Universiteit
Leiden
The Netherlands

Multi-omics in research: epidemiology, methodology, and advanced data analysis

Faquih, T.O.

Citation
Faquih, T. O. (2023, March 28). *Multi-omics in research: epidemiology, methodology, and advanced data analysis*. Retrieved from https://hdl.handle.net/1887/3589838

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3589838

Note: To cite this publication please use the final published version (if applicable).

Chapter 3
A Workflow for Missing Values Imputation of Untargeted Metabolomics Data

Tariq Faquih [1], Maarten van Smeden [2], Jiao Luo [1], Saskia le Cessie [1,3], Gabi Kastenmüller [4,5], Jan Krumsiek [6], Raymond Noordam [7], Diana van Heemst [7], Frits R. Rosendaal [1],Astrid van Hylckama Vlieg [1], Ko Willems van Dijk [8,9,10] and Dennis O. Mook-Kanamori [1,11,12,*]

[1] Department of Clinical Epidemiology, Leiden University Medical Center, Postal Zone C7-P, PO Box 9600, 2300 RC Leiden, The Netherlands; T.O.Faquih@lumc.nl (T.F.); J.Luo@lumc.nl (J.L.); S.le_Cessie@lumc.nl (S.l.C.); F.R.Rosendaal@lumc.nl (F.R.R.); A.van_Hylckama_Vlieg@lumc.nl (A.v.H.V.)

[2] Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, 8, 3584 Utrecht, The Netherlands; M.vanSmeden@umcutrecht.nl

[3] Department of Biomedical Data Sciences, Section Medical Statistics and Bioinformatics, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands

[4] Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, 85764 Neuherberg, Germany; g.kastenmueller@helmholtz-muenchen.de

[5] Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, 85764 Neuherberg, Germany

[6] Department of Physiology, Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10065, USA; jak2043@med.cornell.edu

[7] Department of Internal Medicine, Section of Gerontology and Geriatrics,
Leiden University Medical Center, 2333ZA Leiden, The Netherlands; R.Noordam@lumc.nl (R.N.); D.van_Heemst@lumc.nl (D.v.H.)

[8] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl

[9] Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands

[10] Department of Human Genetics, Leiden University Medical Center, 2, 2333 Leiden, The Netherlands

[11] Department of Public Health and Primary Care, Leiden University Medical Center, 2, 233 Leiden, The Netherlands

[12] Metabolon Inc., Morrisville, NC 27560, USA

***** Correspondence: D.O.Mook@lumc.nl

Metabolites. **2020**. 10:486. doi:10.3390/metabo10120486 PMID: 33256233

1 ABSTRACT:

Metabolomics studies have seen a steady growth due to the development and implementation of affordable and high-quality metabolomics platforms. In large metabolite panels, measurement values are frequently missing and, if neglected or sub-optimally imputed, can cause biased study results. We provided a publicly available, user-friendly *R* script to streamline the imputation of missing endogenous, unannotated, and xenobiotic metabolites. We evaluated the multivariate imputation by chained equations (MICE) and k-nearest neighbors (kNN) analyses implemented in our script by simulations using measured metabolites data from the Netherlands Epidemiology of Obesity (NEO) study (n = 599). We simulated missing values in four unique metabolites from different pathways with different correlation structures in three sample sizes (599, 150, 50) with three missing percentages (15%, 30%, 60%), and using two missing mechanisms (completely at random and not at random). Based on the simulations, we found that for MICE, larger sample size was the primary factor decreasing bias and error. For kNN, the primary factor reducing bias and error was the metabolite correlation with its predictor metabolites. MICE provided consistently higher performance measures particularly for larger datasets (n > 50). In conclusion, we presented an imputation workflow in a publicly available *R* script to impute untargeted metabolomics data. Our simulations provided insight into the effects of sample size, percentage missing, and correlation structure on the accuracy of the two imputation methods.

Keywords: imputation; multiple imputation using chained equations; k-nearest neighbors; untargeted metabolomics; metabolon; simulation; workflow

2 INTRODUCTION

Metabolomics studies have seen a steady growth due to the development and implementation of affordable and high-quality metabolomics platforms. These platforms can be split into two categories: targeted and untargeted metabolomics platforms based on their approach to metabolite identification [1,2]. Targeted platforms are focused on a known prespecified set of metabolites, while untargeted platforms aim to detect as many metabolites as possible in the sample without the need for explicit prior knowledge of their identity. The metabolite signatures detected (i.e., mass to charge ratio, m/z, or retention times) are subsequently matched in a metabolite library to determine their identity. Currently, both targeted and untargeted platforms can detect over 1000 metabolites in a single biological sample (e.g., blood, saliva, and urine). A typical issue with both these platform types is missing values from the measurement.

Missing values in metabolomics data are problematic for subsequent analyses, may be neglected, and are often mishandled or ignored. A common misconception is that missing values in metabolomics data are exclusively due to metabolites with a very low concentration, i.e., below the limit of detection of the instrument. Although in many circumstances the majority of missing values can be due to low concentrations, it has been shown that missing values can also be caused by biological and/or technical variation [3–5]. Based on the assumption that not reaching the limit of detection exclusively causes missingness, missing values are often handled with one or more of the following procedures:

For each metabolite the missing values are replaced (“imputed”) with a single value, such as the minimum detection level or half the minimum detection level. This approach results in overrepresentation of a single value in the population distribution. This may affect subsequent analyses and may cause biased results, regardless of the cause of missing values [5,6]. Furthermore, metabolites could be missing in some individuals because they are not biologically present in their

system. Therefore, imputing these missing values will cause bias in the analysis. For example, if the metabolites for metformin are imputed, both diabetic patients who use the drug as well as and non-diabetic individuals who do not use the drug will have values for the it. This is a prominent issue in platforms such as Metabolon™ (Metabolon Inc., Durham, NC, USA) that include xenobiotic metabolites (e.g., metabolites from external sources such as medications).

Metabolites with a missing percentage above an arbitrary cut-off value (for example 20%) are removed from the dataset due to "too much missingness" regardless of the metabolite identity. By applying a cut-off above which metabolites are removed from the dataset, or, in the most extreme case only using the complete cases, data are unnecessarily discarded, that could have been of importance to the research question. Furthermore, this exclusion can affect further pathway analysis, such as metabolite set enrichment analysis, that explore possible pathway connections for the measured metabolites [7].

Several studies have evaluated imputation methods for metabolomics data. The consensus from these studies has so far been that imputation using half the minimum value leads to more bias than other methods and, consequently, this method is discouraged [3,8]. One alternative imputation method that has been recommended for metabolomics is the k-nearest neighbors (kNN) imputation [6,9]. An extensive simulation was performed that evaluated and compared 31 methods of imputation in a simulated untargeted metabolomics data provided by the Metabolon™ platform [6]

These methods included univariate methods such as half-minimum imputation and multivariate methods such as variations of kNN and multivariate imputation by chained equations (MICE). Two methods were concluded to have the best performance:

kNN on observations with variable pre-selection ("kNN-obs-sel"), a two-step method that incorporates the standard kNN algorithm with a preselection of a group of metabolites that are most correlated with the metabolite with missing values (i.e., auxiliary metabolites). Therefore, the neighbors selected by kNN will have similar metabolomic profiles [6].

MICE using the predictive mean matching method ("MICE-pmm"). Like kNN-obs-sel, the most correlated metabolites were used for the imputation. The imputed values are then selected from distribution of possible values to produce multiple imputed datasets [10–12].

In this paper, we expand upon the meticulous evaluation of the imputation methods by Do et al. [6], which was performed on an older version of the metabolomics platform that detects a smaller set of metabolites (n = 517). Furthermore, we set out to take unannotated (i.e., unidentified metabolites in the library) and xenobiotic metabolites into account. The recent Metabolon™ panel in use (Discovery HD4) has increased the number of metabolites to >1000, which includes more unannotated and xenobiotic metabolites. As more scientists are using metabolomics data in their research, it is helpful to have a user-friendly workflow for imputation using the best available methods. We provided this imputation workflow and a user-friendly *R* script to streamline the imputation of the Metabolon™ HD4 panel using kNN-obs-sel and MICE-pmm. Furthermore, we evaluated the imputations by the script in several scenarios with different missingness conditions by a resampling simulation analysis using measured metabolomics data from the Netherlands Epidemiology of Obesity (NEO) study.

3 RESULTS

3.1 Metabolomic Data Characteristics

Metabolomics measurements in 599 individuals between the ages of 45 and 65 with normal BMI distribution from the NEO study identified 1365 metabolites. Detailed information regarding the population are provided in the methods section and Appendix A. Known metabolites were annotated with their chemical name, super pathway, sub pathway, compound identifiers from various metabolite databases, and information regarding their biochemical properties. A total of 840 metabolites were from various endogenous pathways, 229 metabolites were characterized as xenobiotics, and 296 metabolites were unannotated (lacking information regarding chemical name and pathway). Of the 1365 identified metabolites, 800 (58.6%) contained missing values and the median number of missing metabolites per observation was 228 (38%) (Table 1).

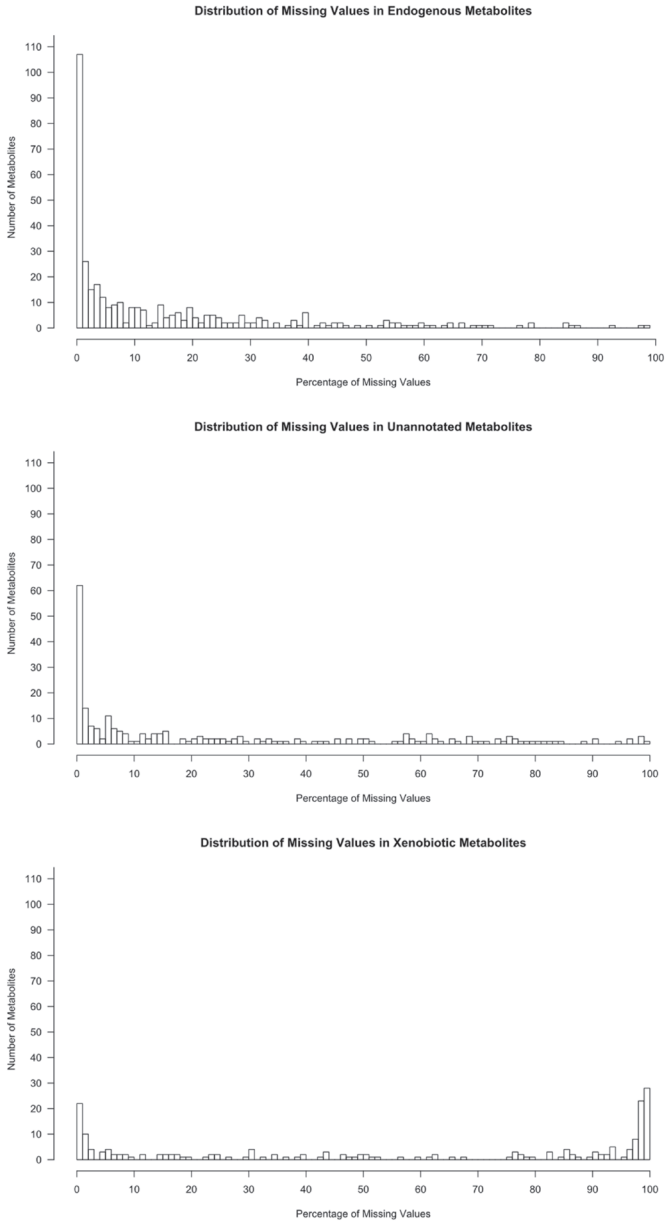
Table 1. Summary of missing data in the Netherlands Epidemiology of Obesity (NEO) study.

Missing Data	Metabolite Groups			
	Endogenous (*n* = 840)	**Unannotated (*n* = 296)**	**Xenobiotics (*n* = 229)**	**Total (*n* = 1365)**
Metabolites with missing values, *n* (%)	367 (43.7)	236 (79.7)	197 (86.0)	800 (58.6)
Missing metabolites per observation, median (range)	57 (23–94)	59 (31–112)	110 (79–149)	228 (152–343)

In the NEO study, 1365 metabolites were measured in 599 individuals (observations).

We plotted the distribution of missing values in each metabolite group (Figure 1). The distribution of the number of missing values of the unannotated metabolites was similar to that of the endogenous metabolites rather than the xenobiotic metabolites. This suggests that most unannotated metabolites are most likely from an endogenous source, similar to the annotated endogenous metabolites, and are most likely expected to be present in all our participants.

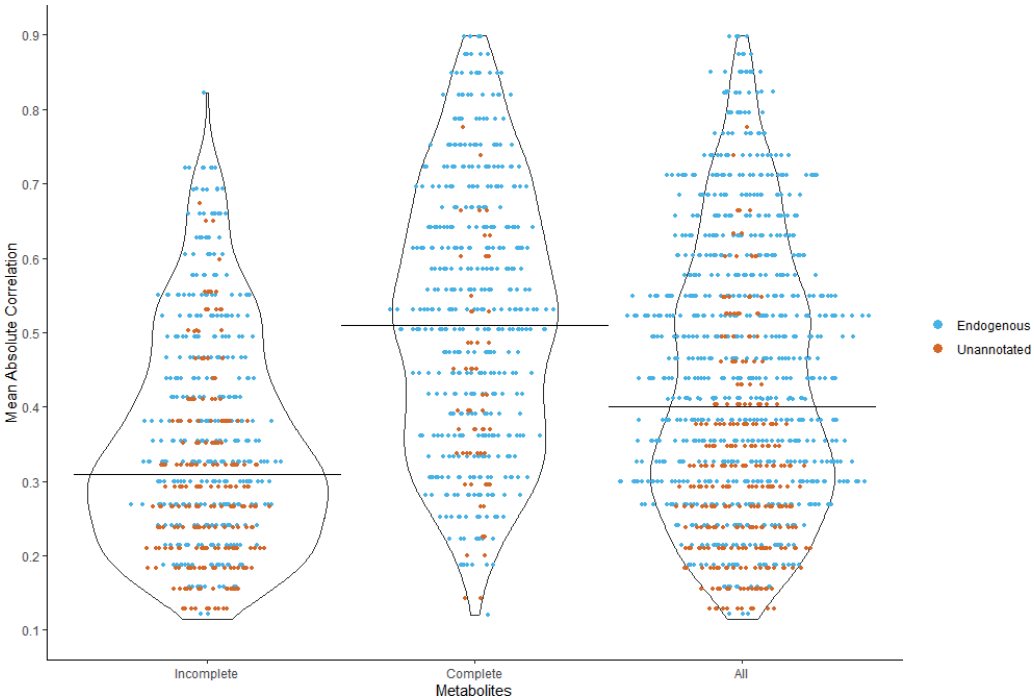
Figure 1. Distribution of the missing values in each metabolite group.



The Pearson's pairwise-complete correlation matrix for the endogenous and unannotated metabolites was calculated using all the metabolites (complete with no missing values and incomplete). For each incomplete metabolite we selected up to 10 complete metabolites with the highest absolute Pearson's correlation (auxiliary metabolites). If the metabolite was not correlated with 10 metabolites (due to high missingness), then we selected the available correlated metabolites. We then calculated the mean value of the Pearson's correlations for these metabolites. Figure 2 shows the distribution of the mean of the auxiliary absolute correlations with further details in Table B1.

The 82% of the incomplete metabolites had a mean absolute Pearson's correlation coefficient lower than 0.5 with their auxiliary metabolites. Overall, the median of the median absolute Pearson's correlation coefficient was 0.4 (0.09–0.89), indicating a generally low intercorrelation between the metabolites.

Figure 2. Distribution of the mean absolute correlations for the complete (without missing values) and incomplete (with missing values) endogenous and unannotated metabolites in the NEO dataset.



3.2 Availability

The imputation script [13] streamlines the workflow by calculating the correlation matrix, selecting the auxiliary metabolites, and imputing the missing values of the metabolites using the provided data from the user. The script requires a dataset, a list of xenobiotic and non-xenobiotic metabolites (endogenous/unannotated), and a choice for the method of imputation (MICE-pmm or kNN-obs-sel). The script and example files can be found at: https://github.com/tofaquih/imputation_of_untargeted_metabolites.

3.3 Performance Evaluation

To evaluate our imputation framework, we applied it to impute metabolites with missing values in the measured NEO dataset (n = 599) using kNN-obs-sel and MICE-pmm. All metabolites were imputed apart from 12 metabolites (3 endogenous, 9 unannotated) in the dataset that had >90% missingness and were subsequently treated as xenobiotic and imputed to 0. As mentioned in the methods section, extremely high missingness limits the amount of data needed to impute the metabolites and to find auxiliary metabolites. High missingness in the 3 endogenous metabolites could have been caused by technical or biological issues, or they

could represent misannotated xenobiotic metabolites. The 9 unannotated metabolites were likely xenobiotic metabolites.

Simulations were performed to compare the performance of the imputation method (MICE- pmm or kNN-obs-sel). As detailed in the Methods section, we generated 144 resampling simulation scenarios, using four metabolites from independent pathways and varying mean correlations with auxiliary metabolites (PC(32:2) (mean absolute correlation = 0.64), urate (mean absolute correlation = 0.49), glutamate (mean absolute correlation = 0.49), succinylcarnitine (mean absolute correlation = 0.36)), three sample sizes (50, 150, 599), three percentage of missing (15%, 30%, 60%), and two missing mechanisms missing mechanisms (missing completely at random (MCAR) and probabilistic limit of detection (PLoD)). The percentage biases from the simulation are presented in Figure 3 and Table 2. Root mean squared errors (RMSE) are shown in Figure 4, Table B5, and Table B6. The mean and standard deviation of the estimates from the simulation are provided in Table B3 and Table B4 using MCAR and PLoD mechanisms, respectively. We used nested loop plots [14] to produce all the figures.

Figure 3. Nested loop plot of the percentage bias of the four metabolites from the simulation. The horizontal axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection.

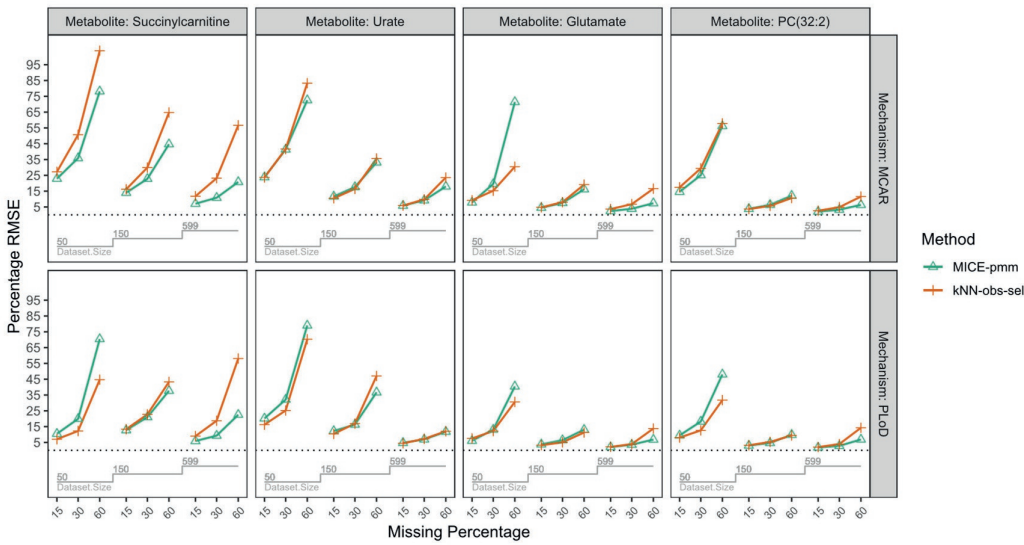


Figure 4. Nested loop plot of the root mean squared error (RMSE) of the four metabolites from the simulation. To simplify comparability in the plot we converted the RMSE values to a percentage by subtracting then dividing the RMSE values by the corresponding true estimates (in sample size n = 599). The horizonal axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection.

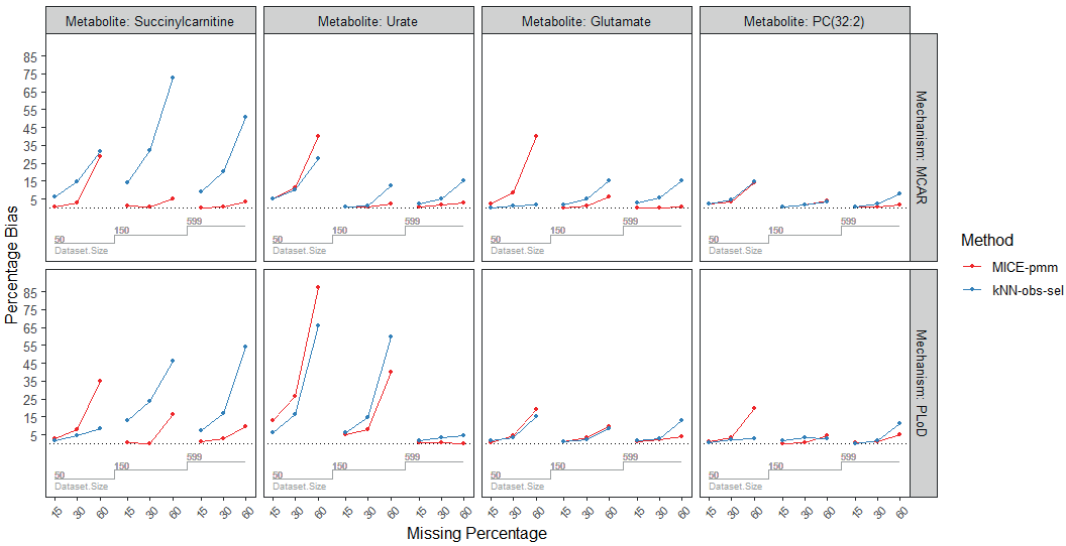


Table 2. Percentage bias of the imputation methods across different parameters on different metabolites including multivariate imputation by chained equations (MICE)-pmm with a single imputation.

Missing Mechanism	Sample Size	Missing Percentage	Metabolites/Imputation Method							
			PC(32:2)		Succinylcarnitine		Glutamate		Urate	
			MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel
MCAR	*n* = 50	15%	2.0	2.1	0.3	6.0	2.2	0.2	4.9	4.9
		30%	3.2	4.5	2.8	14.4	8.6	1.1	11.3	10.2
		60%	13.9	14.5	28.9	31.3	40.2	2.0	39.9	27.6
	n = 150	15%	0.7	0.5	1.4	13.8	0.2	1.9	0.7	0.5
		30%	1.9	1.8	0.4	31.9	1.1	5.1	0.7	1.3
		60%	3.9	3.2	5.2	72.7	6.2	15.0	2.09	12.5
	n = 599	15%	0.3	0.8	0.1	9.3	0.2	2.6	0.6	2.2
		30%	0.7	2.4	0.6	20.1	0.1	5.6	1.5	4.9
		60%	1.9	7.9	3.5	50.9	0.7	15.1	2.6	15.0
PLoD	*n* = 50	15%	1.3	0.8	2.8	1.8	0.7	1.8	12.7	6.2
		30%	3.5	2.2	7.7	4.3	4.2	3.6	26.3	16.5
		60%	19.5	3.0	34.8	8.3	19.3	15.0	87.4	66.1
	n = 150	15%	0.2	1.6	0.7	12.8	1.3	0.8	5.3	6.3
		30%	0.3	3.3	0.2	23.4	3.0	2.3	8.0	14.9
		60%	4.2	2.6	16.2	46.1	9.4	8.6	39.8	59.7
	n = 599	15%	0.5	0.1	0.9	7.3	0.9	1.4	0.5	1.9
		30%	1.1	1.6	2.5	16.7	2.1	2.9	0.6	3.3
		60%	4.9	11.5	9.7	54.3	4.1	13.1	0.1	4.3

Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection.

3.3.1 Sample Size

We observed a decrease of bias and RMSE as the sample size increased (Figures 3 and 4). This trend was consistent for MICE-pmm for each metabolite, with the percentage bias median (range):

8.2 (0.3–87.4) in *n* = 50 decreasing to median (range): 0.8 (0.1–9.7) in *n* = 599. However, increasing sample size did not improve imputation with kNN-obs-sel. Overall, percentage bias was median (range): 4.7 (0.2–66.1) in *n* = 50, median (range): 5.7 (0.5–72.7) in *n* = 150, and median (range): 5.2 (0.1– 54.3). Furthermore, in some scenarios, bias and RMSE increased in larger sample sizes even with the same missing percentage and missing mechanism; this was particularly noticeable for succinylcarnitine (mean absolute correlation = 0.36) where the percentage bias increased in 60% missing from 8.3% in *n* = 50, to 46.1% in *n* = 150, and 54.3% in *n* = 599. Finally, for sample sizes of *n* = 50 and 150, MICE-pmm had lower bias than kNN-obs-sel but a RMSE higher or similar to kNN-obs- sel.

3.3.2 Percentage of Missing

In scenarios with 15% and 30% missing, MICE-pmm and kNN-obs-sel showed low bias and RMSE across all sample sizes. At 15% missing, MICE-pmm had a percentage bias of median (range):

0.7 (0.1–12.7), while kNN-obs-sel had a percentage bias of median (range): 1.9 (0.1–13.8). At 30% missing, MICE-pmm had a percentage bias of median (range): 2.0 (0.1–26.2) and kNN-obs-sel had a percentage bias of median (range): 4.4 (1.1–31.9). Finally, in 60% missing MICE-pmm had a percentage bias of median (range): 7.8 (0.1–87.4) and kNN-obs-sel had a percentage bias of median (range): 14.7 (1.9–72.7). Overall, MICE-pmm had lower bias in all missing percentages than kNN-obs- sel. However, the percentage bias for kNN-obs-sel was often lower than that of MICE-pmm at 30% and 60% missing in *n* = 50.

3.3.3 Correlation Strength with the Auxiliary Metabolites

We compared the percentage bias and RMSE of both imputation methods for the four metabolites to assess the influence of correlation strength of the auxiliary metabolites as shown in Tables 2 and B2 and Figures 3 and 4. We observed that availability of auxiliary metabolites with higher correlation for the imputation greatly reduced the bias and RMSE in both methods. In PC(32:2), the metabolite with the highest mean correlation (mean absolute correlation = 0.64), had the lowest bias overall. Percentage bias was median (range): 1.9 (0.2–19.5) with the MICE-pmm imputation and median (range): 2.3 (0.1–14.5) with kNN-obs-sel imputation. Glutamate (mean absolute correlation = 0.49) had median (range): 2.1 (0.1–40.2) percentage bias with MICE-pmm imputation and median (range): 2.7 (0.2–15.1) with kNN-obs-sel. Similarly, imputation of urate (mean absolute correlation = 0.49) using MICE-pmm had median (range): 3.8 (0.1–87.4) percentage bias and median (range): 6.2 (0.5–66.1) using kNN-obs-sel. In contrast, the percentage bias was much higher for the metabolite with the lowest mean correlation, Succinylcarnitine (mean absolute correlation = 0.36), with median (range): 2.6 (0.1–34.8) percentage bias using MICE-pmm imputation and median (range): 15.5 (1.8–72.7) with kNN-obs-sel. Moreover, the bias reached very high percentages in urate and succinylcarnitine compared to PC(32:2) and glutamate in the *n* = 50 subset.

3.3.4 Missing Mechanisms

We used two mechanisms for missingness, MCAR and PLoD, in our simulations. Since PLoD is fundamentally missing not at random (MNAR), causing lower concentrations to have a higher likelihood of missingness, we examined how PLoD affects the performance of MICE-pmm and kNN-obs-sel compared to MCAR scenarios. MCAR scenarios had a percentage bias median (range): 1.9 (0.1–40.2) with MICE-pmm imputation and median (range): 5.3 (0.2–72.7) with kNN-obs-sel. PLoD scenarios had a percentage bias median (range): 3.2 (~0–87.4) with MICE-pmm imputation and median (range): 4.3 (0.1–66.1) with kNN-obs-sel. However, the RMSE (Figure 4, Tables B5 and B6) was lower in PLoD for MICE-pmm (median (range): 11.1 (1.8–79)) than in MCAR (median (range):

14.3 (2.0–78.0)) and similarly lower for kNN-obs-sel in PLoD scenarios (median (range): 12.1 (2.1–70.3)) than MCAR (median (range): 16.5 (2.6–103.9)). Overall, imputing in PLoD scenarios lead to higher bias but lower RMSE compared to MCAR.

4 DISCUSSION

Several simulation studies have evaluated different imputation methods for missing data in metabolomic datasets [3,6,9,15,16]. Nevertheless, the "half the minimum" method of imputation remains in use despite studies showing its sub-optimal performance [3,6,9,15,16]. In this study, we followed up on previous work and provided a framework and complementary *R* script on GitHub

[13] that streamlines the imputation of untargeted metabolomics data. The script provides univariate imputation of zero for missing values considered to be truly absent in xenobiotics and two options of multivariate imputation methods for the remaining metabolites.

Overall, for the four metabolites we used in the simulation, we observed several factors that influenced the performance of each imputation method with different degrees. In the four metabolites we used, MICE-pmm performed better overall across different simulated scenarios. This performance is especially better in PLoD, which represents a missing mechanism similar to that of real metabolomics data [6]. MICE-pmm performance decreased the most in smaller sample sizes, somewhat less by the metabolite auxiliary correlation and the least by the missing percentage. Interestingly, the negative effect of missing percentage diminished as the sample size increased (n = 150 and n = 599). On the other hand, unlike MICE-pmm, kNN-obs-sel performance was decreased most by a higher percentage of missingness and low metabolite auxiliary correlation, which was not improved by increased sample size. A possible explanation is the nature of the kNN-obs-sel method. kNN-obs-sel focused on finding the nearest neighbors based on the correlated metabolites. If it failed to find strongly correlated metabolites, due to the metabolite naturally a having low correlation or due to a large amount of missing values, it selected weak neighbors. Therefore, even at larger sample sizes (150 and 599) the performance of the kNN-obs-sel method remained poor if the missing percentage was large and the metabolite had poor correlation.

4.1 Advantages and Disadvantages of MICE-pmm for Metabolomics

Unlike kNN imputations, we found few papers in the literature regarding the use of MICE imputation for metabolomics. The MICE-pmm imputation is a more intricate method for generating the imputation values. First, the imputation is repeated multiple times in order to assess the uncertainty of the imputation and provide standard errors of the estimates. Second, MICE-pmm

imputation is more compatible with both normally distributed and skewed metabolites than kNN [10]. Third, MICE imputation utilizes discrete and continues variables for imputation. Therefore, MICE-pmm can include additional biologically relevant predictors and the outcome of the analysis of interest, improving the quality of the imputation [17]. These features explain the robustness of MICE-pmm in situations with low correlated auxiliary variables and high missingness.

However, MICE-pmm has some disadvantages. First, small sample sizes negatively affected the performance of MICE-pmm because this forces duplication and reuse of the same individuals [10]. Second, MICE imputation may require more computational run time and is somewhat more complicated to use than kNN because multiple imputed datasets are generated that require a pooling step for the analysis. We shortened computational time by using the latest *MICE R* package and by setting the number of multiple imputations to 5, which has been shown to be a suitable number of imputations [10]. This caused the running time for the complete imputation using MICE-pmm to be equal to that of kNN-obs-sel for the NEO dataset (n = 599). Furthermore, to test the speed of the script, we duplicated and stacked the NEO dataset to create larger datasets (n = 5400 and n = 20,000); MICE- pmm completed the imputations faster than kNN-obs-sel (Table B7). Third, with MICE-pmm it is not possible to apply further analysis such as lasso regression or random forest, which are common analysis methods used in metabolomics [7,18,19]. This is because MICE-pmm uses multiple datasets with Rubin's Rules to pool the estimates of the analysis per dataset. One solution is to use the kNN- obs-sel method, as it always creates a single dataset for analysis. A second alternative would be to use MICE-pmm with a single imputation [m = 1], which can be specified in our script, and use that single dataset in the multivariate analysis. It should be noted that MICE-pmm with m = 1 still performed better than kNN-obs-sel for the larger sample sizes (see Tables B2 and B6 and Figures B1 and B2).

4.2 Limitations

Several methodological issues should be considered. Firstly, our evaluation was done using 599 samples, limited by available metabolomics data in the NEO study. Although this number is not particularly small, future research should be performed in larger datasets. Secondly, we assumed that all missing xenobiotics values are truly missing and replaced them by zero. This could be explored further by incorporating MICE-pmm or kNN-obs-sel to specifically impute xenobiotic metabolites from the same medication sources in persons taking the medication. Furthermore, it could be possible to use questionnaire and clinical data as imputation predictors in MICE-pmm to impute related xenobiotic metabolites. Thirdly, we did not explore alternative methods for MICE to handle small data sizes, such as regulation and penalization. Fourth, our simulation did not evaluate the variance estimators such as type-I and type-II errors or confidence interval coverage. Fifth, metabolites with very large missingness will have high bias and error in the imputation and should be interpreted with caution. Finally, the data do not provide the explicit cause of the missing values and, therefore, we could only assume if the values were truly missing, missing completely at random, or missing due to other reasons. Future studies which explore the causes of missingness will also allow us to impute the missingness more effectively.

5 MATERIALS AND METHODS

5.1 Population Characteristics

The resampling simulation analyses were performed in the NEO study. This study has been extensively described elsewhere [20] and in Appendix A. The NEO study was accepted by the Medical Ethics committee of the Leiden University Medical Center under protocol P08.109. The study is also registered at clinicaltrials.gov under number NL21981.058.08 / P08.109. All participants gave written informed consent [20]. Fasting state serum samples from a sub-population (*n* = 599) of the NEO study were sent for untargeted metabolomics measurements at Metabolon Inc. (Durham, NC, USA) using their Metabolon™ Discovery HD4 platform. In brief, this process involves four independent ultra-high-performance liquid chromatography mass spectrometry (UHPLC-MS/MS) platforms [21,22]. Two platforms used positive ionization reverse phase chromatography, one used negative ionization reverse phase chromatography, and one used hydrophilic interaction liquid chromatography (HILIC) negative ionization [22]. In total, 1365 serum metabolites were measured which included 840 endogenous, 296 unannotated, and 229 xenobiotic metabolites.

5.2 Imputation Methods

Following our examination of the missing data distribution in the NEO study (Figure 2), we decided the xenobiotic metabolites and non-xenobiotic metabolites (endogenous/unannotated) with different imputations. For xenobiotic metabolites, we assumed missing values are truly missing values. For example, when a medication metabolite concentration is missing, it is most likely that the participant is not taking the medication. Therefore, we decided to impute xenobiotic metabolites to zero, as imputing the values (with MICE, kNN, or half-min) would cause bias due to skewed distribution and false positives. For the non-xenobiotic metabolites (endogenous/unannotated), the missing pattern suggests that the unannotated metabolites are most likely endogenous. Therefore, we decided to impute the endogenous and the unannotated metabolites as a single group using the multivariate imputation methods of MICE-pmm and kNN-obs-sel. For these two multivariate methods, we first estimated a correlation matrix for all applicable/non-xenobiotic metabolites from which to select 10 auxiliary metabolites to be used for imputation.

For non-xenobiotic metabolites, we assumed that they are metabolites with truly missing values only if less than 90% of values were missing. This cut-off was necessary for multiple reasons: 1) it became nearly impossible to find auxiliary metabolites for imputation, 2) unannotated metabolites with high missing values are likely xenobiotic and therefore most likely truly missing, and 3) it became statistically problematic to perform multivariate imputation with such high missingness— particularly in small sample sizes [23].

In this study, we used MICE-pmm with 10 auxiliary metabolites to impute the missing values and generated 5 imputed datasets (m = 5). In addition to the auxiliary metabolites, we included further predictors by adding the clinical variables for the outcome (BMI) and the covariates (age and sex) used in the analysis model for the MICE-pmm imputation. The addition of these variables is required in MICE imputations to avoid bias in the results [23,24]. We used kNN-obs-sel only with 10 auxiliary metabolites to impute the missing values. Details regarding the imputation methods are provided in Appendix A. In our script, we incorporated the *R* package *mice* version 3.6.0 [10] for the MICE-pmm imputations and the package *VIM* version 4.8.0 [25] in the kNN-obs-sel imputations.

5.3 Evaluation Analysis and Missing Value Simulation

For the simulation, the analysis of interest was an ordinary least squares regression model with body mass index (BMI) as the outcome and age, sex, and a selected metabolite as the exposures. For the purpose of our study, BMI was used as the outcome for two reasons: (1) BMI is a variable that was measured in all our participants, and (2) BMI is strongly associated with many metabolites and commonly studied in metabolomics [26].

Four metabolites were used, selected based on the following criteria: (1) the metabolite had no missing values in the original NEO dataset, (2) the metabolite must have a strong association with BMI in our Metabolon™ data as well as in the literature using Metabolon™ [26], (3) the four metabolites must be from different biological pathways, and (4) the metabolites must have different mean correlations with their auxiliary metabolites. We found 6 out of 473 complete endogenous metabolites in NEO that fulfilled these criteria. We then narrowed the selection to one metabolite per pathway. Accordingly, we selected four metabolites: PC(32:2) (mean absolute correlation 0.64) from the lipid super pathway; succinylcarnitine (mean absolute correlation = 0.36) from the energy super pathway, the nucleotide urate (mean absolute correlation 0.49), and the amino acid glutamate (mean absolute correlation 0.49). Information regarding the metabolites is provided in Table 3.

Table 3

Metabolite Full Name	Mean Absolute Correlation	Super Pathway	Sub Pathway	Estimate $n = 599$	Estimate $n = 150$	Estimate $n = 50$
PC(32:2)	0.64	Lipid	Plasmalogen	-4.18×10^{-7}	-3.64×10^{-7}	-4.38×10^{-7}
Urate	0.49	Nucleotide	Purine Metabolism	1.39×10^{-8}	9.58×10^{-9}	9.69×10^{-9}
Glutamate	0.49	Amino Acid	Glutamate Metabolism	1.83×10^{-7}	2.89×10^{-8}	1.66×10^{-8}
Succinylcarnitine	0.36	Energy	TCA Cycle	2.84×10^{-6}	1.53×10^{-6}	4.53×10^{-6}

Abbreviations. Mean absolute correlation: mean of the 10 absolute Pearson's correlations from the metabolite correlation matrix. Estimate is the regression coefficient from the model
BMI~age + sex + metabolite. Therefore, the estimates are the mean increase in BMI per 1 unit increase of the metabolite.

We compared the performance of the two imputation methods by simulating missing values using the NEO dataset ($n = 599$). All simulations were performed on three datasets: the original dataset of 599 participants, and on two randomly sampled sub datasets of size $n = 150$ and $n = 50$. The distribution of age, sex, and BMI was maintained in the sub datasets of 50 and 150 individuals. We used the same sub datasets for the all corresponding simulation scenarios. Generating the subsets with different random sampling did not change the estimates drastically (not shown). It should bepointed out that the selected auxiliary metabolites differed slightly between the sub datasets. Metabolite levels were log transformed and standardized (mean of 0 and variance of 1). We calculated the estimates for each metabolite in the complete datasets separately to be used later for the bias and RMSE calculations. In the different simulation scenarios, we induced different percentages of missingness (15, 30, and 60%), and under two different mechanisms, MCAR and PLoD. In the PLoD missing scenarios, the odds of a value being missing increased as the concentration decreases. The total number of missing values was divided per quantile of the

metabolite as follows: 40% into the lower quantile, 50% into the middle quantile, and 10% in the upper quantile.

The evaluation was done by (1) performing the linear regression analysis and obtaining the estimate of the regression coefficient using the complete metabolites data in each subset (Table 3), (2) simulating missing values, (3) imputing missing values using the two imputation methods, (4) estimating the regression coefficient using the imputed data, and (5) evaluating the difference between the estimate of the complete data for that subset and the estimate using the imputed methods, (6) repeating step 2 to 5 1000 times per simulation scenario. The performance of the imputation methods was evaluated using the following measures: raw bias, which is the difference between the real estimate and the mean of the simulations estimates, which can be a positive or a negative value; percentage bias, which is the raw bias divided by real estimate for easier interpretation and comparison [27]; the RMSE, which is the square root of the mean squared difference between estimated; and true value, this measure combines the bias and variance of the simulated estimates into a single measure and represents the precision of the method [28] (Appendix A).

Thus, in total, we used three datasets (n = 50, 150, 599), four metabolites (PC(32:2), succinylcarnitine, urate, glutamate), three missingness percentages (15%, 30%, 60%), two missing mechanisms (MCAR and PLoD), and evaluation by two imputation methods (kNN-obs-sel and MICE-pmm) for a total of 144 possible scenarios. Each of these scenarios was repeated 1000 times.

5.4 Imputation Workflow

To simplify the procedure of imputing missing data, we wrote an *R* script that calculates the correlation matrix between the different metabolites, selects the auxiliary metabolites with the largest correlation, imputes the xenobiotic metabolites with univariate imputation, and imputes the endogenous metabolites with a multivariate imputation (either kNN-obs-sel or MICE-pmm), which can be found on our GitHub repository [13].

6 CONCLUSIONS

In conclusion, we provided a workflow for handling missing values in untargeted metabolomics data using univariate imputation for xenobiotics and multivariate imputation using MICE-pmm or kNN-obs-sel for endogenous and unannotated metabolites. We further evaluated MICE-pmm and kNN-obs-sel in different simulated scenarios. Our evaluation showed that the performance of both methods is affected by three different factors, namely the metabolite mean correlation with auxiliary metabolites, the sample size, and the missing percentage. For MICE-pmm, sample size was the primary factor affecting bias and error inversely. For kNN-obs-sel, the primary factor affecting bias and RMSE was the metabolite correlation with the predictors, which, when high, can provide low bias and RMSE even in small sample sizes (n = 50). Since most of our metabolites had low mean correlation, MICE-pmm provided consistently higher performance measures than kNN-obs-sel and, as a result, we suggest using MICE-pmm imputation for untargeted metabolomics, particularly for larger datasets (n > 50).

Author Contributions: Conceptualization, T.F., A.v.H.V., K.W.v.D., D.O.M.-K., G.K. and J.K.; formal analysis, T.F.; software, T.F.; funding acquisition, D.v.H., R.N., K.W.v.D. and F.R.R.; methodology, S.l.C., M.v.S. and T.F.;

resources, R.N. and F.R.R.; supervision A.v.H.V., K.W.v.D., D.O.M.-K. and M.v.S.; validation, J.L.; writing— original draft, T.F.; writing—review & editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: The NEO study is supported by the participating Departments, the Division and the Board of Directors of the Leiden University Medical Centre, and by the Leiden University, Research Profile Area 'Vascular and Regenerative Medicine'. The analyses of metabolites are funded by the VENI grant (ZonMW-VENI Grant 916.14.023) of D.O.M.-K. D.v.H. and R.N. were supported by a grant of the VELUX Stiftung [grant number 1156].

J.L. was supported by the China Scholarship Counsel [No. 201808500155]. T.F. was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283].

Acknowledgments: The authors of the NEO study thank all individuals who participated in the Netherlands Epidemiology in Obesity study, all participating general practitioners for inviting eligible participants, and all research nurses for collection of the data. We thank the NEO study group, Pat van Beelen, Petra Noordijk and Ingeborg de Jonge for the coordination, lab, and data management of the NEO study.

Conflicts of Interest: Dennis Mook-Kanamori is a part-time clinical research consultant for Metabolon, Inc. All other co-authors have no conflicts of interest to declare.

7 APPENDIX A

A1. NEO Study Design

The Netherlands Epidemiology of Obesity (NEO) study is a population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of overweight individuals or individuals with obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m^2 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited, irrespective of their BMI. Recruitment of participants started in September 2008 and was completed at the end of September 2012. In total, 6671participants were included, of whom 5217 had a BMI of 27 kg/m^2 or higher. The NEO study was accepted by the Medical Ethics committee of the Leiden University Medical Center under protocol P08.109. The study is also registered at clinicaltrials.gov under number NL21981.058.08 / P08.109. All participants gave written informed consent. Participants were invited to come to the NEO study center of the LUMC for one baseline study visit after an overnight fast. A blood sample of 108 mL was taken from the participants after an overnight fast of at least 10 h [20]. From the Leiderdorp subpopulation (n = 1671) we selected 599 Caucasian individuals with normal BMI distribution and sent their serum samples for metabolomics analysis using the Metabolon platform and for examination in this paper.

A1.1. Evaluation Measures

In addition to using bias and RMSE we also converted these measures to percentages. This was necessary because the estimates of the analysis model of the metabolites in complete NEO dataset varied in magnitude and scale. For example, in sample size n = 599, the estimate for urate was 1.39×10^{-8} and for succinylcarnitine was 2.84×10^{-6} (full details in Table 3 of the main manuscript). Percentage bias was calculated by dividing the bias in each sample size set by the estimate calculated for the respective sample size. RMSE percentage was calculated by subtracting then dividing all scenarios for each metabolite by the corresponding true coefficient in sample size n = 599 and multiplying by 100.

A1.2. Imputation Methods

The first step in our workflow was creating a correlation matrix for the metabolites in the dataset. For each metabolite with missing values (X), we selected the ten metabolites without missing values with the strongest absolute correlation $|r|$ to X from the correlation matrix. Our metabolomics dataset was generated on the latest measuring platform which greatly expanded the number of metabolites but reduced the overall intercorrelation of the data. This reduction of the intercorrelation is partly explained by the inclusion of remote metabolites in smaller pathways.

In standard kNN, distances are used to select closest neighbors to the observation with missing values. In kNN-obs-sel, for each metabolite we used up to 10 auxiliary metabolites as predictors and imputed the missing values by taking the average of the 10 nearest neighbors (K = 10) observations.Multiple imputation using chained equations (MICE) is used for incomplete data in multiple variables and may use discrete, categorical, and continuous variables of different units for the imputation [29]. When using the option predicted mean matching, it yields several different datasets with imputed values obtained from observed cases. The analysis of interest is then performed on each of the imputed datasets separately and the results are pooled afterwards as described by White et al., (2010) [23] and other articles [29–31]. Given that kNN-obs-sel calculates the mean from the auxiliary variables, it was only possible to use metabolites (with the same units and scale) for the imputation. In contrast, we used clinical variables sex and age in addition to the auxiliary metabolites as predictors. Furthermore, the outcome, BMI, was added as well. Adding the outcome and the covariates is essential in MICE imputations to avoid bias and underestimation in the imputation results as shown in simulation studies [24] and discussed in several sources [17,23]. Adding the clinical variables and the outcome in our study was an additional step that was not used in the simulation study by Do et al. [6].

8 APPENDIX B

Table B1. Distribution of the mean correlation for the incomplete endogenous and unannotated metabolites in NEO.

Mean Correlation	0.1–0.19, *n* (%)	0.2–0.29, *n* (%)	0.3–0.39, *n* (%)	0.4–0.49, *n* (%)	0.5–0.59, *n* (%)	0.6–0.69, *n* (%)	0.7–0.79, *n* (%)	0.8–0.89, *n* (%)	Total
Endogenous	32 (8.79)	86 (23.63)	98 (26.92)	58 (15.93)	53 (14.56)	29 (7.97)	7 (1.92)	1 (0.27)	364
Unannotated	53 (23.35)	83 (36.56)	50 (22.03)	22 (9.69)	15 (6.61)	1 (0.44)	3 (1.32)	0 (0)	227
Combined	85 (14.38)	169 (28.6)	148 (25.04)	80 (13.54)	68 (11.51)	30 (5.08)	10 (1.69)	1 (0.17)	591

Approximately 80% of the metabolites have a mean correlation below 0.5 with their respective top 10 correlated metabolites in the correlation matrix.

Table B2. Percentage bias of the imputation methods across different parameters on different metabolites including MICE-pmm with a single imputation.

Missing Mechanism	Sample Size	Missing Percentage	Metabolites/Imputation Method											
			PC(32:2)			Succinylcarnitine			Glutamate			Urate		
			MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]
MCAR	*n* = 50	15%	2.0	2.1	1.4	0.3	6.0	0.4	2.2	0.2	2.3	4.9	4.9	5.5
		30%	3.2	4.5	3.4	2.8	14.4	2.6	8.6	1.1	8.8	11.3	10.2	10.1
		60%	13.9	14.5	13.2	28.9	31.3	29.7	40.2	2.0	40.7	39.9	27.6	40.9
	n = 150	15%	0.7	0.5	0.9	1.4	13.8	1.7	0.2	1.9	0.3	0.7	0.5	0.5
		30%	1.9	1.8	2.2	0.4	31.9	0.9	1.1	5.1	1.1	0.7	1.3	0.1
		60%	3.9	3.2	3.4	5.2	72.7	0.6	6.2	15.0	6.0	2.09	12.5	2.1
	n = 599	15%	0.3	0.8	0.3	0.1	9.3	0.2	0.2	2.6	0.3	0.6	2.2	0.4
		30%	0.7	2.4	0.8	0.6	20.1	0.3	0.1	5.6	0.1	1.5	4.9	1.8
		60%	1.9	7.9	1.7	3.5	50.9	4.6	0.7	15.1	0.8	2.6	15.0	2.4
PLoD	*n* = 50	15%	1.3	0.8	1.0	2.8	1.8	3.4	0.7	1.8	0.8	12.7	6.2	12.5
		30%	3.5	2.2	2.8	7.7	4.3	7.5	4.2	3.6	4.1	26.3	16.5	25.5
		60%	19.5	3.0	18.6	34.8	8.3	34.8	19.3	15.0	19.6	87.4	66.1	89.4
	n = 150	15%	0.2	1.6	0.2	0.7	12.8	0.7	1.3	0.8	1.4	5.3	6.3	5.4
		30%	0.3	3.3	0.6	0.2	23.4	0.4	3.0	2.3	3.1	8.0	14.9	8.7
		60%	4.2	2.6	4.2	16.2	46.1	17.2	9.4	8.6	9.1	39.8	59.7	38.2
	n = 599	15%	0.5	0.1	0.5	0.9	7.3	1.0	0.9	1.4	0.9	0.5	1.9	0.5
		30%	1.1	1.6	1.2	2.5	16.7	2.9	2.1	2.9	2.2	0.6	3.3	0.6
		60%	4.9	11.5	4.9	9.7	54.3	9.2	4.1	13.1	4.4	0.1	4.3	0.6

Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

Table B3. Mean and standard deviation of the estimates of the imputation methods using the MCAR missing mechanism, the three sample sizes, the three missing percentages, and four metabolites.

Sample Size	Missing Percentage	Metabolites/Imputation Method/Mean Estimate (SD)									
		PC(32:2)		Succinylcarnitine		Glutamate			Urate		
		MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm		kNN-obs-sel MICE-pmm			kNN-obs-sel
n = 50	0%	-4.18×10^{-7}		2.84×10^{-6}		1.83×10^{-7}			1.39×10^{-8}		
	15%	-4.46×10^{-7} (6.04×10^{-8})	-4.47×10^{-7} (7.24×10^{-8})	4.52×10^{-6} (6.38×10^{-7})	4.8×10^{-6} (7.38×10^{-7})	2.82×10^{-8} (1.29×10^{-9})	2.88×10^{-8} (1.69×10^{-9})	9.11×10^{-9} (3.29×10^{-9})	1.01×10^{-8} (3.27×10^{-9})		
	30%	-4.52×10^{-7} (1.04×10^{-7})	-4.57×10^{-7} (1.21×10^{-7})	4.39×10^{-6} (1.02×10^{-6})	5.15×10^{-6} (1.24×10^{-6})	2.64×10^{-8} (2.61×10^{-9})	2.86×10^{-8} (2.79×10^{-9})	8.5×10^{-9} (5.66×10^{-9})	1.06×10^{-8} (5.72×10^{-9})		
	60%	-3.77×10^{-7} (2.28×10^{-7})	-5.01×10^{-7} (2.34×10^{-7})	3.22×10^{-6} (1.73×10^{-6})	5.96×10^{-6} (2.58×10^{-6})	1.73×10^{-8} (6.06×10^{-9})	2.83×10^{-8} (5.57×10^{-9})	5.75×10^{-9} (9.36×10^{-9})	1.22×10^{-8} (1.13×10^{-8})		
n = 150	0%	-3.64×10^{-7}		1.53×10^{-6}		2.89×10^{-8}			9.58×10^{-9}		
	15%	-3.66×10^{-7} (1.57×10^{-8})	-3.62×10^{-7} (1.53×10^{-8})	1.54×10^{-6} (4.09×10^{-7})	1.77×10^{-6} (4.17×10^{-7})	1.66×10^{-8} (8.24×10^{-10})	1.69×10^{-8} (8.15×10^{-10})	9.76×10^{-9} (1.62×10^{-9})	9.74×10^{-9} (1.46×10^{-9})		
	30%	-3.71×10^{-7} (2.57×10^{-8})	-3.58×10^{-7} (2.29×10^{-8})	1.52×10^{-6} (6.43×10^{-7})	1.99×10^{-6} (6.94×10^{-7})	1.64×10^{-8} (1.37×10^{-9})	1.74×10^{-8} (1.21×10^{-9})	9.75×10^{-9} (2.45×10^{-9})	9.82×10^{-9} (2.29×10^{-9})		
	60%	-3.78×10^{-7} (4.87×10^{-8})	-3.52×10^{-7} (4.3×10^{-8})	1.48×10^{-6} (1.2×10^{-6})	2.7×10^{-6} (1.48×10^{-6})	1.56×10^{-8} (2.79×10^{-9})	1.91×10^{-8} (2.47×10^{-9})	9.89×10^{-9} (4.61×10^{-9})	1.09×10^{-8} (4.81×10^{-9})		
n = 599	0%	-4.38×10^{-7}		4.53×10^{-6}		1.66×10^{-8}			9.69×10^{-9}		
	15%	-4.2×10^{-7} (8.44×10^{-9})	-4.22×10^{-7} (1.05×10^{-8})	2.84×10^{-6} (1.96×10^{-7})	3.1×10^{-6} (1.94×10^{-7})	1.84×10^{-8} (4.43×10^{-10})	1.88×10^{-8} (4.77×10^{-10})	1.4×10^{-8} (8.08×10^{-10})	1.42×10^{-8} (7.59×10^{-10})		
	30%	-4.21×10^{-7} (1.33×10^{-8})	-4.28×10^{-7} (1.76×10^{-8})	2.85×10^{-6} (3.08×10^{-7})	3.43×10^{-6} (3.35×10^{-7})	1.84×10^{-8} (7.05×10^{-10})	1.94×10^{-8} (7.16×10^{-10})	1.41×10^{-8} (1.27×10^{-9})	1.46×10^{-8} (1.19×10^{-9})		
	60%	-4.26×10^{-7} (2.52×10^{-8})	-4.51×10^{-7} (3.54×10^{-8})	2.79×10^{-6} (5.88×10^{-7})	4.34×10^{-6} (7.2×10^{-7})	1.82×10^{-8} (1.34×10^{-9})	2.11×10^{-8} (1.27×10^{-9})	1.43×10^{-8} (2.47×10^{-9})	1.6×10^{-8} (2.55×10^{-9})		

Estimates are the regression coefficient from the model BMI ~ age + sex + metabolite. Therefore, the estimates are the mean increase in BMI per 1 unit increase of the metabolite. The 0% rows are the estimates from the real data before amputing and imputing the missing values.

Table B4. Mean and standard deviation of the estimates of the imputation methods using the PLoD missing mechanism, the three sample sizes, the three missing percentages, and four metabolites.

Sample Size	Missing Percentage	Metabolites/Imputation Method/Mean Estimate (SD)							
		PC(32:2)	kNN-obs-sel	Succinylcarnitine	kNN-obs-sel	Glutamate	kNN-obs-sel	Urate	kNN-obs-sel
		MICE-pmm		MICE-pmm		MICE-pmm		MICE-pmm	
n = 50	0%	-4.18×10^{-7}		2.84×10^{-6}		1.83×10^{-7}		1.39×10^{-8}	
	15%	-4.32×10^{-7} (3.91×10^{-8})	-4.34×10^{-7} (3.35×10^{-8})	4.4×10^{-6} (2.56×10^{-7})	4.62×10^{-6} (1.79×10^{-7})	2.87×10^{-8} (1.09×10^{-9})	2.94×10^{-8} (1.29×10^{-9})	8.36×10^{-9} (2.53×10^{-9})	8.99×10^{-9} (2.19×10^{-9})
	30%	-4.22×10^{-7} (7.47×10^{-8})	-4.28×10^{-7} (5.19×10^{-8})	4.15×10^{-6} (4.4×10^{-7})	4.72×10^{-6} (2.83×10^{-7})	2.77×10^{-8} (2.07×10^{-9})	2.99×10^{-8} (1.99×10^{-9})	7.07×10^{-9} (3.7×10^{-9})	8×10^{-9} (3.14×10^{-9})
	60%	-3.52×10^{-7} (1.82×10^{-7})	-4.51×10^{-7} (1.32×10^{-7})	3.05×10^{-6} (1.18×10^{-6})	4.96×10^{-6} (1.19×10^{-6})	2.33×10^{-8} (4.89×10^{-9})	3.32×10^{-8} (3.59×10^{-9})	1.21×10^{-9} (7.1×10^{-9})	3.25×10^{-9} (7.47×10^{-9})
n = 150	0%	-3.64×10^{-7}		1.53×10^{-6}		2.89×10^{-8}		9.58×10^{-9}	
	15%	-3.65×10^{-7} (1.26×10^{-8})	-3.58×10^{-7} (1.22×10^{-8})	1.51×10^{-6} (3.88×10^{-7})	1.72×10^{-6} (3.61×10^{-7})	1.64×10^{-8} (6.63×10^{-10})	1.67×10^{-8} (5.85×10^{-10})	9.18×10^{-9} (1.62×10^{-9})	9.08×10^{-9} (1.31×10^{-9})
	30%	-3.65×10^{-7} (1.96×10^{-8})	-3.52×10^{-7} (1.86×10^{-8})	1.49×10^{-6} (6.11×10^{-7})	1.86×10^{-6} (5.51×10^{-7})	1.61×10^{-8} (1.07×10^{-9})	1.7×10^{-8} (8.47×10^{-10})	8.91×10^{-9} (2.14×10^{-9})	8.25×10^{-9} (1.87×10^{-9})
	60%	-3.79×10^{-7} (3.83×10^{-8})	-3.74×10^{-7} (3.83×10^{-8})	1.66×10^{-6} (1.05×10^{-6})	2.11×10^{-6} (9.8×10^{-7})	1.5×10^{-8} (1.8×10^{-9})	1.8×10^{-8} (1.51×10^{-9})	5.83×10^{-9}(3.36×10^{-9})	3.91×10^{-9} (3.08×10^{-9})
n = 599	0%	-4.38×10^{-7}		4.53×10^{-6}		1.66×10^{-8}		9.69×10^{-9}	
	15%	-4.2×10^{-7} (7.17×10^{-9})	-4.18×10^{-7} (8.61×10^{-9})	2.82×10^{-6} (1.59×10^{-7})	3.05×10^{-6} (1.5×10^{-7})	1.82×10^{-8} (3.37×10^{-10})	1.86×10^{-8} (3.1×10^{-10})	1.4×10^{-8} (6.57×10^{-10})	1.42×10^{-8} (5.63×10^{-10})
	30%	-4.23×10^{-7} (1.1×10^{-8})	-4.25×10^{-7} (1.52×10^{-8})	2.79×10^{-6} (2.52×10^{-7})	3.34×10^{-6} (2.34×10^{-7})	1.8×10^{-8} (5.37×10^{-10})	1.89×10^{-8} (4.81×10^{-10})	1.4×10^{-8} (9.58×10^{-10})	1.44×10^{-8} (8.97×10^{-10})
	60%	-4.39×10^{-7} (2.02×10^{-8})	-4.66×10^{-7} (3.56×10^{-8})	2.58×10^{-6} (5.62×10^{-7})	4.37×10^{-6} (6.13×10^{-7})	1.76×10^{-8} (9.99×10^{-10})	2.07×10^{-8} (8.04×10^{-10})	1.39×10^{-8} (1.64×10^{-9})	1.45×10^{-8} (1.57×10^{-9})

Estimates are the regression coefficient from the model BMI ~ age + sex + metabolite. Therefore, the estimates are the mean increase in BMI per 1 unit increase of the metabolite. The 0% rows are the estimates from the real data before amputing and imputing the missing values.

Table B5. RMSE of the imputation methods across different parameters on different metabolites.

Missing Mechanism	Sample Size	Missing Percentage	Metabolites/Imputation Method							
			PC(32:2)		Succinylcarnitine		Glutamate		Urate	
			MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel	MICE-pmm	kNN-obs-sel
MCAR	*n* = 50	15%	6.56×10^{-8}	7.48×10^{-8}	6.5×10^{-7}	7.73×10^{-7}	3.32×10^{-9}	3.3×10^{-9}	1.44×10^{-9}	1.69×10^{-9}
		30%	1.04×10^{-7}	1.22×10^{-7}	1.02×10^{-6}	1.44×10^{-6}	5.76×10^{-9}	5.8×10^{-9}	3.59×10^{-9}	2.8×10^{-9}
		60%	2.33×10^{-7}	2.49×10^{-7}	2.22×10^{-6}	2.95×10^{-6}	1.01×10^{-8}	1.16×10^{-8}	1.31×10^{-8}	5.59×10^{-9}
	n = 150	15%	1.71×10^{-8}	1.69×10^{-8}	3.97×10^{-7}	4.6×10^{-7}	1.62×10^{-9}	1.46×10^{-9}	8.24×10^{-10}	8.75×10^{-10}
		30%	2.63×10^{-8}	2.49×10^{-8}	6.48×10^{-7}	8.5×10^{-7}	2.45×10^{-9}	2.29×10^{-9}	1.38×10^{-9}	1.47×10^{-9}
		60%	5.23×10^{-8}	4.78×10^{-8}	1.27×10^{-6}	1.84×10^{-6}	4.62×10^{-9}	4.96×10^{-9}	2.97×10^{-9}	3.51×10^{-9}
	n = 599	15%	8.64×10^{-9}	1.08×10^{-8}	$2{,}00 \times 10^{-7}$	3.35×10^{-7}	8.12×10^{-10}	8.17×10^{-10}	4.45×10^{-10}	6.71×10^{-10}
		30%	1.28×10^{-8}	1.95×10^{-8}	3.08×10^{-7}	6.6×10^{-7}	1.28×10^{-9}	1.36×10^{-9}	7.05×10^{-10}	1.24×10^{-9}
		60%	2.63×10^{-8}	5.04×10^{-8}	5.91×10^{-7}	1.61×10^{-6}	2.5×10^{-9}	3.29×10^{-9}	1.35×10^{-9}	3.04×10^{-9}
PLoD	*n* = 50	15%	3.99×10^{-8}	3.23×10^{-8}	2.98×10^{-7}	$2{,}00 \times 10^{-7}$	2.81×10^{-9}	2.26×10^{-9}	1.11×10^{-9}	1.38×10^{-9}
		30%	7.36×10^{-8}	5.36×10^{-8}	5.66×10^{-7}	3.48×10^{-7}	4.47×10^{-9}	3.51×10^{-9}	2.4×10^{-9}	2.24×10^{-9}
		60%	1.87×10^{-7}	1.25×10^{-7}	$2{,}00 \times 10^{-6}$	1.27×10^{-6}	1.1×10^{-8}	9.79×10^{-9}	7.42×10^{-9}	5.62×10^{-9}
	n = 150	15%	1.33×10^{-8}	1.36×10^{-8}	3.62×10^{-7}	3.8×10^{-7}	1.7×10^{-9}	1.45×10^{-9}	6.97×10^{-10}	5.99×10^{-10}
		30%	1.96×10^{-8}	2.15×10^{-8}	5.97×10^{-7}	6.46×10^{-7}	2.27×10^{-9}	2.36×10^{-9}	1.18×10^{-9}	9.27×10^{-10}
		60%	4.42×10^{-8}	3.98×10^{-8}	1.07×10^{-6}	1.23×10^{-6}	5.11×10^{-9}	6.55×10^{-9}	2.39×10^{-9}	2.07×10^{-9}
	n = 599	15%	7.54×10^{-9}	8.52×10^{-9}	1.67×10^{-7}	2.56×10^{-7}	6.6×10^{-10}	6.22×10^{-10}	3.77×10^{-10}	4.03×10^{-10}
		30%	1.17×10^{-8}	1.64×10^{-8}	2.64×10^{-7}	5.32×10^{-7}	9.61×10^{-10}	1.01×10^{-9}	6.55×10^{-10}	7.2×10^{-10}
		60%	2.96×10^{-8}	5.82×10^{-8}	6.4×10^{-7}	1.65×10^{-6}	1.64×10^{-9}	1.68×10^{-9}	1.25×10^{-9}	2.53×10^{-9}

Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

Table B6. Percentage RMSE of the imputation methods across different parameters on different metabolites including MICE-pmm with a single imputation.

Missing Mechanism	Sample Size	Missing Percentage	Metabolites/Imputation Method											
			PC(32:2)			Succinylcarnitine			Glutamate			Urate		
			MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]	MICE-pmm	kNN-obs-sel	MICE-pmm [m = 1]
MCAR	*n* = 50	15%	14.6	17.4	18.3	22.9	27.2	29.0	7.9	9.2	10.9	23.9	23.7	29.7
		30%	25.1	29.4	31.3	35.9	50.7	45.1	19.6	15.3	24.6	41.4	41.7	50.6
		60%	56.2	57.9	71.2	78.2	103.9	96.1	71.4	30.5	78.0	72.5	83.3	89.1
	n = 150	15%	3.8	3.7	5.0	14.0	16.2	17.1	4.5	4.8	5.9	11.6	10.5	14.1
		30%	6.4	5.7	8.2	22.8	29.9	27.7	7.5	8.0	9.8	17.6	16.5	21.1
		60%	12.1	10.7	15.6	44.7	64.8	51.8	16.2	19.1	20.0	33.2	35.6	39.3
	n = 599	15%	2.0	2.6	2.7	7.0	11.8	8.9	2.4	3.7	3.2	5.8	5.9	7.0
		30%	3.3	4.8	4.0	10.8	23.2	13.3	3.8	6.8	4.7	9.2	9.8	11.6
		60%	6.3	11.6	8.0	20.8	56.7	27.1	7.4	16.6	8.8	18.0	23.6	21.4
PLoD	*n* = 50	15%	9.4	8.1	12.5	10.5	7.0	15.5	6.1	7.5	8.3	20.2	16.2	23.9
		30%	18.2	12.6	23.3	19.9	12.3	25.5	13.1	12.2	17.0	32.1	25.2	37.2
		60%	48.1	31.8	61.0	70.4	44.7	81.0	40.4	30.6	48.9	79.0	70.3	90.5
	n = 150	15%	3.0	3.2	4.1	12.7	13.4	15.7	3.8	3.3	4.9	12.2	10.4	13.7
		30%	4.7	5.3	6.2	21.0	22.7	25.2	6.4	5.1	7.9	16.3	17.0	19.2
		60%	9.9	9.4	13.4	37.7	43.3	49.3	13.0	11.3	16.7	36.7	47.1	42.5
	n = 599	15%	1.8	2.1	2.4	5.9	9.0	7.7	2.1	2.2	2.6	4.7	4.5	6.0
		30%	2.8	3.9	3.7	9.3	18.7	12.6	3.6	3.9	4.4	6.9	7.3	8.6
		60%	6.9	14.3	8.3	22.5	58.1	27.4	6.8	13.8	8.1	11.8	12.1	15.2

We converted the RMSE values to a percentage by subtracting then dividing the RMSE values by the corresponding true estimates (in sample size *n* = 599). Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

Table B7. Runtime for MICE-pmm and kNN-obs-sel imputations using different datasets.

Imputation Method	Dataset Sizes		
	***n* = 599**	***n* = 5400**	***n* = 20,000**
MICE-pmm (minutes)	1.9	13.4	138.9
kNN-obs-sel (minutes)	0.7	16.2	210.7

Imputation was applied to the actual NEO dataset (*n* = 599; 58% metabolites contain missing values) and two oversampled datasets generated from the NEO data. With *n* = 599, kNN-obs-sel was slightly faster. However, MICE-pmm imputation took a shorter time to complete the imputations in larger datasets.

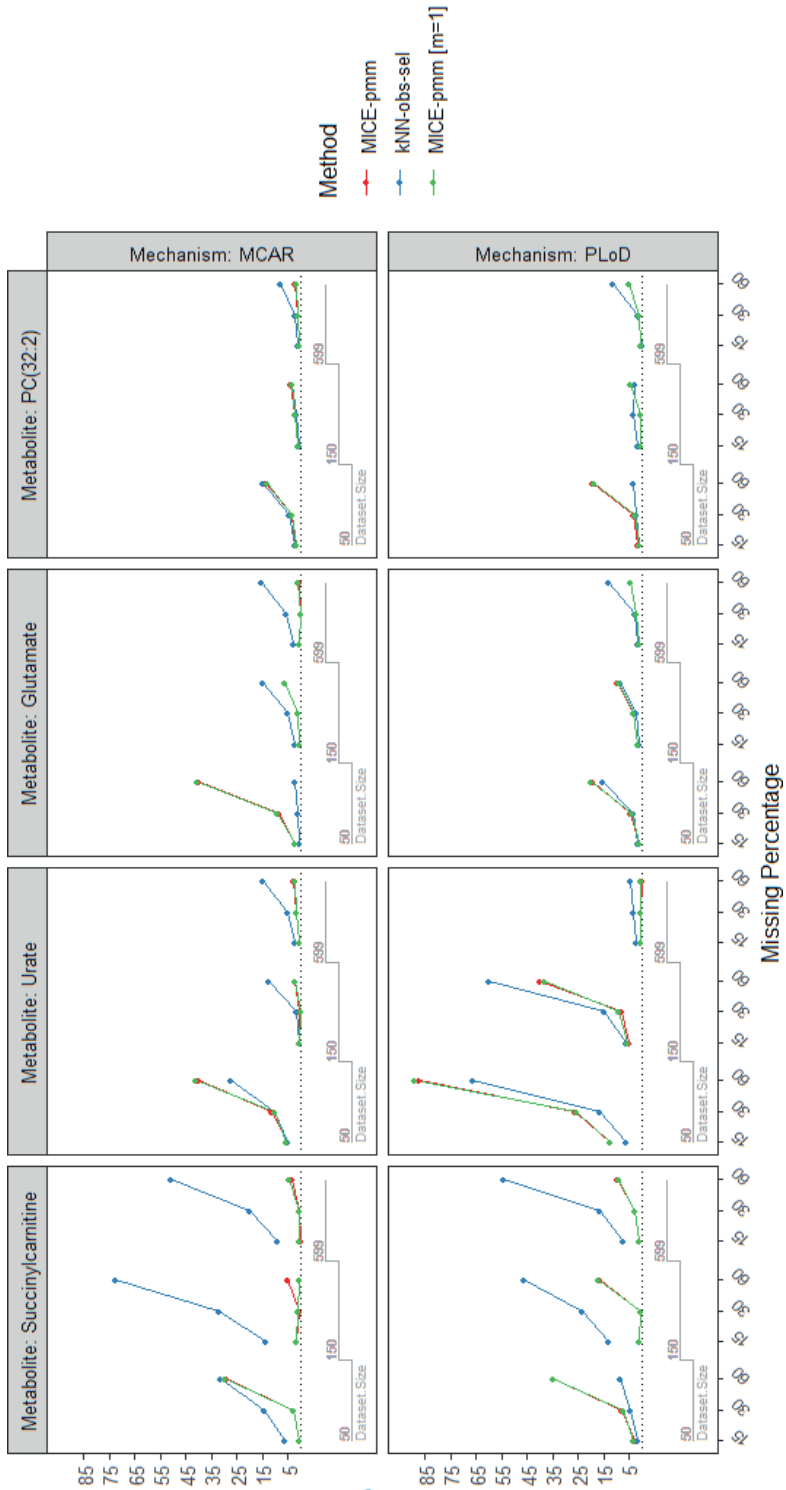


Figure B1. Nested loop plot of the percentage bias of the four metabolites from the simulation including MICE-pmm with a single imputation. The X axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

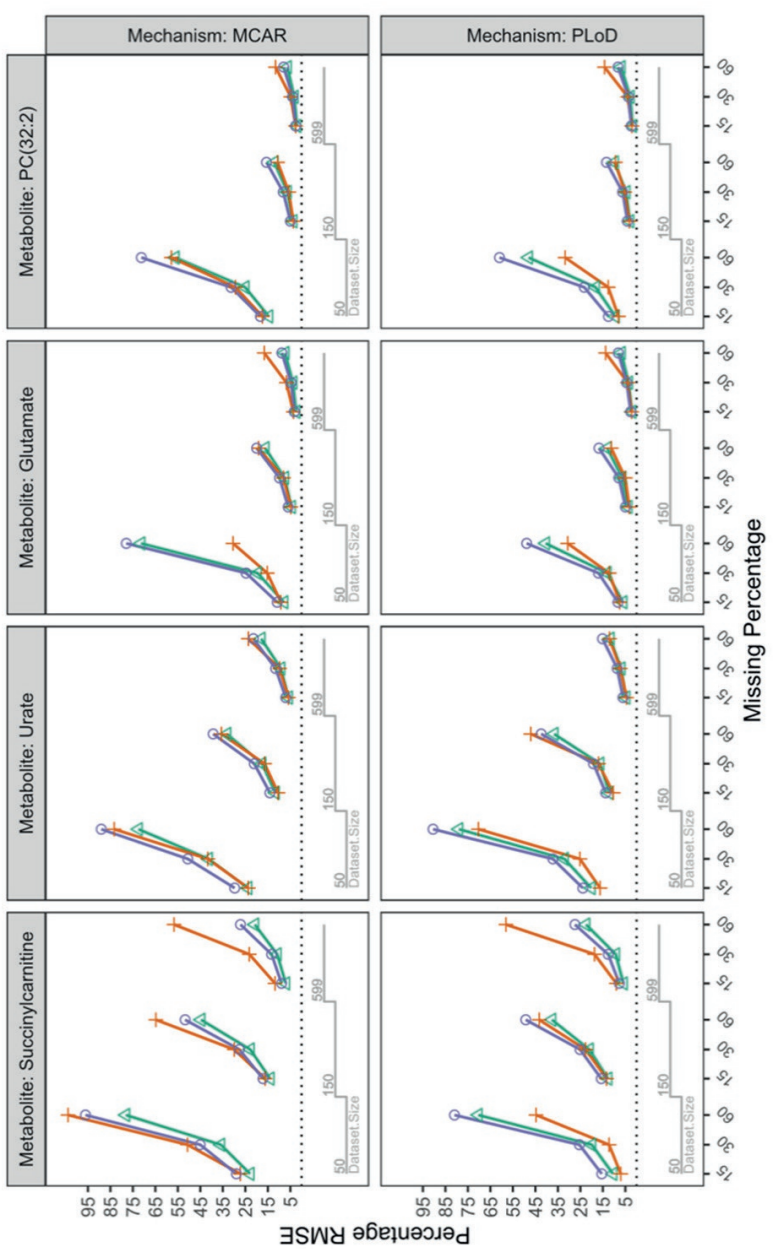


Figure B2. Nested loop plot of the percentage RMSE of the four metabolites from the simulation including MICE-pmm with a single imputation. To simplify comparability in the plot we converted the RMSE values to a percentage by subtracting then dividing the RMSE values by the corresponding true estimates (in sample size *n* = 599). The horizonal axis in each box represents the missing percentage and is split per sample size. Abbreviations: MCAR: missing completely at random; PLoD: probabilistic limit of detection; MICE-pmm [m = 1]: MICE-pmm with a single imputation.

9 REFERENCES

1. Suhre, K.; Meisinger, C.; Döring, A.; Altmaier, E.; Belcredi, P.; Gieger, C.; Chang, D.; Milburn, M.V.; Gall, W.E.; Weinberger, K.M.; et al. Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. *PLoS ONE* **2010**, *5*, e13953, doi:10.1371/journal.pone.0013953.
2. Schrimpe-Rutledge, A.C.; Codreanu, S.G.; Sherrod, S.D.; McLean, J.A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1897–1905, doi:10.1007/s13361-016-1469-y.
3. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 1–10, doi:10.1038/s41598-017-19120-0.
4. Karpievitch, Y.V.; Dabney, A.R.; Smith, R.D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinform.* **2012**, *13*, S5, doi:10.1186/1471-2105-13-s16-s5.
5. Hrydziuszko, O.; Viant, M.R. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. *Metabolomics* **2012**, *8*, 161–174, doi:10.1007/s11306-011-0366-4.
6. Do, K.T.; Wahl, S.; Raffler, J.; Molnos, S.; Laimighofer, M.; Adamski, J.; Suhre, K.; Strauch, K.; Peters, A.; Gieger, C.; et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **2018**, *14*, 128.
7. Alonso, A.; Marsal, S.; Juliã, A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3*, 23, doi:10.3389/fbioe.2015.00023.
8. Deng, Y.; Chang, C.; Ido, M.S.; Long, Q. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Sci. Rep.* **2016**, *6*, 21689, doi:10.1038/srep21689.
9. Gromski, P.S.; Xu, Y.; Kotze, H.L.; Correa, E.; Ellis, D.I.; Armitage, E.G.; Turner, M.L.; Goodacre, R. Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. *Metabolites* **2014**, *4*, 433–452, doi:10.3390/metabo4020433.
10. Van Buuren, S. *Flexible Imputation of Missing Data*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
11. Little, R.J.A. Missing-Data Adjustments in Large Surveys. *J. Bus. Econ. Stat.* **1988,** *6*, 287–296.
12. Rubin, D.B. Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *J. Bus. Econ. Stat.* **1986**, *4*, 87–94.
13. Faquih, T. *Imputation of Untargeted Metabolites Official Release*, Version v1.3; Zenodo: Meyrin, Switzerland, 2020. Available online: https://zenodo.org/record/4167193 (accessed on 31 October 2020).
14. Rücker, G.; Schwarzer, G. Presenting simulation results in a nested loop plot. *BMC Med. Res. Methodol.* **2014**, *14*, 129, doi:10.1186/1471-2288-14-129.
15. Shah, J.; Rai, S.N.; DeFilippis, A.P.; Hill, B.G.; Bhatnagar, A.; Brock, G. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinform.* **2017**, *18*, 114, doi:10.1186/s12859-017-1547-6.
16. Di Guida, R.; Engel, J.; Allwood, J.W.; Weber, R.J.M.; Jones, M.R.; Sommer, U.; Viant, M.R.; Dunn, W.B. Non-targeted UHPLC-MS metabolomic data processing methods: A comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **2016**, *12*, 1–14, doi:10.1007/s11306-016-1030-9.
17. Molenberghs, G.; Kenward, M. *Missing Data in Clinical Studies*; Wiley: Hoboken, NJ, USA, 2007.
18. Wang, J.; Li, Z.F.; Chen, J.; Zhao, H.; Luo, L.; Chen, C.; Xu, X.; Zhang, W.; Gao, K.; Li, B.; et al. Metabolomic identification of diagnostic plasma biomarkers in humans with chronic heart failure. *Mol. BioSyst.* **2013**, *9*, 2618, doi:10.1039/c3mb70227h.
19. Yousri, N.A.; Bayoumy, K.; Elhaq, W.G.; Mohney, R.P.; Al Emadi, S.; Hammoudeh, M.; Halabi, H.; Masri, B.; Badsha, H.; Uthman, I.; et al. Large Scale Metabolic Profiling identifies Novel Steroids linked to Rheumatoid Arthritis. *Sci. Rep.* **2017**, *7*, 1–9, doi:10.1038/s41598-017-05439-1.

20. De Mutsert, R.; Heijer, M.D.; Rabelink, T.J.; Smit, J.W.A.; Romijn, J.A.; Jukema, J.W.; De Roos, A.; Cobbaert, C.M.; Kloppenburg, M.; Le Cessie, S.; et al. The Netherlands Epidemiology of Obesity (NEO) study: Study design and data collection. *Eur. J. Epidemiol.* **2013**, *28*, 513–523, doi:10.1007/s10654-013-9801-3.
21. Evans, A.; Bridgewater, B.; Liu, Q.; Mitchell, M.; Robinson, R.; Dai, H.; Stewart, S.; DeHaven, C.; Miller, L.J.M. High Resolution Mass Spectrometry Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry in High-Throughput Profiling Metabolomics. *J. Postgenomics Drug Biomark. Dev.* **2014**, *4*, 1–7, doi:10.4172/2153-0769.1000132.
22. Rhee, E.P.; Waikar, S.S.; Rebholz, C.M.; Zheng, Z.; Perichon, R.; Clish, C.B.; Evans, A.M.; Avila, J.; Denburg, M.R.; Anderson, A.H.; et al. Variability of Two Metabolomic Platforms in CKD. *Clin. J. Am. Soc. Nephrol.* **2019**, *14*, 40.
23. White, I.R.; Royston, P.; Wood, A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **2011**, *30*, 377–399, doi:10.1002/sim.4067.
24. Moons, K.G.; Donders, R.A.; Stijnen, T.; Harrell, F.E. Using the outcome for imputation of missing predictor values was preferred. *J. Clin. Epidemiol.* **2006**, *59*, 1092–1101, doi:10.1016/j.jclinepi.2006.01.009.
25. Kowarik, A.; Templ, M. Imputation with the R Package VIM. *J. Stat. Softw.* **2016,** *74*, 16.
26. Cirulli, E.T.; Guo, L.; Swisher, C.L.; Shah, N.; Huang, L.; Napier, L.A.; Kirkness, E.F.; Spector, T.D.; Caskey, C.T.; Thorens, B.; et al. Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. *Cell Metab.* **2019**, *29*, 488–500.e2, doi:10.1016/j.cmet.2018.09.022.
27. Demirtas, H.; Freels, S.A.; Yucel, R.M. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *J. Stat. Comput. Simul.* **2008**, *78*, 69–84, doi:10.1080/10629360600903866.
28. Morris, T.P.; White, I.R.; Crowther, M.J. Using simulation studies to evaluate statistical methods. *Stat. Med.* **2019**, *38*, 2074–2102, doi:10.1002/sim.8086.
29. Van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67, doi:10.18637/jss.v045.i03.
30. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons, Inc.: New York, NY, USA, 1987.
31. Rubin, D.B. Multiple Imputation After 18+ Years. *J. Am. Stat. Assoc.* **1996**, *91*, 473–489.