 patients with ADPKD in the HALT Progression of Polycystic Kidney Disease (HALT), NCT00283686, <http://clinicaltrials.gov> clinical trial. We used 207 negative mode peaks; overall missingness was 0.59%. Intensity values were total quantity normalized for each sample to the median total ion counts across all samples. We restricted our analysis to patients with the two most common genetic forms: PKD 1 ($n = 372$) and PKD 2 ($n = 82$). For these samples, 166 compounds with no missing values were used in our studies.

Simulation studies

We conducted two simulation studies. Study 1 compared results of statistical tests to detect differentially regulated compounds using imputation and imputation-free methods under a wide range of missingness levels. Study 2 explored the

interplay of sample size and level of missingness on statistical testing.

Study 1

In Study 1, we used all three real data sets to simulate a two-group comparison study reflecting a case-control design. Groups for our data sets were PKD: HC versus PKD; RCC: HC versus RCC; HALT: PKD 1 versus PKD 2. To create analytical data sets from the real data sets, we first permuted group labels to make the mean difference between groups 0 across all compounds. Then, for each metabolite, zSD was added to each sample in one group (HC or PKD 1) where SD was the pooled standard deviation for the metabolite. By first creating a data set with no difference on average between groups, we could experimentally create group differences of a specified magnitude allowing us to manipulate the numbers of differentially and non-differentially regulated compounds found in the complete data sets. The value for z was selected to yield power to detect statistically significantly different compounds between two groups of approximately 50 or 80% at a significance level of 5% in two sets of simulations. For the 50% power simulations, z was 0.75 for the PKD data set, 0.5 for RCC and 0.25 for HALT, and for 80% power, z was 1.0, 0.75 and 0.33 for PKD, RCC and HALT, respectively. At 50% power, 45–64% of the compounds were statistically significantly different (raw P -value < 0.05) and at 80% power 67–89% of the compounds were significant (Supplementary Table S1 available online at <http://bib.oxfordjournals.org/>). Over all compounds in a dataset, the mean effect size (ES) defined as the difference in means between the two experimental groups (Delta) divided by SD was zSD but ES varied among compounds (Supplementary Figure S1 available online at <http://bib.oxfordjournals.org/>). We simulated 100 complete analytical data sets from each real data set.

Using the complete sets, we induced missing values incrementally to yield overall missingness percentages from 1 to 70% across all samples and compounds using a strategy similar to Scheel et al. [25]. To generate $x\%$ missingness, we identified the pre-specified y th quantile of the data set. We then randomly selected values below y th quantile such that $x\%$ of the values in the entire data set were missing. For 1, 5, 10, 20, 30, 40, 50, 60 and 70% missingness, the y th quantiles used were 2, 10, 20, 40, 50, 60, 70, 80 and 90. This approach resulted in more missing values at low intensities consistent with the existence of a detection limit but did not impose a hard threshold. Resultant data sets were similar to real data sets consisting of some compound with no missing values and the remaining compounds with a range of missingness. (Supplementary Figure S2 available online at <http://bib.oxfordjournals.org/>). Figure 1 shows the process for simulating data for Study 1.

We analyzed the simulated data with missing values using the imputation-free methods and imputation methods by imputing with each imputation method followed by a two-sample t -test (parametric test) and Wilcoxon rank sum test (non-parametric test) to identify differentially regulated compounds. We compared overall performance of analytical approaches using a metric called the Biomarker List Concordance Index (BLCI) defined as $BLCI = \text{sensitivity} + \text{specificity} - 1$ [14]. A true positive is defined as a compound identified as being significant ($P < 0.05$) in the true complete data set and a true negative as being non-significant. We worked with raw P -values for a direct comparison of how these methods affected significance testing independent from a particular multiple testing adjustment method. In this context, sensitivity is the percentage of true positives that were correctly identified as being significant

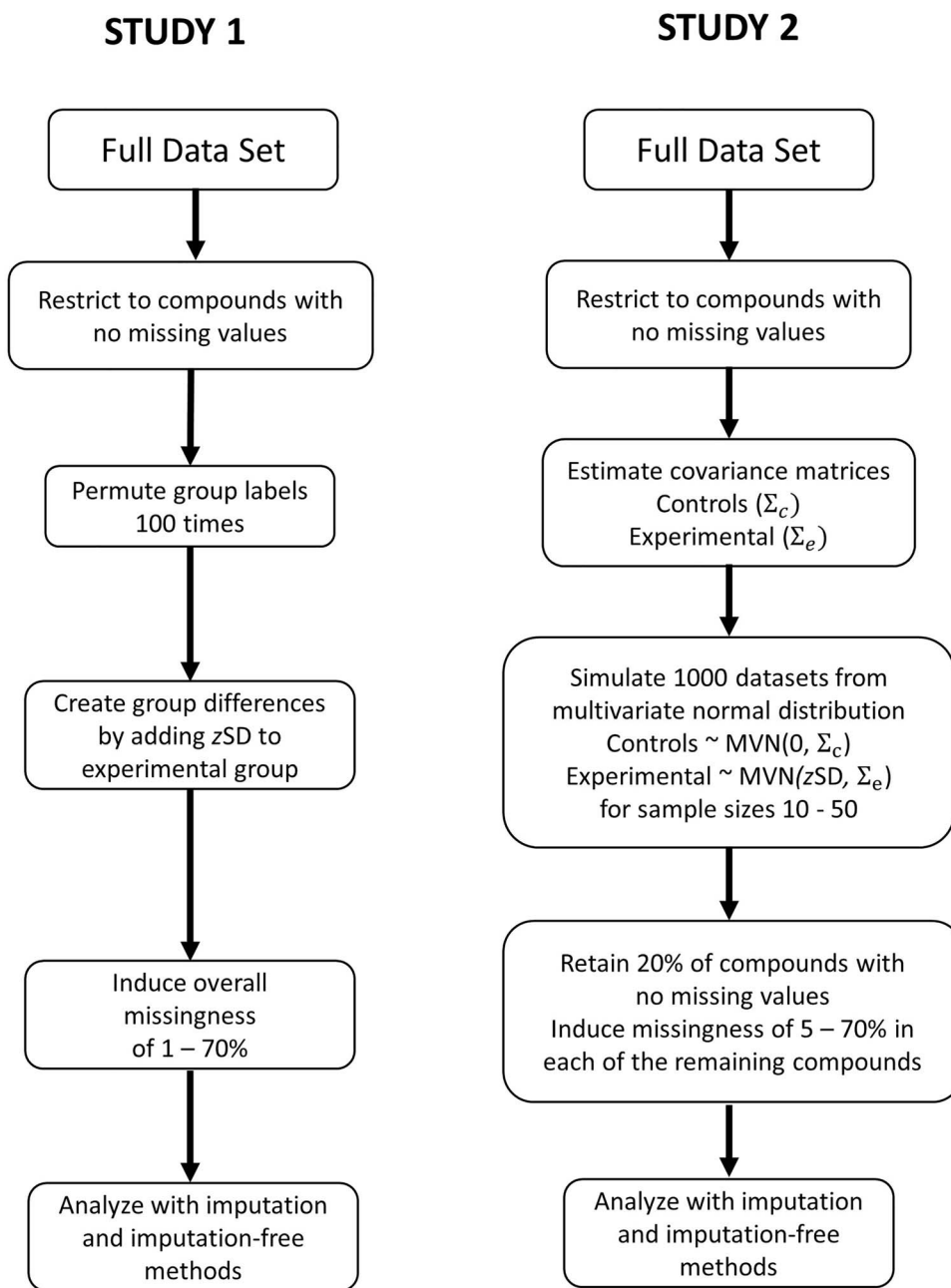


Figure 1. Process for simulating data for Study 1 and 2. Differences between experimental and control groups were created by adding zSD to each compound in the experimental group where SD was the pooled standard deviation and z was selected to achieve pre-specified statistical power based on a two-sample t-test. For Study 2, data were simulated from a multivariate normal distribution (MVN).

in the simulated data with induced missingness by the method under consideration. Similarly, specificity is the percentage of true negatives that were correctly identified as being not significant in the simulated data set with induced missing values. We used the BLCI metric for the overall assessment of a method's statistical performance because it reflects true discovery ability of each method in regard to both true positives and true negatives. Imputation results were compared to the corresponding test (t-test or Wilcoxon) results for the complete data set. Results from imputation-free methods were only compared to t-tests because for all of the imputation-free methods, the underlying statistical models assume a log

normal distribution for the quantitative component of the models. Complete data were log transformed for t-tests resulting in the distributions of most compounds following an approximately normal distribution. Compounds with no missing values were not included in calculating performance metrics.

For imputation methods, we also investigated the effect of imputation on ES by calculating the bias between ES, Delta and SD in the complete data sets versus imputed values. Analyses were conducted using R versions 3.6.2, 4.0.2 and 4.0.3 because simulations were conducted over many months and on different computers.

Table 1. Overall percentage of missing values in the entire datasets for Study 2 by percentage missing induced per compound

Compound % missing	Overall % of missingness in an entire dataset	
	PKD-based	HALT-based
5% Missing	4.0	4.0
10% Missing	8.0	8.0
20% Missing	16.0	15.9
30% Missing	23.9	23.9
40% Missing	31.9	31.8
50% Missing	39.9	39.8
60% Missing	47.9	47.7
70% Missing	55.8	55.7

There were 168 metabolites in PKD and 166 lipids in HALT. In the simulated data, 20% of all compounds were retained with no missing values. Missing percentages refer to the percentage of missing values in compounds with missing values.

Study 2

Study 1 provided the most realistic assessment of how the analysis procedures would perform when applied to real data. However, Study 1 could not answer questions related to sample size and thresholds of missingness for retaining compounds in an analysis.

Study 2 was specifically designed to evaluate sample size and levels of missingness, which necessitated fully simulating data. Because KNN, RF and BPCA use information from all the data to impute missing values, it was important that the simulated data conserved a realistic covariance structure. Thus, we estimated the between-compound covariance matrices for cases and controls separately for the PKD and HALT data sets. For HALT, a random sample of controls was used to provide a balanced design. These covariance matrices were then used to simulate data from a multivariate normal distribution. We again considered a two-group comparison setting and simulated data with mean 0 for a control group. For the cases, means varied depending on the sample size but were selected to provide an ES for which a two-sample t-test has approximately 50% power at a significance level of 5%. For sample sizes of 10, 20, 30, 40 and 50 per group, the means for compounds/lipids in the simulated cases group were 0.89SD, 0.62SD, 0.51SD, 0.44SD and 0.39SD, respectively, where SD was the respective pooled standard deviation. We generated 1000 data sets for each sample size at each sample based on PKD and HALT data.

In our data sets, 19–36% of the compounds had no missing values. In our simulated data sets, we retained 20% of the compounds with no missing values. In the remaining compounds, we induced missingness levels of 5–70% in each compound using the same approach as in Study 1. This approach also allowed us to evaluate the performance of imputation-free methods when all compounds had fixed, known and constant levels of missing values. Table 1 shows the overall percentage of missing values across all samples and compounds including those with no missing values in the simulated data. Figure 1 shows the process for simulating data for Study 2.

Results

Study 1

As the level of missingness increased, BLCI decreased for all methods (Figure 2) and was affected at low percentages of missingness (e.g. $\leq 5\%$). QR or HM imputation with a non-parametric

Wilcoxon test provided the highest BLCI for PKD with $\leq 5\%$ missing and $\leq 10\%$ missing in RCC. At higher levels of missingness ($\geq 10\%$), RF yielded the highest BLCI for PKD, RCC and HALT. However, the statistical test coupled with RF mattered. For HALT, RF with Wilcoxon had a higher BLCI than RF with a t-test but vice versa for PKD and RCC. Although QR and HM with a Wilcoxon test performed well at low percentage of missingness, statistical testing with a t-test was worse than other methods for missingness $> 5\%$. KNN performed relatively well with either testing procedure at missingness $\geq 20\%$, often yielding the second highest BLCI. The relative performance of BPCA varied. For PKD and RCC, the BLCI of BPCA was one of the worst with a Wilcoxon test but it was relatively better for HALT, comparable to KNN. With a t-test, BPCA was typically intermediate to other imputation methods. Supplementary Table S2 available online at <http://bib.oxfordjournals.org/> provides median values of BLCI for each procedure and level of missingness.

Among the imputation-free methods, BLCI was high for the AFT model at the smallest missingness levels ($\leq 5\%$; Figure 2). The TP and MM procedures had similar BLCI and were best among the imputation-free methods for missingness $\geq 20\%$. The DASEV method requires at least one missing and one non-missing observation per group and at least 10 non-missing observations total in order to obtain the prior distribution for the variance. At low missingness, some compounds contained no missing values in one of the groups, and hence no results were produced for the DASEV method in these instances. This resulted in low BLCI values because sensitivity and specificity were calculated as the number of true positives or negatives identified by DASEV divided by all true positives or true negatives in the complete data, respectively. At higher missingness levels, DASEV improved but remained at or below the MM model. Overall, imputation-free methods yielded similar BLCI values to imputation approaches with t-tests, which generally were worse than imputation with Wilcoxon tests (Figure 2).

The BLCI depends on both specificity and sensitivity and the methods differ in their effects on these components as missingness increases. Patterns in sensitivity and specificity were broadly similar across data sets (Supplementary Figures S3 and S4 available online at <http://bib.oxfordjournals.org/>). Sensitivity with QR and HM imputation and the imputation-free methods declined sharply with increasing missingness in all data sets dropping below 50% when missingness reached 20–40% depending on the data set. RF and KNN had more stable sensitivities with varying missingness and were markedly higher than QR and HM imputation. Sensitivity of RF and KNN remained above 75 and 50%, respectively, in nearly all cases (Supplementary Figure S3 available online at <http://bib.oxfordjournals.org/>). In the PKD data set, for BPCA imputation, sensitivity declined with increasing missingness. However, in HALT and RCC data sets, BPCA sensitivity tended to remain high ($> 75\%$) until the percentage of missing values exceeded 30% or more. These patterns generally switch for specificity with QR, HM and imputation-free methods having high ($> 75\%$), stable specificity, while KNN and RF show declining specificity with increasing missingness. BPCA had a complex pattern with specificity often among the worst for low levels of missingness (e.g. up to about 30%) but then performing well for higher levels of missingness up to 60 or 70%.

For imputation methods with a t-test, the changes in sensitivity and specificity, and hence BLCI, reflect how imputation changes the ES (Figure 3; Supplementary Figures S5 and S6 available online at <http://bib.oxfordjournals.org/>). ES is determined by both Delta and SD and both are affected

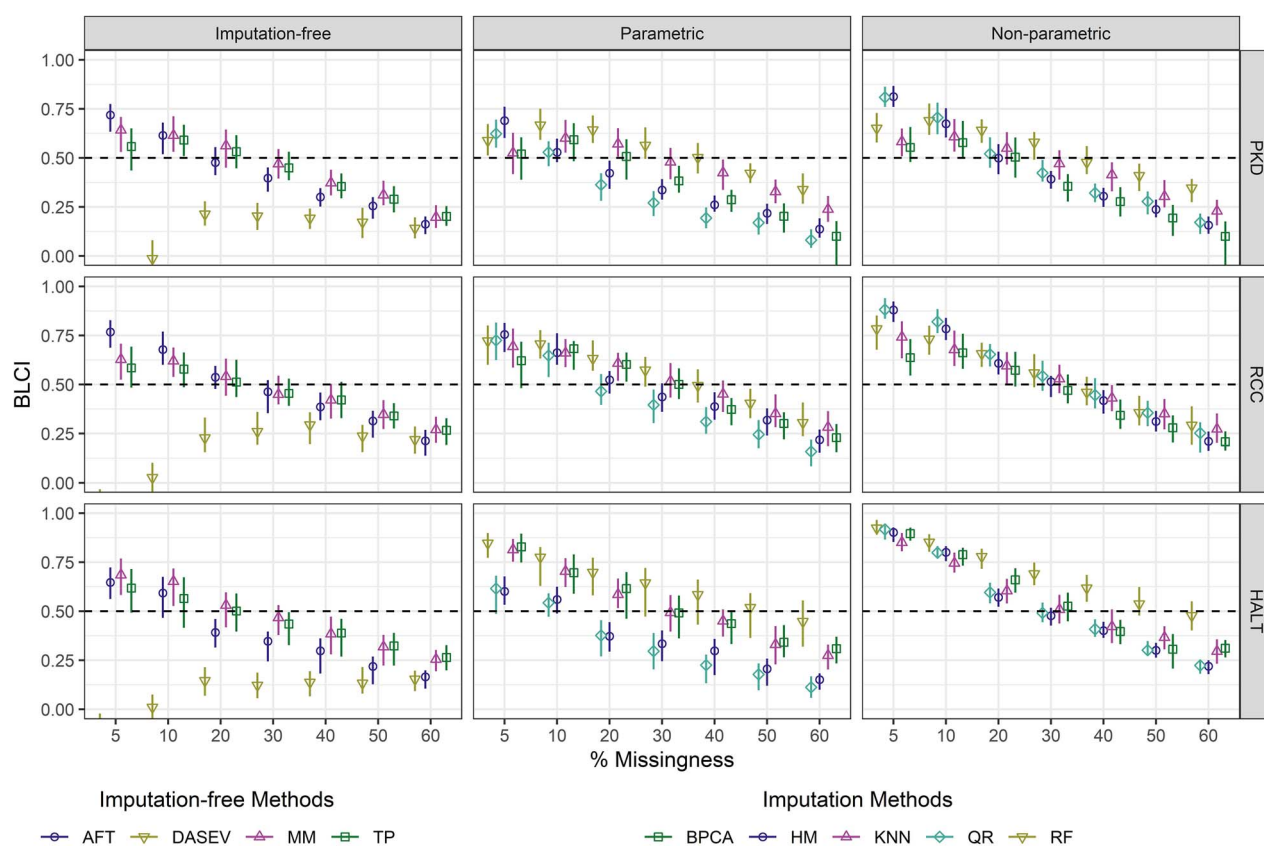


Figure 2. Median and quartiles of BLCI versus overall percentage of missing values for 100 simulated data sets using PKD, RCC and HALT data with ESs of 0.75SD, 0.5SD and 0.25SD respectively, in Study 1. Groups were compared using imputation-free methods [accelerated failure time model (AFT), differential abundance analysis with Bayes shrinkage estimation of variance method (DASEV), mixture model (MM), two-part model (TP)], or with imputing missing values using BPCA, HM, KNN, QR or RF followed by inferential testing with a parametric (two-sample t-test) or non-parametric (Wilcoxon rank sum) test.

by imputation. QR imputation increased both Delta and SD of the imputed data relative to the complete data with the magnitude of the bias strongly increasing with missingness (Supplementary Figures S7–S12 available online at <http://bib.oxfordjournals.org/>). Because QR imputation uses the distribution of compounds within a subject as the basis for imputation, the overall level of missingness was strongly influential. HM also increased Delta in the imputed data when the level of missingness in individual compounds was less than 50% but reduced Delta at higher missingness as more values were imputed as one-half the minimum; the SD also increased with increasing missingness. KNN was largely unbiased with respect to Delta and SD although both were biased slightly smaller for compounds with 10–30% missingness. BPCA and RF tended to reduce Delta and SD of the imputed data with larger effects as compound-level missingness increased. Alterations in ES are explored in more detail in Study 2. These patterns of results were broadly similar with the 80% power simulations (see Supplementary Table S3 and Supplementary Figures S13–S24 available online at <http://bib.oxfordjournals.org/>).

Study 2

Study 2 focused on the joint effects of sample size and missingness. As with Study 1, the level of missingness strongly affected results of statistical testing of all methods with BLCI declining as missingness increased. With 5% missingness, all imputation

methods with t-tests or Wilcoxon tests have BLCI greater than 0.75. BLCI drops to about 0.5 by 30% missingness and below 0.5 by 70% missingness. The imputation-free methods followed a similar pattern (Figure 4; Supplementary Figure S25 and Supplementary Table S4 available online at <http://bib.oxfordjournals.org/>).

The overall effect of sample size on BLCI was small. For imputation methods except HM, BLCI increased a small amount with increasing sample size when missingness was 50% or higher. For example, at 50% missingness, median BLCI for KNN with a t-test was about 0.5 with a sample size of 10 and this increased to about 0.55 with a sample size of 50. In interpreting these results, it is important to remember that to simulate data for Study 2, we decreased the ES with increasing sample size in order to keep the number of statistically significant compounds in the complete data sets approximately the same for each sample size (Supplementary Table S1 available online at <http://bib.oxfordjournals.org/>). With this approach, we avoided potential confounding from increased statistical power with increasing sample size.

Results of Study 2 generally corroborated findings from Study 1 in terms of results of statistical tests for these procedures. QR with Wilcoxon had the highest or penultimate median BLCI for all sample sizes for missingness $\leq 30\%$. With higher missing levels, KNN and RF were typically best in terms of BLCI as was found in Study 1 (Supplementary Table S4 available online at <http://bib.oxfordjournals.org/>). RF tended to have higher BLCI than KNN in Study 1 while KNN was higher in

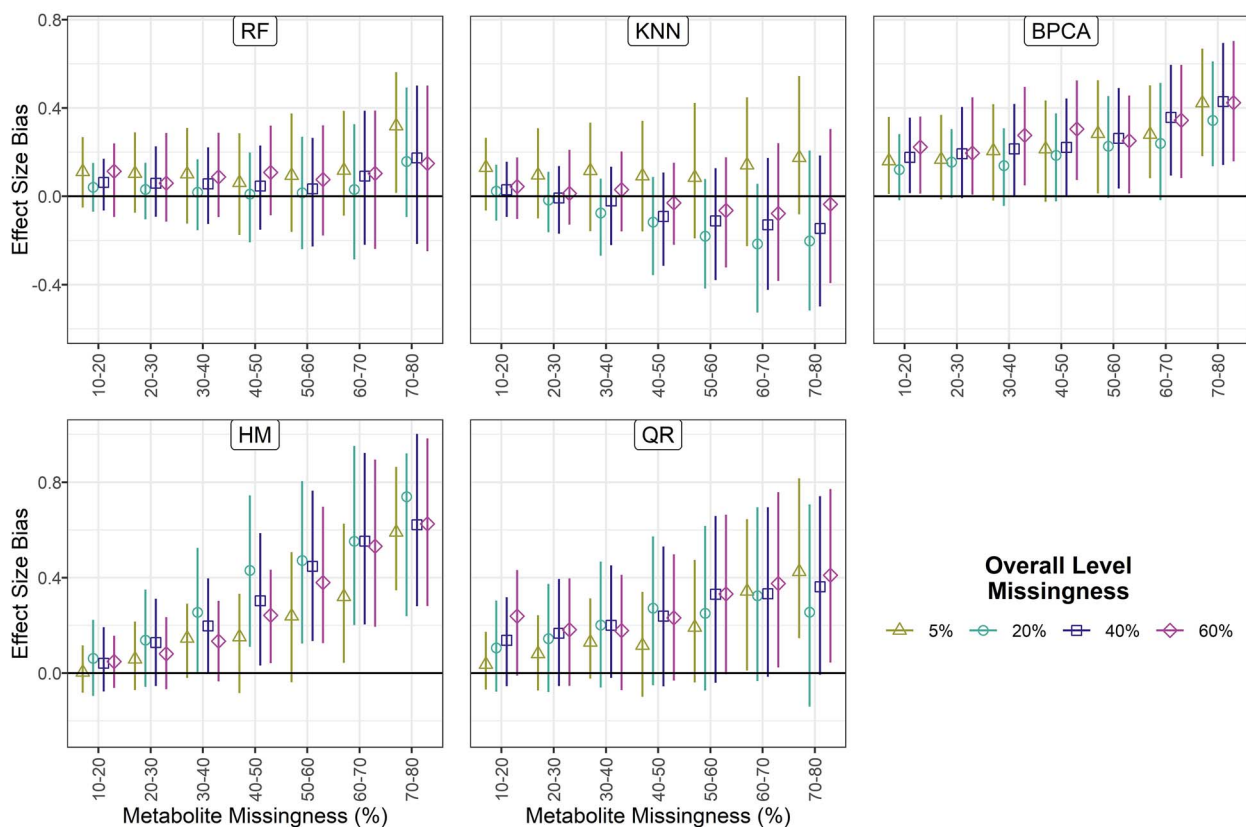


Figure 3. Bias (median and quartiles) in ES of each imputation method for varying levels of missingness in each compound and overall level of missingness in 100 data sets simulated in Study 1 from the PKD data set with an ES of 0.75SD. Missing values were imputed with RF, KNN, BPCA, HM or QR. Negative numbers indicate larger ES in imputed versus complete data.

Study 2 reflecting less variation in KNN's performance relative to sample size than for RF. QR and HM had higher BLCI in combination with a Wilcoxon test than a t-test. However, for KNN, RF and BPCA, BLCI results were largely similar with the two statistical testing procedures. The imputation-free methods were intermediate to the imputation methods, performing better than some imputation methods and worse than others (Figure 4; Supplementary Figure S25 and Supplementary Table S4 available online at <http://bib.oxfordjournals.org/>).

Because all compounds had the same level of missing values for a given simulation in Study 2, discerning the effects of the imputation methods on ES was more straightforward than in Study 1. The methods varied substantially in their effects on SD and Delta (Supplementary Figures S27–S30 available online at <http://bib.oxfordjournals.org/>). KNN yielded little bias in SD, only slightly reducing the SD regardless of sample size and missingness (median bias <0.05 for PKD and <0.03 for HALT). With RF and BPCA, SD was decreased a little more than KNN but bias remained low (less than about <0.1 for both PKD and HALT) across the range of sample sizes and missingness levels. In contrast, QR and HM imputation markedly increased SD with bias increasing with missingness regardless of sample size. Median increase in SD was up to about 0.45 for HM and 0.4 for QR (Supplementary Figures S29 and S30 available online at <http://bib.oxfordjournals.org/>). Sample size had little effect for QR imputation since this method imputes on a per-sample basis, but for HM, increasing sample size increased the bias

in SD (Supplementary Figures S29 and S30 available online at <http://bib.oxfordjournals.org/>).

Under KNN, Delta remained largely unbiased (Supplementary Figures S27 and S28 available online at <http://bib.oxfordjournals.org/>) regardless of sample size and missingness. For the PKD-based simulations, BPCA and RF had larger but still modest effects on Delta relative to KNN tending to reduce Delta with increasing missingness and smaller sample sizes. With the HALT-based simulations, bias in Delta was similar for BPCA while the bias in Delta for RF was less than with the PKD-based simulations. QR and HM imputation increased the difference in means between the groups and the bias in Delta increased markedly as the level of missingness increased. With the PKD-based simulations, for HM at the highest levels of missingness ($\geq 60\%$), the bias declined because many missing values were imputed to the same value but this was not seen in the HALT-based simulations.

Distortion of ES by imputation was greatest at the smallest sample size ($N=10$) with bias decreasing as sample size increased (Figure 5; Supplementary Figure S26 available online at <http://bib.oxfordjournals.org/>). Bias also generally increased with increasing missingness with the largest bias for high missing percentages and small sample sizes. With the PKD-based simulations, QR and KNN were largely unbiased up to about 60 and 40% missingness, respectively. This held true for KNN in the HALT-based simulations but not for QR, which reduced ES at 60% missingness (Supplementary Figure S26 available online at <http://bib.oxfordjournals.org/>). In the PKD-based simulations,

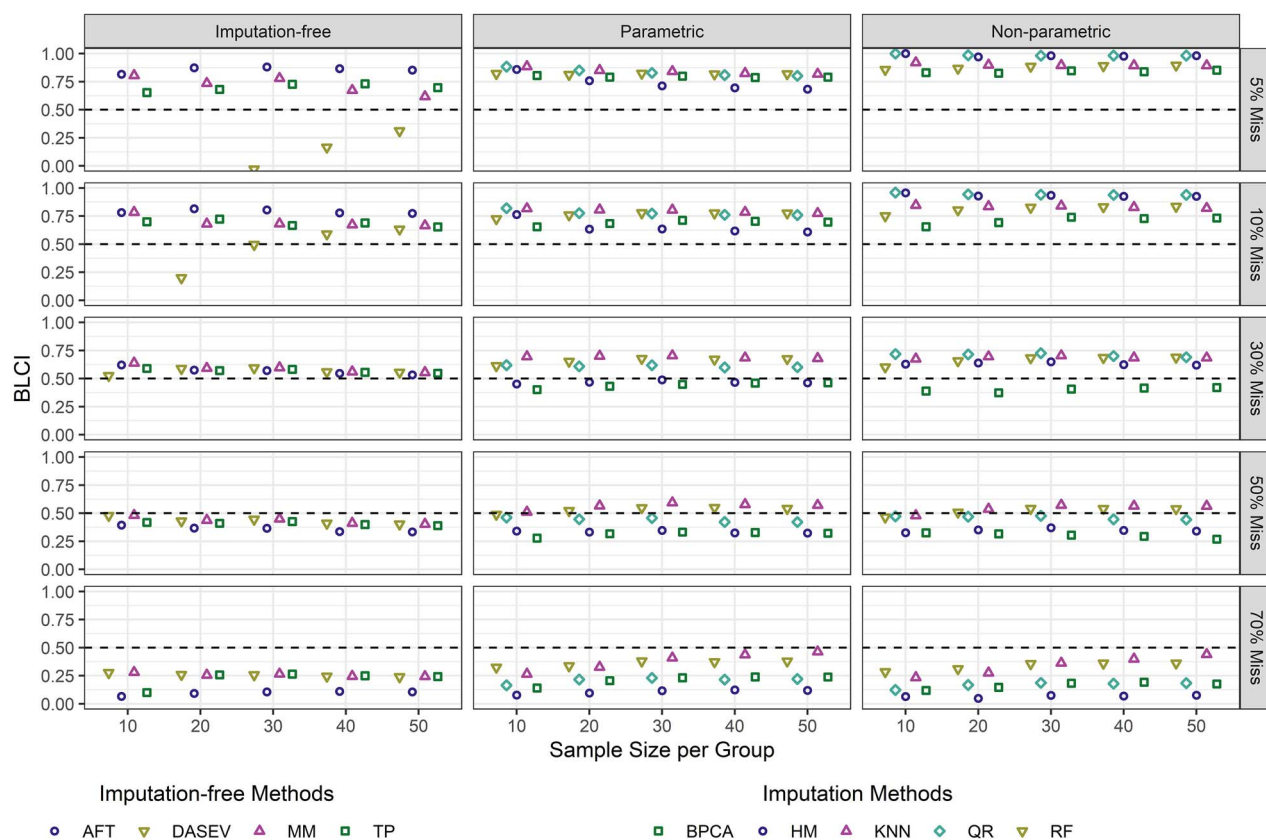


Figure 4. Median BLCI versus sample size (sample size per group) for 1000 simulated data sets based on the PKD covariance matrix in Study 2 at varying levels of missing values. Groups were compared using imputation-free methods [accelerated failure time model (AFT), differential abundance analysis with Bayes shrinkage estimation of variance method (DASEV), mixture model (MM), two-part model (TP)], or with imputing missing values using BPCA, HM, KNN, QR or RF followed by inferential testing with a parametric (two-sample t-test) or non-parametric (Wilcoxon rank sum) test.

HM, BPCA and RF generally reduced ES (Figure 5). The bias in ES increased with the level of missingness but decreased with larger sample sizes (Figure 5). Bias was greatest for HM and smallest for RF (Figure 5); these effects were less marked in the HALT-based simulations (Supplementary Figure S26 available online at <http://bib.oxfordjournals.org/>). For QR, the increase in SD and Delta were largely offsetting resulting in ES remaining unbiased on average but highly variable. Although KNN did not have much effect on Delta and SD, the combined small changes in Delta and SD increased ES at high levels of missingness (>40%). RF increased ES for HALT-based simulations but decreased it in PKD reflecting slightly different magnitudes of effects on the Delta in these data sets. In PKD-based simulations, BPCA tended to reduce both the SD and Delta, which combined to reduce ES (PKD-based simulations) or maintain ES similar to the complete data (HALT-based simulations).

In Study 2, data were simulated from a multivariate normal distribution, which might not accurately reflect the distribution of real data. Thus, we repeated Study 2 by resampling from the HALT data to generate data sets with desired sample sizes. With the exception of QR, the results of this resampling study were qualitatively and quantitatively similar to the results of Study 2 described above. BLCI and bias were notably worse (lower BLCI and greater bias) for QR in the resampling study than found in Study 2. In Study 2, samples were drawn from a multivariate normal distribution, which is consistent with the distributional assumption of the QR imputation method. As a result, QR performance was better in Study 2 than in the resampling study where

the distribution of compounds within a sample could have deviated from normality. The methods and graphical presentation of results from the resampling study are provided in Supplementary Materials 3 (Study 2 with Data Sets Generated through Resampling) available online at <http://bib.oxfordjournals.org/>.

Discussion

A consistent pattern in our evaluations was a marked decline in BLCI with increasing missingness. In Study 2, the median BLCI for all methods was below 0.6 once the level of missingness per compound reached 40–50% regardless of sample size. Because 20% of compounds in Study 2 had no missing values, these percentages represent overall missingness of 32–40%. Sensitivity and specificity were differentially affected by the level of missingness among the methods. Because imputation tended to reduce the ES with increasing missingness, sensitivity was more affected than specificity. Nevertheless, sensitivity remained relatively high for KNN, RF, TP model and MM model as missingness increased but declined for the other methods while specificity tended to show the opposite pattern. Thus, in selecting an analytical approach, investigators should consider their tolerance for false positives and false negatives, particularly when pre-screening for differentially regulated compounds for a large number of compounds for downstream analysis such as pathway analysis and functional enrichment analysis. High sensitivity favors correctly identifying truly differentially regulated compounds but potentially with many false positives while

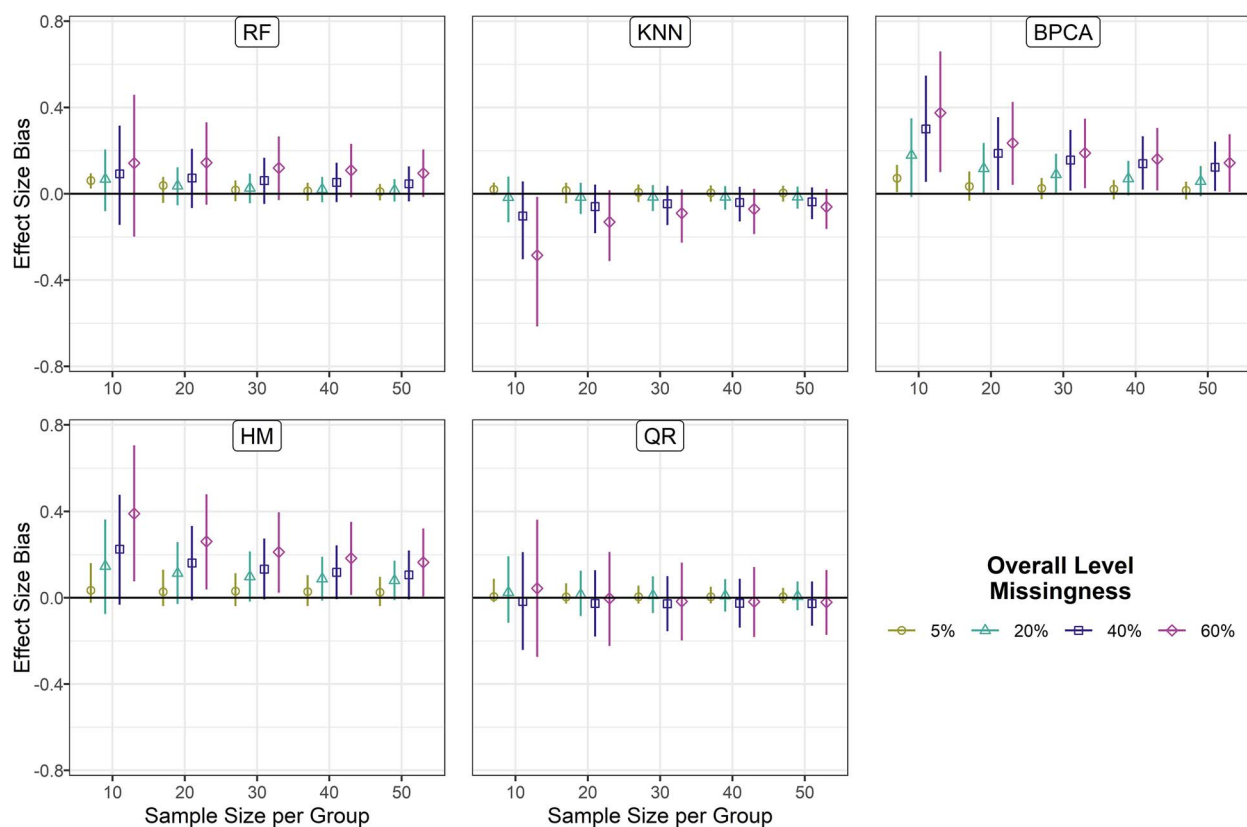


Figure 5. Bias (median and quartiles) in ES of each imputation method versus sample size (N per group) for 1000 simulated data sets based on the PKD covariance matrix in Study 2 at varying levels of missing values. Missing values were imputed with RF, KNN, BPCA, HM or QR. Negative numbers indicate larger ES in imputed versus complete data.

high specificity favors correctly identifying compounds that are not truly differentially regulated at the expense of missing some true positives.

For the imputation methods, changes in sensitivity and specificity reflect the effects of imputation on ES. Sensitivity declined if imputation reduced ES but remained high for methods that increased ES. Conversely, specificity remained high if ES was reduced but declined among methods that increased ES.

Wilcoxon tests yielded higher BLCI than t -tests following QR and HM imputation. These methods assume that data are missing entirely due to detection limit censoring. Therefore, imputed values are small and clustered around a presumed detection limit. With QR, imputation is done on a per sample basis and imputes values below all other values in the sample. For HM, imputed values are half the minimum observed value for each compound across all the samples. These approaches substantially alter a compound's distribution but because non-parametric inferential methods are less affected by actual values than parametric methods, QR and HM imputation performed well with the rank-based Wilcoxon test but not with a t -test.

The distortion in Delta and SD caused by QR and HM is an important consideration in using these methods. Often investigators report a fold change as a measure of the relative magnitude of the difference between groups. Given the effects of QR and HM on Delta, the fold change would be affected which could affect interpretation of the results as well as analyses of regulatory pathways. Multivariate analysis methods such as PCA are also commonly used to discriminate groups in MS analyses,

and the effectiveness of these methods could be compromised by the effects of HM and QR imputation on Delta and SD.

To refine our recommendations, we first identified methods that provided median sensitivity and specificity greater than 80% based on Study 2 and secondarily relaxed the specificity threshold to $>60\%$ (Supplementary Tables S5–S8 available online at <http://bib.oxfordjournals.org/>). No methods provided both sensitivity and specificity $>80\%$ when the percentage of missingness was $>40\%$ for HALT-based simulations and $>30\%$ for PKD-based simulations. For $\leq 20\%$ missingness, KNN, RF and QR with Wilcoxon had 80% specificity and sensitivity for both HALT- and PKD-based simulations. At 30% missingness, QR with Wilcoxon met these criteria for both data sets. With HALT-based simulations, KNN with Wilcoxon met criteria for $\leq 30\%$ missingness and for 40% missingness when $N > 20$. The same methods remained strong performers with relaxing specificity to 60%. Notably however, KNN with Wilcoxon provided 80% sensitivity and 60% specificity for missingness levels of 50 or 60% depending on sample size. Finally, it is important to recognize that at $N = 10$, sensitivity and specificity were only acceptable at the lowest levels of missingness, which suggests that a minimal number of observed data is required to produce reliable results particularly for small studies.

Comparison with other imputation evaluations

The imputation methods we considered have been previously studied. In a microarray setting with sample sizes ranging from 8 to 60, Chiu *et al.* [14] found KNN to be one of the best methods

while BPCA was poor. Missingness was relatively low $\leq 20\%$ and was generated completely at random (MCAR) in this study. Do et al [9] explored performance under different causes of missingness including a fixed censoring level, probabilistic censoring and completely random; missingness ranged from 10 to 70% but with a very large sample size ($N=1750$). They concluded that KNN was the most robust method. Hrydziuszko and Viant [1] evaluated several imputation methods in a MS metabolomics context with missingness up to 20% either MCAR or not at random (MNAR). HM did poorly while BPCA and KNN did well; overall KNN was recommended. Kokla et al. [10] simulated data with up to 30% missingness data under different combinations of MCAR, missing at random (MAR) and MNAR missingness. RF followed by KNN were their best performers overall although under all MNAR missingness, imputing with the minimum observed value was better than other methods. Lazar et al. [11] had similar results with KNN doing well with more MCAR missingness and minimum imputation methods performing better with higher amounts of MNAR. Finally, Wei et al. [12] found RF best when MCAR and MAR mechanisms predominated but for left-censored MNAR data, QR was favored.

Our results are largely consistent with previous studies. We found RF and KNN to be strong, consistent performers across a range of conditions; they were often the best with missingness greater than 20%. BPCA was an intermediate performer consistent with [10, 14] although Hrydziuszko and Viant [1] reported good results with BPCA. We also found QR to be effective for statistical inference when paired with a non-parametric (rank-based) test for missingness levels up to about 30%.

Comparison with previous evaluations of imputation-free methods

Only a few studies have considered the imputation-free methods. Tekwe et al. [6] compared the AFT model to row-mean, KNN and probabilistic PCA imputation and found the AFT model to identify the most differentially expressed proteins as the percentage of censored values increased to 45%. In our results, the AFT model had high sensitivity ($>80\%$) primarily for missingness levels $\leq 20\%$ but performance declined with higher missingness. In their modeling, all missing values arose due to censoring and thus the AFT model correctly reflected the missing data mechanism. In real MS data and reflected in our simulations, missing values arise due to a combination of censoring at a detection limit and random technical issues making the AFT model less effective than other methods as missingness increased. Further, the authors did not report false positives.

Huang et al. [21] reported DASEV to outperform the MM, AFT and TP models in inferential testing. In our results, when the percentage of missingness was 30% or greater, DASEV yielded higher BLCI than both AFT and TP models but was similar to the MM model. With less than 20% missingness, DASEV did not perform well in our simulations. At $\leq 10\%$ missingness and at small sample sizes, DASEV generated an error for many compounds because the simulated data did not have at least one missing value in each group. We calculated sensitivity and specificity based on all compounds with at least one missing value even if DASEV returned an error. This approach resulted in very low values for DASEV for some simulation settings.

Taylor et al. [4] compared the MM model to the AFT model and KNN with a t-test for missing values of 25, 50 and 75%. Missing values consisted of compounds absent from a sample plus censored values. The AFT model had higher power than the MM model in this work while KNN with t-test had lower

power. Here we found the AFT model to have higher BLCI than the MM model for missingness up to 20 or 30% but for the MM model to perform better at higher levels of missingness. Further, we found KNN to outperform the AFT and MM models. The difference is likely due to how missing values were simulated in the two studies. Taylor et al. [4] expressly modeled missing values as a point mass in combination with censored values, whereas in the current study, missing values could arise from random technical variation or censoring although censoring did not entail a hard threshold. KNN assumes MAR or MCAR which was not the setting in [4]. Further, in [4], every compound had the specified level of missingness, whereas here some compounds had no missing values. Because KNN uses information from other compounds and samples, the lack of complete compounds could have affected this method's performance in the prior study.

Overall empirical recommendations

Sample size did not strongly affect the results except at $N=10$ where even moderate levels of missingness markedly reduced statistical power. None of the imputation-free methods performed consistently better than the imputation methods nor were there particular conditions under which they were favored (e.g. high missingness or small or moderate sample sizes). QR imputation with Wilcoxon test often had good inferential testing outcomes but this method substantially distorted Delta and SD. We only recommend this method for compounds with $\leq 10\%$ missingness. For greater missingness, KNN and RF are good choices for statistical testing with little distortion in Delta and SD.

A common practice is to drop compounds with more than a set level of missingness, often 30–50% of samples per compound. We found that no method retained at least 80% sensitivity and specificity with missingness greater than 30% lending support for that threshold. With relaxing minimum specificity to 60%, KNN yielded 80% sensitivity for missingness levels up to 60% depending on the sample size. Thus, compounds with up to 50% missingness could be retained if there is tolerance for more false positives. Beyond 50% missingness, sensitivity and/or specificity were low for most methods, and thus, analysis results should be interpreted with caution.

Key Points

- Missing values are common in high-throughput mass spectrometry data.
- Two general strategies are available to address missing values: (1) eliminate or impute the missing values and apply standard statistical methods for complete data and (ii) use statistical methods that specifically account for missing values.
- Statistical testing with methods under both approaches is strongly affected by the amount of missing values but is less influenced by sample size.
- Random forest and k -nearest neighbor imputation combined with a Wilcoxon test perform well for statistical testing with up to 50% missingness. Quantile regression imputation with a Wilcoxon test has good statistical testing outcomes but substantially distorts the mean differences.
- Imputation-free methods do not perform consistently or markedly better than imputation methods.

Data availability

No new data were generated or analyzed in support of this research.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgements

We thank Dr Robert Weiss (University of California Davis) for providing the PKD, RCC and HALT data for this study.

Funding

National Center for Advancing Translational Sciences (UL1 TR001860); National Institute of Child Health and Human Development (P50 HD103526); National Institute of Aging (P01AG062817); National Institute of Environmental Health Sciences (P30ES023513).

References

- Hrydziuszk O, Viant MR. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics* 2011;**8**(S1):161–74.
- Wang X, Anderson GA, Smith RD, et al. A hybrid approach to protein differential expression in mass spectrometry-based proteomics. *Bioinformatics* 2012;**28**(12):1586–91.
- Webb-Robertson BJ, Wiberg HK, Matzke MM, et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* 2015;**14**(5):1993–2001.
- Taylor SL, Leiserowitz GS, Kim K. Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies. *Stat Appl Genet Mol Biol* 2013;**12**(6):703–22.
- Clough T, Key M, Ott I, et al. Protein quantification in label-free LC-MS experiments. *J Proteome Res* 2009;**8**:5275–87.
- Tekwe CD, Carroll RJ, Dabney AR. Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data. *Bioinformatics* 2012;**28**(15):1998–2003.
- Taylor SL, Pollard KS. Hypothesis tests for point-mass mixture data with application to 'omics data with many zero values. *Stat Appl Genet Mol Biol* 2009;**8**(1):Article 8.
- Karpievitch Y, Stanley J, Taverner T, et al. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* 2009;**25**(16):2028–34.
- Do KT, Wahl S, Raffler J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* 2018;**14**(10):128.
- Kokla M, Virtanen J, Kolehmainen M, et al. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics* 2019;**20**(1):492.
- Lazar C, Gatto L, Ferro M, et al. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J Proteome Res* 2016;**15**(4):1116–25.
- Wei R, Wang J, Su M, et al. Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci Rep* 2018;**8**(1):663.
- Liu M, Dongre A. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Brief Bioinform* 2021;**22**(3):1–22.
- Chiu C-C, Chan SY, Wang CC, et al. Missing value imputation for microarray data: a comprehensive comparison study and a web tool. *BMC Syst Biol* 2013;**7**:S12.
- Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;**17**(6):520–5.
- Oba S, Sato MA, Takemasa I, et al. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 2003;**19**(16):2088–96.
- Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;**28**(1):112–8.
- Lee M, Rahbar MH, Brown M, et al. A multiple imputation method based on weighted quantile regression models for longitudinal censored biomarker data with missing values at early visits. *BMC Med Res Methodol* 2018;**18**(1):8.
- Muñoz JF, Rueda M. New imputation methods for missing data using quantiles. *J Comput Appl Math* 2009;**232**(2):305–17.
- Lachenbruch PA. Comparisons of two-part models with competitors. *Stat Med* 2001;**20**(8):1215–34.
- Huang Z, Lane AN, Fan TW, et al. Differential abundance analysis with Bayes shrinkage estimation of variance (DASEV) for zero-inflated proteomic and metabolomic data. *Sci Rep* 2020;**10**(1):876.
- Kim K, Mall C, Taylor SL, et al. Mealtime, temporal, and daily variability of the human urinary and plasma metabolomes in a tightly controlled environment. *PLoS One* 2014;**9**(1):e86223.
- Kim K, Taylor SL, Ganti S, et al. Urine metabolomic analysis identifies potential biomarkers and pathogenic pathways in kidney cancer. *OMICS* 2011;**15**(5):293–303.
- Kim K, Trott JF, Gao G, et al. Plasma metabolites and lipids associate with kidney function and kidney volume in hypertensive ADPKD patients early in the disease course. *BMC Nephrol* 2019;**20**(1):66.
- Scheel I, Aldrin M, Glad IK, et al. The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics* 2005;**21**(23):4272–9.