Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies

Kieu Trinh Do[1¶], Simone Wahl[2,3,4¶], Johannes Raffler[5], Sophie Molnos[2,3,4], Michael Laimighofer[1], Jerzy Adamski[6,7], Karsten Suhre[9], Konstantin Strauch[10,11], Annette Peters[2,3], Christian Gieger[2,3], Claudia Langenberg[12], Isobel D. Stewart [12], Fabian J. Theis[1,13], Harald Grallert[2,3,4], Gabi Kastenmüller[4,5#], Jan Krumsiek[1,4,14#]

1 Institute of Computational Biology, Helmholtz-Zentrum München, Neuherberg, Germany, **2** Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, **3** Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, **4** German Center for Diabetes Research (DZD e.V.), Neuherberg, Germany, **5** Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, Neuherberg, Germany, **6** Institute of Experimental Genetics, Genome Analysis Center Helmholtz Zentrum München, Neuherberg, Germany, **7** Lehrstuhl für Experimentelle Genetik, Technische Universität München, Freising-Weihenstephan, Germany, **8** German Center for Cardiovascular Disease Research (DZHK e.V.), partner-site Munich, Germany, **9** Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Doha, Qatar, **10** Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany, **11** Chair of Genetic Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich, Germany, **12** MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, **13** Department of Mathematics, Technische Universität München, Garching, Germany **14** Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Department of Physiology and Biophysics, Weill Cornell Medicine, New York, USA

[¶] These authors contributed equally to this work.

[#] **Corresponding authors**:

Dr. Gabi Kastenmüller, Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, Neuherberg, Germany, Phone: +49 89 3187-3578, Fax: +49 89 3187-3585, E-mail: g.kastenmueller@helmholtz-muenchen.de

Dr. Jan Krumsiek, Institute of Computational Biology, Helmholtz-Zentrum München, Neuherberg, Germany, Phone: +49 89 3187-3641, Fax: +49 89 3187-3369, E-mail: jan.krumsiek@helmholtz-muenchen.de

Abstract

BACKGROUND: Untargeted mass spectrometry (MS)-based metabolomics data often contain missing values that reduce statistical power and can introduce bias in epidemiological studies. However, a systematic assessment of the various sources of missing values and strategies to handle these data has received little attention. Missing data can occur systematically, e.g. from run day-dependent effects due to limits of detection (LOD); or it can be random as, for instance, a consequence of sample preparation.

METHODS: We investigated patterns of missing data in an MS-based metabolomics experiment of serum samples from the German KORA F4 cohort (n = 1750). We then evaluated 31 imputation methods in a simulation framework and biologically validated the results by applying all imputation approaches to real metabolomics data. We examined the ability of each method to reconstruct biochemical pathways from data-driven correlation networks, and the ability of the method to increase statistical power while preserving the strength of established genetically metabolic quantitative trait loci.

RESULTS: Run day-dependent LOD-based missing data accounts for most missing values in the metabolomics dataset. Although multiple imputation by chained equations (*MICE*) performed well in many scenarios, it is computationally and statistically challenging. K-nearest neighbors (*KNN*) imputation on observations with variable pre-selection showed robust performance across all evaluation schemes and is computationally more tractable.

CONCLUSION: Missing data in untargeted MS-based metabolomics data occur for various reasons. Based on our results, we recommend that *KNN*-based imputation is performed on observations with variable pre-selection since it showed robust results in all evaluation schemes.

Keywords: untargeted metabolomics, missing values imputation, limit of detection, batch effects, runday effects, *MICE*, K-nearest neighbor, mass spectrometry

Key messages

- Untargeted MS-based metabolomics data show missing values due to both batch-specific LOD-based and non-LOD-based effects.
- Statistical evaluation of multiple imputation methods was conducted on both simulated and real datasets.
- Biological evaluation on real data assessed the ability of imputation methods to preserve statistical inference of biochemical pathways and correctly estimate effects of genetic variants on metabolite levels.
- *KNN*-based imputation on observations with variable pre-selection and $K = 10$ showed robust performance for all data scenarios across all evaluation schemes.

Introduction

In epidemiological studies, metabolomics is an established tool that provides insights into disease mechanisms (1), as metabolite profiles generate a molecular readout that is closely linked to the (patho-)phenotype (2,3). Recent metabolomics studies have identified many metabolites as candidate biomarkers for various health conditions, such as diabetes (4–6) and cardiovascular diseases (7,8). Mass spectrometry (MS)-based metabolomics measurements can be performed either in a targeted or untargeted manner (9). In the former, only a limited number of already known and biochemically annotated metabolites are captured. In the latter, the measurements are not limited to predefined signals and offer discovery of novel compounds. While missing values in targeted MS-based data occur rarely, untargeted MS-based techniques typically produce 20-30% missing values, affecting more than 80% of the measured compounds (10–13).

There are various reasons why metabolite concentrations can be missing in an untargeted metabolomics dataset. First, it is possible that the molecules are truly absent from the sample, a situation that may occur e.g. for drug metabolites that only appear in a subset of people taking that medication. On the other hand, there are several technical reasons that could result in missing values, including: (i) instrument sensitivity thresholds, below which concentrations of a specific metabolite might not be detectable in a sample (i.e., below the limit of detection, LOD); (ii) matrix effects that impede the quantification of a metabolite in a sample through other co-eluting compounds and ion suppression; (iii) declining separation ability of the chromatographic column and increasing contamination of the MS instrument; and (iv) limitations in computational processing of spectra, such as poor selection and alignment of the spectral peaks across samples (14).

Commonly, observed patterns of missing data are categorized as either missing completely at random (MCAR), missing at random (MAR), or missing but not at random (MNAR) (15). In the MCAR category, the probability of missing values does not depend on observed or unobserved measurements. In contrast, the occurrence of MAR depends on other observed measurements (for

94 instance, resulting from technical effects, such as overlapping peaks). MNAR describes the
95 occurrence of missing values that depend on unobserved measurements (for instance, due to issues
96 with the performance of the machine).

97 Although it is clear that the handling of missing values affects all downstream analyses, it is
98 less clear how to appropriately handle their occurrence statistically. A simple *ad hoc* approach is
99 known as complete case analysis (*CCA*), which only considers samples that do not contain any missing
100 values in the metabolites analyzed in each statistical analysis step. However, missing data may occur
101 in some systematic way (i.e., they are dependent on external factors). For example, if all cases in a
102 case-control study have more missing data than the controls, removing observations that are missing
103 will lead to bias in biological interpretation (16). Furthermore, *CCA* can cause severe loss of
104 information and statistical power by excluding a majority of observations if multivariate methods,
105 such as principal component analysis or partial correlation networks, are to be performed.

106 A widely used and flexible class of missing data strategies is imputation, which involves the
107 replacement of missing values by reasonable substitute values. The most commonly used imputation
108 approaches for metabolomics data assume that missing data occur because they are below the limit
109 of detection (left-censoring, a variant of MNAR). Therefore, all missing entries of a metabolite are
110 replaced by a low constant value, such as the actual LOD (if known), zero, or the smallest value found
111 in the dataset for that metabolite (13). Another LOD-based substitution strategy assumes a
112 parametric left-truncated normal distribution and performs likelihood-based parameter estimation
113 on the observed values to reconstruct the truncated part of the distribution. Missing values are then
114 replaced by numbers drawn from this estimated part (16,17). Additional imputation-based
115 substitution approaches assume MCAR and replace missing values by the mean or median per
116 metabolite (12). Advanced approaches use multivariate statistical methods for imputation, including
117 multiple imputation by chained equations (MICE) (18) and K-nearest neighbors (KNN) imputation
118 (19,20).

119 Several previous studies have investigated the occurrence and effects of different strategies
120 for missing values in metabolomics data. Taylor *et al.* (21) reported that no single imputation method
121 was universally superior, but constant substitution methods consistently showed poor performance.
122 Gromski *et al.* (12) recommended imputation by Random Forests (RFs) for GC/MS metabolomics data
123 after evaluating the outputs of supervised and unsupervised learning approaches. Di Guida *et al.* (15)
124 investigated various combinations of different preprocessing steps to determine which were the
125 most appropriate for univariate and multivariate analyses of UHPLC-MS metabolomics data. The
126 authors recommended RF and *KNN*-based imputation for PCA and PLS-DA, respectively (15).
127 Armitage *et al.* (10) studied missing values in CE/MS metabolomics data and reported *KNN*
128 imputation to be more effective compared with simpler substitution-based imputation methods.
129 Finally, in a study by Hrydziuszko and Viant (11), a *KNN*-based imputation approach also
130 outperformed competing strategies in an investigation of direct infusion Fourier transform ion
131 cyclotron resonance (DI-FTICR) MS-based metabolomics data.

132 Despite these advances in our understanding of the effects of imputation on metabolomics
133 data analysis, several aspects have not been addressed by those previous studies. (i) A detailed
134 statistical description of the patterns of missing values in MS-based metabolomics data has not yet
135 been published. Most previous studies evaluated imputation strategies assuming only random or
136 LOD-based missing values without assessing whether this applies to real metabolomics datasets. In
137 particular, the influence of batch effects on the occurrence of missing values has not been
138 investigated in any study. If a cohort comprises a large number of samples, the MS runs usually are
139 spread across multiple days, which is known to influence metabolite measurements due to variation
140 in instrument sensitivity. Here, the LOD itself is also expected to vary across run days, an assumption
141 that has not been explicitly accounted for in any studies. (ii) In addition, a simulation framework that
142 reflects realistic data situations is needed to provide an unbiased evaluation of strategies for handling
143 missing values. Evaluation of previous studies has been biased in the sense that "complete"
144 measured data (created by excluding all variables with missing values) with artificially introduced

145 missing values were simulated, which most likely does not mirror realistic missing value patterns. (iii)
146 Finally, biological validation and biochemical interpretation of the data have not been addressed in
147 the majority of papers. Only Hrydziuszko *et al.* evaluated the ability of different imputation strategies
148 to preserve metabolic differences between biological groups, which then were related to KEGG
149 pathways (11).

150 In the present study, we analyzed patterns of missing data and evaluated the performance of
151 various imputation strategies for untargeted MS-based metabolomics data from serum samples of
152 the German Cooperative Health Research in the Region of Augsburg (KORA) F4 cohort. Data were
153 measured on a typical, widely used untargeted MS-based metabolomics platform (Metabolon, Inc.,
154 USA) and should be representative of many untargeted population-scale metabolomics studies. The
155 study consisted of three steps: (i) We described and analyzed patterns of missing values and their
156 possible underlying mechanisms in a real untargeted metabolomics dataset. In particular, we
157 investigated the occurrence of missing values within and across batches of measurements. (ii) The
158 insights gained from these analyses were used to introduce realistic patterns of missing data into
159 simulated data. We applied 31 imputation methods to the datasets and evaluated them with respect
160 to their ability to achieve correct statistical estimates and hypothesis test results in various data
161 scenarios. (iii) Finally, the imputation methods were applied to real metabolomics data (KORA F4),
162 followed by two biologically-driven evaluation schemes. First, we assessed how accurately real
163 biochemical pathways were reconstructed in data-driven correlation networks inferred from the
164 imputed data. Second, we verified whether imputation led to a gain in statistical power, while
165 preserving effects of genetic variants on metabolite levels. The study workflow is visualized in Figure
166 1.

Results

Characterization of missing data patterns in KORA F4 untargeted metabolomics data

We used an untargeted metabolomics dataset from the KORA F4 study, which was generated from fasting serum samples measured on three platforms: LC/MS in both positive (LC/MS+) and negative modes (LC/MS−), as well as a GC/MS platform. After log-transformation and outlier handling (see Methods), 1757 samples and 516 metabolites were available for analysis.

The dataset contained 19.41% missing values, with 416 (80.6%) metabolites and all observations showing at least one missing value. The majority (301) of these 416 metabolites had fewer than 10% missing values (Figure 2A). For only 9.9% (51) of the metabolites, more than 70% of the measurements were missing. The amount of missing values per observation ranged from 11.4% to 32.2%, with an average of 19.6% (Figure 2B).

LOD-based missing values

For metabolomics data, a common assumption is that missing values occur because of low concentrations that are below the limit of detection. To explore this assumption, we analyzed missing values of a metabolite using a second, strongly correlated metabolite, which we term the *auxiliary* metabolite. The auxiliary metabolite is defined as the metabolite with the highest correlation (r) to the given metabolite. Due to its strong correlation, we assume that insights into the pattern of missing values of a metabolite can be gained from the corresponding non-missing observations of its auxiliary metabolite. For example, assuming that metabolite A has missing values in certain observations for which its auxiliary metabolite B has measurements. If these measurements in B are low then a missing value in A most likely occurred because the actual concentrations were below the LOD. We required a minimum correlation of $r = 0.3$ for auxiliary metabolites, but other values gave qualitatively similar results (File S1).

191 Overall, an auxiliary metabolite was available for 56.6% of the metabolites. Of those, 62.0%
192 showed a clear tendency for missing values to below the LOD (see Methods and File S1). An example
193 for a clear LOD-tendency is shown for 7-methylxanthine in Figure 2C. This compound is a metabolite
194 of caffeine metabolism that is correlated with 3-methylxanthine. The majority of observations with
195 missing data in 7-methylxanthine showed low values for 3-methylxanthine, indicating that the 7-
196 methylxanthine values were most probably below the LOD. An example for a metabolite pair that
197 does not show an LOD-based missingness pattern is provided in Figure 2D for 1-
198 arachidonoylglycerophosphocholine (1-AGPC) and its auxiliary metabolite 1-
199 docosahexaenoylglycerophosphocholine (1-DGPC). Unlike the previous example, observations with
200 missing data for 1-AGPC showed values varying over the whole range of 1-DGPC. Consequently, this
201 suggests that LOD does not adequately explain the pattern of missing values for 1-AGPC. Scatterplots
202 of investigated metabolites and their corresponding auxiliary metabolites, as well as boxplots of
203 concentrations in the auxiliary metabolites for missing and non-missing observations in the
204 investigated metabolites can be found in File S1.

205 Although the LOD-tendency was observed for many metabolites, there was no clear LOD threshold
206 separating missing and observed measurements across all metabolites (Figure 2C), which would have
207 been the case if LOD was the only underlying mechanism for missing data. Instead, the values of the
208 auxiliary metabolites with missing values in the investigated metabolites were spread broadly over a
209 range of lower values, indicating a blurred rather than a single fixed LOD for all metabolites.

210 ***Run day-dependent missing values***

211 Batch (run day) effects also can drive systematic patterns of missing data due to daily variation in
212 instrument sensitivity. To examine whether missing data depended on overall run day quality, we
213 examined the amount of missing values per run day for each platform (LC/MS+, LC/MS–, or GC/MS).
214 Subsequently, we investigated whether metabolites were affected differently by runday quality.

215 The KORA F4 samples were measured on 53 run days with 34 samples on average per day. If
216 missing values were dependent on run day quality due to variation in instrument performance (e.g.,
217 caused by LC or GC column decline), we would expect there to be some days for which samples
218 overall contained more ("bad" run day) or fewer ("good" run day) missing values compared with the
219 average. Indeed, we observed such "bad" and "good" run days for all three platforms (Figure 3A).
220 While the run day-specific amount of missing values tended to be correlated between LC/MS− and
221 LC/MS+ (correlation of the run day-specific median of missing values between the two platforms was
222 $r = 0.36$), there was no correlation between LC/MS+/− and GC/MS. This suggests that changes in
223 instrument performance, rather than global effects (such as those that could originate from sample
224 preparation) were responsible for differences in run day quality.

225 Although there was an overall effect of run day quality on the pattern of missing values, we
226 observed considerable differences in the standard deviations (SD) of run day-specific missing values
227 for metabolites with the same amount of missing data (Figure 3B). This suggests that metabolites
228 were affected differently by run day quality. For example, the bile acid ursodeoxycholate (46% total
229 missing data) showed relatively low variation in run day missing data (SD = 0.12) (Figure 3**Figure 3**C).
230 However, for gamma-glutamylisoleucine (Figure 3D), a metabolite with a similar total amount of
231 missing values (42%), the observed variation in missing data across run days was substantially larger
232 (SD = 0.22).

233 ***Run day-dependent LOD mechanism***

234 The observed run day-dependent pattern of missing data, together with the blurred LOD-based
235 pattern, suggests that different run days may exhibit different LODs, which contributed to the blurred
236 global LOD effect. To verify this, we calculated the correlation between run day mean and run day
237 missingness for all metabolites. A histogram of the correlation coefficients is shown in Figure 4A. The
238 majority of metabolites displayed a strong tendency for negative correlations. An example for run
239 day-specific LODs is shown in Figure 4B–C: for 7-methylxanthine, the correlation of run day mean and

240 the run day-specific amount of missing values is $r = -0.68$ (Figure 4B). Run days with low means
241 tended to have a higher amount of missing values (Figure 4C). Density plots for all metabolites before
242 and after run day normalization can be found in File S2.
243
244 Taken together, we observed that batch (run day) effects on the limit of detection can result in a
245 blurred LOD-effect after run day normalization, which can explain patterns of missing values in most,
246 but not all, metabolites.
247

248 Evaluation of imputation approaches in a simulation framework

249 As shown in the previous analyses, not all of the missing data in MS-based metabolomics studies can
250 be attributed to run day-dependent LOD-based missing data. Thus, the optimal imputation approach
251 should perform well across all possible patterns. We conducted a simulation study to compare
252 statistical estimates between imputed and complete data. We simulated incomplete data according
253 to the patterns of missing values observed in the real metabolomics data and imputed these data
254 using various imputation approaches. We then evaluated these approaches for recovering correct
255 statistical estimates after conducting correlation and regression analyses.

256 *Simulation setup and evaluation criteria*

257 We simulated six mechanisms for missing data derived from observations in the real data (see
258 Methods, File S3, and Figure 5A–E): (i) *Fixed LOD,* as an extreme form of systematic missing values
259 below a global LOD; (ii) *Probabilistic LOD*, where the probability of a missing value increases at lower
260 values, which should resemble the blurred LOD-based patterns observed in the real data; (iii) *Run*
261 *day-specific fixed LOD*, where LOD is assumed to vary across run days; (iv) *Run day-specific*
262 *probabilistic LOD*, where a probabilistic form of LOD is assumed to occur across run days; (v)
263 *Unsystematic (random) missingness*, for missing data with an unknown reason; and (vi) *Mixtures of*
264 *LOD-based and unsystematic missingness*. Based on these 6 mechanisms, we created various

265 parameter scenarios resembling realistic conditions. For each scenario, we conducted 250
266 simulations to assess whether the imputation methods could reconstruct statistical estimates of
267 Pearson correlation, partial correlation, linear regression (results shown in File S3), and logistic
268 regression. To this end, we calculated type 1 error as the proportion of simulations in which a
269 significant estimate was obtained when the true correlation was equal to zero. In addition, we
270 calculated power as the proportion of significant estimates when the true correlation was unequal to
271 zero. We also estimated bias, which is shown in File S3. A detailed description of the simulation and
272 evaluation framework is also provided in File S3.

273 ***Missing data handling strategies***

274 We applied 31 imputation approaches (see Figure 5F; detailed descriptions in Methods and File S4)
275 on the simulated data. Some were adapted to account for run day-specific missing values. The
276 imputation approaches followed different concepts, which could have one of the following four
277 properties or combinations thereof: (i) approaches that explicitly assume LOD-based missing values,
278 (ii) approaches that consider run day-specific missing values, (iii) multivariate procedures using
279 correlations among variables, and (iv) multiple imputation (MI) strategies. The MI approaches usually
280 comprise imputation, analysis, and pooling steps. In the first step, the incomplete data are imputed
281 *m* times to produce *m* complete datasets. Subsequently, statistical analysis is performed on each of
282 the *m* complete datasets and then the *m* analyses are combined to one final result.

283 ***Simulation results***

284 In the following, we evaluate the performance of the four imputation properties (i)–(iv) introduced
285 above. Simulation results from other data scenarios, all variations of the imputation approaches
286 used, and the combination of parameter settings are available in File S5.

287 Property (i): Methods that explicitly assume LOD-based missing values and perform
288 imputation globally without taking run day information into account (*min*, Richardson & Ciampi (*RC*),
289 imputation by truncated sampling (*ITS*)), showed inflated type 1 error rates and low power for both

correlation and regression analysis. This was expected for three reasons. First, for a data scenario with run day-dependent probabilistic LOD-based missing values, these methods underestimate the LOD for most of the rundays and replace missing entries by too low values (Figure 6A). Second, for a data scenario with random missing values, they expectedly fail since the underlying assumption of an LOD is not met (Figure 6B). Finally, *min* and *RC* impute a metabolite by replacing all of its missing entries by a constant value, which substantially distorts the metabolite distribution (see File S5).

Property (ii): The LOD-based methods that take run days into account (*RC-R, ITS-R*) were expected to perform well in a simulated data scenario with run day effects (Figure 6A). Unexpectedly, we observed an inflated type 1 error rate and decreased power for all three statistical analyses (Pearson correlation, partial correlation, and logistic regression). *RC-R* and *ITS-R* assume that the observed values of a metabolite follow a truncated normal distribution, which is parametrized by maximum likelihood estimation (MLE), in order to replace missing values with randomly drawn values from the truncated part. The instability of MLE due to small sample sizes available within run days could explain the poor performance of these approaches. The same poor performance was observed for scenarios with a mixture of run day-dependent LOD-based and random missing values (Figure 6C). For the dataset with only random missing values, LOD- or run day-based approaches showed the expected strong reduction in power since here the underlying assumption of a truncated normal distribution is false (Figure 6B).

Property (iii): Multivariate approaches (imputation based on chained equations *(ICE)* and *KNN*-based imputation) take into consideration the correlation between variables or observations. *ICE* approaches had high power, but an increased type 1 error rate when missing value proportions increased (Figure 6). *KNN*-based imputation on observations with variable pre-selection and K = 10 (*KNN-obs-sel(10)*) was one of the best performing methods with high power and an overall marginal type 1 error rate, even for a high amount of missing values. The power for *KNN-obs* was also high, but it showed high type 1 error rate and therefore a poor ability to correctly identify truly absent

315 associations. In contrast, *KNN-vars* had a low type 1 error rate, but decreased power, which became
316 more pronounced at higher amounts of missing values.

317 Property (iv): Single imputation procedures often underestimate the variability of statistical
318 estimates, resulting in inflated type 1 error rates. This should be avoided by approaches performing
319 multiple imputations (MI). MI versions based on LOD- (*MITS*) and run day-effects (*MITS-R*) indeed had
320 decreased type 1 error rates, although power was low (Figure 6). *MICE* with Bayesian linear
321 regression (*MICE-norm*) or predictive mean matching (*MICE-pmm*) as imputation model showed
322 negligible type 1 error rates and high power for all scenarios with up to 50% missing values. At higher
323 amounts of missing data, the power decreased considerably, but the type 1 error remained marginal
324 (File S5). A slight modification of the *MICE* algorithm applied widely in the metabolomics field (here
325 termed *MICE-avg*) was performed on each imputed data, and comprised the pooling of the imputed
326 data with subsequent statistical analyses rather than pooling the statistical estimates after analysis.
327 This approach showed high power, but increased type 1 error rates, in particular for >30% missing
328 values.

329 Taken together, when considering all patterns of missing data and all evaluation criteria,
330 *KNN-obs-sel(10)* and *MICE-norm* were the most robust approaches. For higher amounts of missing
331 data (≥50%), *MICE* showed a strong decrease in power with marginal type 1 error, whereas *KNN-obs-*
332 *sel(10)* had only slightly increased type 1 error rates with high power.

333

334 Evaluation of imputation approaches on real MS-based metabolomics data

335 We conducted a biological evaluation of all approaches using the metabolomics data from the KORA
336 F4 population study. An objective criterion for evaluation is challenging to construct, since the true
337 values underlying the missing ones are unknown. We devised two indirect tests that assessed
338 imputed values for biological validity. First, we assessed the ability of imputation methods to
339 statistically reconstruct biochemical pathways in metabolomics data. Second, we evaluated the gain

340 in statistical power while preserving the true effect size of genetic variants (SNPs) on metabolite
341 levels.

342 ***Evaluation based on pathway modularity***

343 GGMs are based on partial correlations and reflect conditional dependencies in multivariate Gaussian
344 distributions (5,22). When applied to metabolomics data, they reconstruct a precise picture of the
345 metabolic network, showing a modular topology with respect to known pathways. In other words,
346 metabolites will tend to be correlated with other metabolites from the same biochemical pathway
347 (5,22,23). We used this pathway-based modularity in a metabolic network as a quality criterion to
348 indicate whether the imputation methods generally were capable of maintaining biochemically valid
349 edges.

350 Each imputation strategy was applied to the KORA F4 metabolomics data, and a GGM was
351 estimated for each obtained dataset. Subsequently, we used *a priori* pathway annotations from
352 Metabolon Inc., where each metabolite was assigned to one pathway (e.g., branched-chain amino
353 acids, lysolipids, xanthines) to calculate pathway-based modularity (Q), according to (22,24). This
354 measure reflects the ratio of metabolite correlations within *versus* across pathways. A high Q value
355 indicates a dense within-pathway correlation compared with cross-pathways. Variability was
356 estimated by bootstrap resampling (see Methods).

357 Across all datasets, we obtained modularity values ranging from 0.384 to 0.434 (Figure 7A).
358 Imputation methods that explicitly considered the LOD-based mechanism and their run day-specific
359 versions (Figure 5, property (ii)) did not outperform alternative approaches. Multivariate, single
360 imputation methods (property (iii)) yielded low Q values, except for *KNN-obs-sel*, which achieved the
361 overall third best result (Q = 0.422 for K = 10) (Figure 5). The performance of *KNN*-based imputation
362 methods strongly depended on the definition of neighbors (variables or observations) and on the
363 number of these neighbors (K). The MI procedures (property (iv)) *MITS*, *MITS-R*, and *MICE-avg*
364 performed poorly, whereas the networks generated on *MICE* imputed data showed the overall

365 highest modularity (Q = 0.434 and Q = 0.424 for *MICE-norm* and *MICE-pmm*, respectively) (Figure 5).
366 Overall, the three best performing approaches were *MICE-norm*, *MICE-pmm*, and *KNN-obs-sel(10)*.

367 ***Evaluation based on metabolite-SNP associations***

368 Using KORA F4 data (n = 1750), we determined the ability of imputation methods to gain statistical
369 power compared with complete case analysis (*CCA*, deleting samples with any missing values) while
370 preserving the effect of genetic variants on metabolite levels in human blood. For the evaluation, we
371 selected a set of metabolite-SNP associations from a previous genome wide association study
372 (GWAS) in the KORA F4 and TwinsUK cohorts, for which a functional connection between the gene
373 and the metabolite was biologically evident (Table S8) (25). For example, GOT2 (*rs4784054*), which
374 was associated with concentrations of phenyllactate, encoded an enzyme that catalyzes the
375 conversion of phenylalanine to phenylpyruvate, which is then converted to phenyllactate (25,26).

376 We investigated the gain in statistical power when using imputed datasets compared with
377 the power obtained with *CCA* for 18 of such metabolite-SNP pairs, where the metabolite had
378 between 10% and 70% missing values. Statistical power gain was calculated as the negative log10 of
379 the ratio of the p-values estimated for the imputed data to the p-values estimated for *CCA* in
380 corresponding linear regression models (detailed results in File S8 and Table S8). A high ratio
381 indicates greater power for imputed data. As a second evaluation criterion, we calculated the log2
382 absolute ratio of the effect sizes obtained from the regression models for imputed data and those
383 derived from *CCA* in KORA F4 (see Methods). A log2 ratio close to zero indicates that the imputation
384 method was able to preserve effect sizes, whereas imputations yielding a highly negative or positive
385 log2 ratios indicate underestimation or overestimation of the effect sizes, respectively.

386 Imputation with LOD-based methods (property (i)) yielded a gain in power for up to seven
387 genetic associations of the 14 metabolites (Figure 7**Figure 7**). For two of these associations
388 (tetradecanedioate and SLCO1B1; and hexadecanedioate and SLCO1B1), effect sizes were
389 underestimated, and for the association between 1-methylurate and NAT2, the effect size was

390 overestimated across all methods, except for *MITS-R*. Run day-specific imputation methods (property
391 (ii)) performed well, with *ITS-R* yielding the highest number of associations (12) with greater
392 statistical power, of which seven showed effect sizes similar to effect sizes derived from *CCA*. The
393 best methods among multivariate approaches (property (iii) and (iv)) were *MICE-avg-norm*, *KNN-obs-*
394 *sel(10)*, *and KNN-obs-sel(20)*, all three of which generated a gain in statistical power for 12
395 associations. These methods also showed good performance in preserving genetic effects and did not
396 show severe overestimation or underestimation of effect sizes. *MICE-norm/-pmm/-adjR* showed only
397 moderate performance with a power gain for seven associations.

398 In an additional analysis, we used results from the EPIC-Norfolk cohort with n = 10 634
399 subjects (27), to assess the ability of imputation methods to preserve effects of genetic variants on
400 metabolites. We hypothesized that the effect sizes would be estimated more accurately in this much
401 larger dataset, and effect sizes obtained with KORA F4 imputed data should approximate effect sizes
402 derived from EPIC-Norfolk. Overall, we observed that the majority of SNP-metabolite pairs showed
403 either an overestimation or an underestimation of effect sizes across all imputation methods. This
404 tendency might reflect differences between the cohorts KORA F4 and EPIC-Norfolk rather than
405 differences between imputation strategies (see detailed results in File S7 and Table S8).

406 Overall, for nearly all metabolite-SNP pairs, this analysis showed that statistical power was
407 increased by imputing missing values and the effect sizes could be preserved. *ITS-R*, *MICE-avg-pmm,*
408 *KNN-obs-sel* with *K* = 10 and *K* = 20 were the imputation methods that generated the highest number
409 of associations (12) and resulted in a gain in statistical power compared with *CCA*.

410

411 ## Discussion

412 In this study, we investigated patterns of missing data in a typical example of untargeted MS-based
413 metabolomics data and their possible underlying mechanisms. Insights gained from these analyses
414 were used to generate simulated data that reflected the real data situation for a comprehensive
415 evaluation of 31 imputation methods. Finally, we applied the imputation strategies to real MS-based
416 metabolomics data from the German KORA F4 study and evaluated them using biological validity
417 measures.

418 For metabolomics data, an intuitive assumption is that missing data occur when metabolite
419 concentrations fall below the machine's LOD. Indeed, we found evidence for systematic patterns of
420 missing data due to LOD- and batch-effects for a large proportion of the analyzed metabolites.
421 Missing data were found to be influenced by run day quality, although metabolites varied in their
422 susceptibility to this effect. Finally, we found a negative correlation between run day mean and
423 missing data per run day, further confirming LOD-based mechanism within run days. The existence of
424 multiple run day-dependent LODs possibly accounted for the blurred rather than fixed global LOD
425 observed in the data. It has been suspected that multiple detection limits arise from factors such as
426 batch (run day) effects (27). However, to the best of our knowledge, this is the first time that these
427 effects have been systematically explored so far.

428 We evaluated 31 imputation methods in an evaluation framework consisting of three
429 schemes: (i) unbiased estimation of statistical estimates and hypothesis test results based on
430 simulated data, (ii) statistical reconstruction of biochemical pathways in metabolic networks, and (iii)
431 the ability to preserve effects of genetic variants on metabolite levels while allowing for a gain in
432 statistical power.

433 *MICE-norm* was the best performing imputation method for evaluation scheme (i) and (ii), but it
434 showed only moderate performances in the metabolite-SNP analysis. One major drawback of this
435 method is that multiple imputations have to be performed, making these approaches statistically and

436 computationally challenging. For *m* imputations, the desired statistical analyses must be performed
437 on each of the *m* imputed datasets, and then the resulting *m* estimates must be combined to one
438 statistical result. A widely applied alternative is to perform *m* multiple imputations and then combine
439 the *m* complete datasets to one final dataset containing the average of the imputed values (*MICE-*
440 *avg*). That is, *MICE-avg* does not require statistical estimates to be pooled, and therefore, it is much
441 easier to apply. However, this simplicity is accompanied by an underestimation of metabolites'
442 variances, resulting in poorer performance of statistical estimation (correlation and regression
443 coefficients) and reconstruction of biochemical pathways.

444 A feasible, but better performing method was *KNN-obs-sel(10)*, which uses *KNN*-based
445 imputation on observations with variable pre-selection and K = 10. This method ranked highly in all
446 evaluation schemes. Other *KNN*-based imputation schemes, including *KNN*-based imputation on
447 variables (*KNN-vars*) and on observations without variable pre-selection (*KNN-obs*), consistently
448 showed poor performance across all evaluation schemes. Our results are in line with observations
449 from previous studies, where *KNN*-based imputation performed well (10,11,15,28). However, we also
450 observed that variations of *KNN* imputation lead to substantially different results, as in previous
451 studies (20,28).

452 Although we observed LOD- and run day-based effects in real metabolomics data, methods
453 that explicitly consider this information did not outperform competing approaches in the first two
454 evaluation schemes. This is likely due to the fact that they perform imputation in a univariate manner
455 without taking the correlation between the variables into account. Moreover, all of these LOD-based
456 methods include maximum likelihood estimation in their imputation process, which was found to
457 perform well only for larger sample sizes in previous studies (27,29). In our study, the number of
458 observations within run days is limited, resulting in considerable instability of the MLE. LOD-based
459 run day-dependent methods performed well with respect to gain in statistical power in the analysis
460 of metabolites–SNP associations.

In summary, we have presented a detailed description of patterns of missing data in untargeted MS-based metabolomics data. In particular, we considered, for the first time, the effects of run days on systematic patterns of missing data. Our work showed that missing data occur in most cases due to LOD effects, which are moreover run day-dependent. Nevertheless, *MICE* and *KNN*-based imputation, methods that do not explicitly consider LOD-based effects, performed best when tested in both statistical and biological evaluation schemes. This is most likely because these methods take into account multivariate dependencies within the data. The two approaches are For future studies, we recommend *KNN*-based imputation on observations with K = 10, since it consistently performed well across all data scenarios and all evaluation schemes, and is computationally non-demanding for daily data analysis.

472 # Material and Methods

473 ## Study cohort, metabolomics and genotype measurements

474 Data from 1768 fasting serum samples of the German Cooperative Health Research in the Region of
475 Augsburg (KORA F4) population cohort (30) was used, comprising 910 females and 858 males. Age
476 distribution was 60.53 ± 8.79 years for females and 61.20 ± 8.78 years for males. Body mass index
477 (BMI) distribution was 27.88 ± 5.24 $\mathrm{kg/m^2}$ for females and 28.46 ± 4.29 $\mathrm{kg/m^2}$ for males.

478 Serum metabolomics measurements were performed on three platforms, LC/MS− (negative
479 mode), LC/MS+ (positive mode), and GC/MS by Metabolon, Inc. (Durham, NC, USA). The 1768 serum
480 samples were measured on 53 different run days, with 34 samples on average per run day. A total of
481 516 metabolites were quantified, of which 303 had an identified chemical structure. A more detailed
482 description of sample acquisition, experimental procedures, and metabolite identification can be
483 found in File S10.

484 Each known metabolite was annotated with one of 68 pathways by Metabolon, Inc. A full list
485 of all measured metabolites, including pathway annotations, can be found in Table S9. For correlation
486 analysis, data were normalized for run day-effects by dividing each metabolite by run day median.
487 Since metabolite measurements were assumed to follow a log-normal distribution, the data were
488 log-transformed for all statistical analyses. The run day-corrected and log-transformed data were
489 used to determine outlier samples. Eleven individuals with a Mahalanobis distance (calculated across
490 the complete dataset) greater than four SD from the mean were considered outliers and excluded
491 from the dataset. For the biological evaluation schemes, age, sex, and BMI were used as standard
492 covariates. Seven samples were excluded due to incomplete information in these phenotypes,
493 resulting in 1750 individuals in total.

494 The KORA F4 cohort was genotyped using the Affymetrix Axiom platform. After quality
495 control, genotype data (measured or imputed according to data from the 1000 genomes project,
496 phase 1 version 3) were available for 1685 of the 1750 individuals.

497 Missing data in KORA F4

498 To explore the mechanism for the missing data of a given metabolite m, a second (auxiliary)
499 metabolite m_{aux} was used. m_{aux} was defined as the metabolite with the strongest Pearson
500 correlation to m (at least 0.3). An LOD-tendency was assumed if the average value of m_{aux} in
501 samples with missing values in m was significantly lower than the average of m_{aux} in samples with
502 measured values in m. Significance was assessed using Wilcoxon–Mann–Whitney tests with $\alpha = 0.05$
503 after Bonferroni correction for multiple testing.

504 For all correlation analyses, only metabolites with more than 10% and less than 70% overall
505 missing values were considered.

506 In order to explore whether missing values varied among run days, the normalized
507 proportions of missing values among the 53 run days were compared within each platform. For a
508 metabolite m and a run day d, the normalized amount of run day-specific missing values was
509 calculated as the number of missing values for m in d divided by the total number of samples
510 measured in d, divided by the median value of missing data of m over all run days.

511 Simulation study

512 Insights gained from the analyses of missing values in real MS-based metabolomics data were used to
513 create artificial data that best mirror reflected patterns of missing data. A brief overview of the
514 simulation framework is provided below, and a detailed description can be found in File S3. For each
515 set of parameters corresponding to a certain data situation, 250 random datasets were generated.
516 For each dataset, two variables were simulated by drawing from a multivariate normal distribution,
517 with sample sizes ranging from 100 to 1000, and with means equal to zero and covariance chosen
518 such that variances were equal to one (representing scaled variables). The Pearson correlation

519 between the two variables was ranged from 0 to 0.4. In addition, for the multivariate analyses and to
520 evaluate imputation methods that apply to a multivariate strategy, auxiliary variables correlated with
521 the two main variables were introduced. Their number and correlation strength were chosen to
522 match the real data (for details, see File S3).

523 Simulated observations were randomly assigned to "run days" with the number of run days
524 chosen such that each run day comprised 34 observations, according to the average number found
525 for the real KORA F4 measurements.

526 A proportion of missing values (10%, 30%, 50%, and 70%) was introduced into the main
527 variable pair according to different mechanisms derived from our observations in the KORA F4
528 Metabolon data (Figure 5, File S3).

529 We used the following parameter settings for the results in the main manuscript: moderate
530 variability of missing data across run days (see File S3), uncorrelated run day-specific missing patterns
531 of the metabolite pair, and varying association of the inverse relation between metabolite
532 concentration and missing values, at $n = 250$ and in the presence of informative auxiliary
533 metabolites. For Pearson and partial correlation analysis, both main variables had the same degree of
534 missing data. For logistic regression analysis, the predictor variable had a mixture of 50% run day-
535 dependent probabilistic LOD-based missing data and 50% non-systematic missing data. Results for
536 more parameter settings can be found in File S5.

537 Imputation approaches

538 A variety of imputation methods (Figure 5**Figure 5**) were selected because they were reported in the
539 context of metabolomics data or were developed and adopted to address characteristics in the
540 current dataset.

541 ***Mean imputation (mean):*** All missing values of each incomplete variable are replaced by the average
542 of the observed values of that metabolite. ***Minimum imputation (min):*** All missing values of each

543 incomplete variable are replaced by the smallest observed value of that metabolite (5,13,16).
544 ***Richardson & Ciampi (RC):*** Assuming that missing values occur due to LOD and the observed
545 metabolite values follow a left-truncated normal distribution, maximum likelihood is used to
546 estimate this distribution. A missing value x is then replaced by the expected value of x conditional
547 on x being below the LOD, $E(x|x \leq LOD)$ (17). ***Imputation by truncated sampling (ITS):*** This is an
548 extension of the *RC* method, where the missing values are replaced by randomly drawn values from
549 the censored part of the estimated truncated normal distribution. ***Multiple imputation by truncated***
550 ***sampling (MITS):*** *ITS* is applied as described above, but multiple imputation is performed according
551 to Rubin's rules (31) using the *R* package *mice*, version 2.25. These rules include: (i) the datasets are
552 imputed m times, (ii) each of the m completed datasets is analyzed separately, and (iii) the m
553 resulting estimates are combined using established procedures (31–33). The number of imputations
554 was set to $m = 20$ for all methods. ***Runday-specific LOD-based methods (RC-R/ITS-R/MITS-R):*** The
555 previously described methods *RC, ITS,* and *MITS* are applied within run days where at least 17
556 observations are available. In *RC-R*, the remaining missing values are set to the mean of all available
557 expected values. For *ITS-R* and *MITS-R*, the remaining missing values are replaced using *ICE-norm* (see
558 below). ***Imputation by chained equations (ICE-norm/-pmm/-adjR)*** was performed using the *R*
559 package *mice*, version 2.25. It uses a repeated chain of equations through the incomplete variables,
560 where in each imputation model, the respective incomplete variable is modeled as a function of the
561 remaining variables (34–36). In *ICE-norm*, a Bayesian linear regression is used as the imputation
562 model, whereas in *ICE-pmm* (predictive mean matching as imputation model), missing values are
563 replaced by a random draw of measured values from other observations with the closest predicted
564 values. In *ICE-adjR*, a model is specified with random intercept per run day, which aims to better
565 utilize run day information. This model assumes that variable values (i.e., metabolite concentrations)
566 have a run day-specific component, which varies randomly following a normal distribution. ***Multiple***
567 ***imputation by chained equations (MICE-norm/-pmm/-adjR)*** was performed using the *R* package
568 *mice*, version 2.25: *MICE-norm, MICE-pmm,* and *MICE-adjR* consisted of $m = 20$ parallel imputation

runs of *ICE-norm, ICE-pmm,* and *ICE-adjR,* respectively. Subsequently, the estimates are combined using Rubin's rules as described above for *MITS*. ***MICE average version (MICE-avg-norm/-pmm)***: *ICE-norm* or *ICE-pmm* is applied multiple ($m = 20$) times in parallel, followed by combining the m imputed datasets to one final dataset as the average of the imputed values. ***K-nearest neighbor imputation (KNN-var(K)/KNN-obs(K)/KNN-obs-sel(K)):*** In *KNN-var* and *KNN-obs*, missing values of each variable are replaced by the weighted average of pre-specified *K* nearest variables and observations, respectively. Distances to neighbors were defined as Euclidean distance and weights were chosen as e^{-d}, where d defines the distances between two variables or observations. In *KNN-obs-sel, KNN-obs* is performed by selecting the strongest correlated variables with $|\rho| \geq 0.2$, but it was constrained to a minimum of 5 and a maximum of 10 variables. The number of neighbors for *K* was set to 3, 5, 10, and 20.

More detailed descriptions of *RC*, *RC-R*, *ITS*, *MITS*, *ICE*, and *KNN*-based methods can be found in File S4. The two best performing methods, *KNN-obs-sel(K)* and *MICE* are available as R code in File S11.

Statistical evaluation of missing data handling strategies in the simulation study

Pearson correlation, partial correlation, linear regression, and logistic regression analysis were performed, and the ability of imputation methods to reconstruct true associations and unbiased hypothesis test results was evaluated. For logistic regression, a dichotomized variable was simulated by discretizing one of the simulated continuous variables: all values above the median were set to 1 and all values below the median were set to 0. This dichotomized variable was used as response and the remaining continuous variable as predictor. For MI strategies, the resulting (correlation or regression coefficient) estimates and their variances were combined using Rubin's rules. The obtained point estimates were then compared with the true underlying values by assessing the validity of hypothesis tests. To this end, type 1 error was calculated as the proportion of significant estimates (at $\alpha = 0.05$) after imputation when there was no true effect. Power was calculated as the

593 proportion of significant estimates (at $\alpha = 0.05$) after imputation in the presence of a true effect.
594 Detailed results can be found in File S5.

595 Evaluation based on pathway modularity

596 This analysis was based on pathway annotations from Metabolon Inc. (see Supporting Information
597 S9). Each imputation strategy was applied to the KORA F4 metabolomics data, resulting in different
598 imputed datasets. All unknown metabolites were excluded since these compounds were not assigned
599 to a pathway. For each imputed dataset, a Gaussian graphical model (GGM) was estimated to infer a
600 network using the *R* package *GeneNet*, version 1.2.12. In previous studies, we have demonstrated
601 that these models correctly reconstruct biochemical pathways from the data (22,25,37). In the case
602 of MIs, a GGM was estimated for each imputed dataset, followed by combining partial correlations
603 using Rubin's rules after a Fisher Z-transformation. The network was constructed using partial
604 correlations that are significantly different from zero after Bonferroni correction for $n * (n-1)/2$,
605 where n is the number of metabolites.

606 The pathway-based network modularity measure Q (22,24) was calculated for each network as

607 $$Q = \sum_{i=1}^{|S|} \left[\frac{A(V_i,V_i)}{A(V,V)} - \left(\frac{A(V_i,V)}{A(V,V)} \right)^2 \right],$$

608 where $|S|$ is the total number of pathways, V is the set of all metabolites, and V_i describes the subset
609 of metabolites annotated with pathway i. $A(V_i, V_j)$ is the number of edges between any two node
610 sets V_i and V_j. The variance of Q was estimated non-parametrically using bootstrapping of the
611 original dataset (R package *boot*, version 1.3-15) with 1000 runs.

612 Evaluation based on metabolite-SNP associations

613 Linear regression was performed using KORA F4 *CCA* and the results were compared with each other.
614 For this analysis, we selected metabolite-SNP pairs for which (i) a genome-wide significant
615 association could be identified in the meta-analysis of KORA F4 and TwinsUK cohorts in a previous
616 GWAS (25) (summary statistics retrieved from http://www.gwas.eu); (ii) the proportion of each

617 metabolite's missing values in KORA F4 was between 10% and 70%; (iii) the metabolite was
618 measured in the EPIC-Norfolk cohort, which we used to further benchmark the preservation of effect
619 sizes; and (iv) a functional connection between the genetic locus of the SNP and the metabolite (e.g.,
620 metabolite is a known substrate of the transporter) was evident according to manual curation of the
621 GWAS results (Table S8). For each imputed dataset, 18 metabolite-SNP pairs were tested for genetic
622 association using age- and sex-corrected linear regression models under the assumption of an
623 additive genetic model ($\text{metabolite} \sim \beta_0 + \beta_1 \times \text{SNP} + \beta_2 \times \text{age} + \beta_3 \times \text{sex}$). To avoid spurious
624 associations, metabolic data points greater than four SDs from the mean were removed prior to
625 computing linear models. For MI approaches, the regression coefficients were pooled using Rubin's
626 rules as provided by the *R* package *mice*, version 2.25. For each metabolite-SNP pair, the variance of
627 the regression coefficients and p-values were estimated using bootstrapping.

628 To explore which imputation approaches increased statistical power, p-values obtained for
629 the effect sizes based on imputed data were compared with p-values obtained from *CCA* by
630 calculating their ratio as $r_p = \frac{-\log_{10}\left(\frac{p_{imp}}{p_{CCA}}\right)}{-\log_{10}(p_{CCA})}$, where p_{imp} was the p-value obtained for imputed data
631 and p_{CCA} was the p-value derived from *CCA*. A ratio less than or equal to zero indicated either no
632 power gain or a power loss, whereas a ratio greater than zero indicated a drop in p-value, which
633 suggested that statistical power increased when imputation was performed.

634 In addition to statistical power gain, the imputation approaches should be able to preserve
635 effect sizes compared to *CCA*. Standardized effect sizes obtained from the imputed data (β_{imp}) were
636 compared with standardized effect sizes estimated for *CCA* (β_{CCA}) based on the KORA F4 data (n =
637 1750) and the EPIC-Norfolk data (n = 10 634), assuming estimates from the EPIC-Norfolk data to be
638 close to true effects. We calculated the ratio $r_\beta = \log_2(|\frac{\beta_{imp}}{\beta_{CCA}}|)$, with a low ratio indicating a similar
639 effect size between the imputed data and *CCA*. A highly negative or positive r_β indicates an

underestimation or overestimation of the effect sizes in imputed data, respectively. A well performing imputation method is assumed to obtain high r_p and low absolute r_β.

643 ## Figures and Tables
644

645 **Figure 1. Flow chart of the study design.** Pre-processed KORA F4 metabolomics data were
646 used to analyze patterns of missing values in the dataset. Possible underlying mechanisms
647 were inferred and implemented in a simulation framework to generate data resembling the
648 observed patterns. Based on these simulated data, imputation methods with different
649 characteristics were applied and evaluated. Finally, the same imputation approaches were
650 evaluated using KORA F4 metabolomics and genomics data.

651 **Figure 2. Overall amounts of missing data and LOD effects.** (A,B) The overall fraction of
652 missing values across metabolites and observations, respectively. (C,D) Scatter plots and
653 boxplots of selected metabolite pairs to illustrate missing data due to LOD and non-LOD
654 effects, respectively. Blue - observed concentrations. Red - observed values of the auxiliary
655 metabolite in observations with missing values of the investigated metabolite. Note that red
656 data points are not part of the x-axis but were plotted in the same scatterplot for clarity. *corr*
657 = correlation, *p* = p-value of correlation, $\boldsymbol{p_{Wst}}$ = p-value of Wilcoxon–Mann–Whitney test.

658 **Figure 3. Run day-dependent effects on missing data.** (A) Normalized amount of missing
659 values per run day in each platform (LC/MS+, LC/MS−, GC/MS). For a given metabolite and
660 run day, the normalized amount of missing data per run day was calculated as the number of
661 missing values for the respective metabolite on the respective run day divided by the total
662 number of observations for that run day, divided by the median amount of missing data of
663 that metabolite over all run days. Thus, a normalized run day-missingness of 1 is the average
664 run day-missingness for a given metabolite. Pearson correlation coefficients were calculated
665 across all pairs of platforms. (B) Standard deviation of missing values across run days,
666 depending on the total amount of missing data for each platform. Each dot in the plot shows
667 the total proportion of missing values and the run day variation for one metabolite. (C)–(D)
668 The distribution of the total amount of missing values is shown for a metabolite with
669 moderate (ursodeoxycholate) and high (gamma-glutamylisoleucine) standard deviation.

670 **Figure 4. Run day-dependent LOD.** (A) Histogram of Pearson correlation coefficients of the
671 percent of missing values and run day means. (B) Scatterplot of run day mean versus percent
672 missing values, with 7-methylxanthine as an example of a negative correlation. (C) Run day
673 distributions of 7-methylxanthine before run day normalization.

674 **Figure 5. Mechanisms of missing data and imputation approaches used in the simulation**
675 **study.** (A)–(E) Mechanisms of missing values used in the simulation study, based on evidence
676 from real metabolomics data. (F) Venn diagram of imputation methods showing different
677 characteristics. Note that the figure contains complete case analysis (*CCA*), which is not an
678 imputation method, and is noted in brackets. *CCA* and *mean* were placed outside the Venn
679 diagram, as they do not comprise any of the four characteristics. LOD: limit of detection.

680 **Figure 6. Simulation results for Pearson, partial correlation, and logistic regression analysis.**
681 Performance of imputation approaches in data scenarios where (A) both variables followed a
682 run day-specific probabilistic LOD mechanism, (B) both variables showed non-systematic
683 patterns of missing data, and (C) one variable with run day-specific probabilistic LOD-based
684 missing data and the other variable showed non-systematic patterns of missing data. Type 1
685 error and power reflect the false positive and true positive rate of hypothesis testing,
686 respectively. Note that power = 1 - type 2 error rate. Note further that due to readability
687 issues, only KNN-based imputation methods with *K* = 3, 10, and 20 were included, whereas
688 KNN imputation with *K* = 1 and 5 can be found in File S5.

689 **Figure 7. Evaluation of imputation approaches on real data.** (A) Pathway-based modularity
690 for each imputation strategy. Modularity Q was calculated based on pathways. Vertical lines
691 represent bootstrap-based confidence intervals (1000 times resampling). (B) The ability to
692 gain statistical power and to preserve real metabolite-SNP associations after imputation.
693 Circle color represents the ability of imputation methods to preserve effect sizes, with red
694 and blue indicating possible overestimation and underestimation, respectively, and yellow
695 corresponding to cases with good preservation of the association. Circle size depicts the gain
696 in statistical power after imputation. The bigger the circle the higher the statistical power
697 gain after imputation compared to *CCA*. Squares correspond to cases where no statistical
698 power was gained. Note that due to readability issues, only KNN-based imputation methods
699 with *K* = 3, 10, and 20 were included, whereas KNN imputation with *K* = 1 and 5 can be found
700 in File S6 and Table S8.

701

702 Funding statement

703 This work was supported by grants from the German Federal Ministry of Education and Research
704 (BMBF), by BMBF Grant no. 01ZX1313C (project e:Athero-MED) and Grant no. 03IS2061B (project
705 Gani_Med). Moreover, the research leading to these results has received funding from the European
706 Union's Seventh Framework Programme [FP7-Health-F5-2012] under grant agreement n° 305280
707 (MIMOmics) and from the European Research Council (starting grant "LatentCauses"). KS is
708 supported by Biomedical Research Program funds at Weill Cornell Medical College in Qatar, a
709 program funded by the Qatar Foundation. The KORA Augsburg studies were financed by the
710 Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg,
711 Germany and supported by grants from the German Federal Ministry of Education and Research
712 (BMBF). Analyses in the EPIC-Norfolk study were supported by funding from the Medical Research
713 Council (MC_PC_13048 and MC_UU_12015/1).

714

715 The funders had no role in study design, data collection and analysis, decision to publish, or
716 preparation of the manuscript.

717

718 ## References

719 1. Fearnley LG, Inouye M. Metabolomics in epidemiology: from metabolite concentrations to
720 integrative reaction networks. Int J Epidemiol. 2016 Apr 26;dyw046.

721 2. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat
722 Rev Mol Cell Biol. 2012 Apr;13(4):263–9.

723 3. Blow N. Metabolomics: Biochemistry's new look. Nature. 2008 Oktober;455(7213):697–700.

724 4. Mook-Kanamori DO, Selim MME-D, Takiddin AH, Al-Homsi H, Al-Mahmoud KAS, Al-Obaidli A, et
725 al. 1,5-Anhydroglucitol in Saliva Is a Noninvasive Marker of Short-Term Glycemic Control. J Clin
726 Endocrinol Metab. 2014 Jan 1;99(3):E479–83.

727 5. Do KT, Kastenmüller G, Mook-Kanamori DO, Yousri NA, Theis FJ, Suhre K, et al. Network-based
728 approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and
729 saliva. J Proteome Res. 2015 Feb 6;14(2):1183–94.

730 6. Urpi-Sarda M, Almanza-Aguilera E, Tulipani S, Tinahones FJ, Salas-Salvadó J, Andres-Lacueva C.
731 Metabolomics for Biomarkers of Type 2 Diabetes Mellitus: Advances and Nutritional
732 Intervention Trends. Curr Cardiovasc Risk Rep. 2015 Feb 17;9(3):1–12.

733 7. Rasmiena AA, Ng TW, Meikle PJ. Metabolomics and ischaemic heart disease. Clin Sci. 2013 Mar
734 1;124(5):289–306.

735 8. Rhee EP, Gerszten RE. Metabolomics and Cardiovascular Biomarker Discovery. Clin Chem. 2012
736 Jan 1;58(1):139–47.

737 9. Wang JH, Byun J, Pennathur S. Analytical Approaches to Metabolomics and Applications to
738 Systems Biology. Semin Nephrol. 2010 Sep 1;30(5):500–11.

739 10. Armitage EG, Godzien J, Alonso-Herranz V, López-Gonzálvez Á, Barbas C. Missing value
740 imputation strategies for metabolomics data. Electrophoresis. 2015 Dec;36(24):3050–60.

741 11. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: an
742 undervalued step in the data processing pipeline. Metabolomics. 2011 Oct 8;8(1):161–74.

743 12. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, et al. Influence of Missing Values
744 Substitutes on Multivariate Analysis of Metabolomics Data. Metabolites. 2014 Jun 16;4(2):433–
745 52.

746 13. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data
747 analysis and interpretation. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W652–60.

748 14. Redestig H, Kobayashi M, Saito K, Kusano M. Exploring Matrix Effects and Quantification
749 Performance in Metabolomics Experiments Using Artificial Biological Gradients. Anal Chem.
750 2011 Jul 15;83(14):5645–51.

751 15. Di Guida R, Engel J, Allwood JW, Weber RJM, Jones MR, Sommer U, et al. Non-targeted UHPLC-
752 MS metabolomic data processing methods: a comparative investigation of normalisation,
753 missing value imputation, transformation and scaling. Metabolomics [Internet]. 2016 [cited
754 2017 Jan 13];12. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4831991/

755 16. Chen H, Quandt SA, Grzywacz JG, Arcury TA. A Distribution-Based Multiple Imputation Method
756 for Handling Bivariate Pesticide Data with Values below the Limit of Detection. Environ Health
757 Perspect. 2011 Mar;119(3):351–6.

758 17. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is
759 constrained by a lower limit. Am J Epidemiol. 2003 Feb 15;157(4):355–63.

760 18. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional
761 specification. Stat Methods Med Res. 2007 Jun;16(3):219–42.

762 19. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value
763 estimation methods for DNA microarrays. Bioinforma Oxf Engl. 2001 Jun;17(6):520–5.

764 20. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor
765 methods. Comput Stat Data Anal. 2015 Oktober;90:84–99.

766 21. Taylor SL, Ruhaak LR, Kelly K, Weiss RH, Kim K. Effects of imputation on correlation: implications
767 for analysis of mass spectrometry data from multiple biological matrices. Brief Bioinform. 2016
768 Feb 19;

769 22. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs
770 pathway reactions from high-throughput metabolomics data. BMC Syst Biol. 2011;5:21.

771 23. Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular
772 structure in biological networks. Nat Rev Genet. 2013 Oct;14(10):719–32.

773 24. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E
774 Stat Nonlin Soft Matter Phys. 2004 Feb;69(2 Pt 2):026113.

775 25. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic
776 influences on human blood metabolites. Nat Genet. 2014 Jun;46(6):543–50.

777 26. Shrawder E, Martinez-Carrion M. Evidence of phenylalanine transaminase activity in the
778 isoenzymes of aspartate transaminase. J Biol Chem. 1972 Apr 25;247(8):2486–92.

779 27. Helsel DR. More than obvious: better methods for interpreting nondetect data. Environ Sci
780 Technol. 2005 Oct 15;39(20):419A–423A.

781 28. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. Distribution based nearest
782 neighbor imputation for truncated high dimensional data with applications to pre-clinical and
783 clinical metabolomics studies. BMC Bioinformatics [Internet]. 2017 Feb 20 [cited 2017 Mar
784 16];18. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319174/

785 29. Helsel DR. Less than obvious - statistical treatment of data below the detection limit. Environ
786 Sci Technol. 1990 Dezember;24(12):1766–74.

787 30. Holle R, Happich M, Löwel H, Wichmann HE, MONICA/KORA Study Group. KORA--a research
788 platform for population based health research. Gesundheitswesen Bundesverb Ärzte Öffentl
789 Gesundheitsdienstes Ger. 2005 Aug;67 Suppl 1:S19-25.

790 31. Rubin DB. Introduction. In: Multiple Imputation for Nonresponse in Surveys [Internet]. John
791 Wiley & Sons, Inc.; 1987 [cited 2016 Feb 1]. p. 1–26. Available from:
792 http://onlinelibrary.wiley.com/doi/10.1002/9780470316696.ch1/summary

32. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009 Jul 28;9:57.

33. D'Angelo GM, Luo J, Xiong C. Missing Data Methods for Partial Correlations. J Biom Biostat [Internet]. 2012 Dec [cited 2016 Feb 28];3(8). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3772686/

34. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;18(6):681–94.

35. Van Hoewyk J, Lepkowski JM, Solenberger P, Raghunathan TE. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol. 2001 Aug 22;27(1):85–95.

36. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R | van Buuren | Journal of Statistical Software. J Stat Softw [Internet]. 2011 Dec 12 [cited 2016 Feb 28];45(3). Available from: https://www.jstatsoft.org/article/view/v045i03

37. Aichler M, Borgmann D, Krumsiek J, Buck A, MacDonald PE, Fox JEM, et al. N-acyl Taurines and Acylcarnitines Cause an Imbalance in Insulin Synthesis and Secretion Provoking β Cell Dysfunction in Type 2 Diabetes. Cell Metab. 2017 Jun 6;25(6):1334–1347.e4.

Supporting information captions

File S1. LOD tendency.

File S2. Runday-dependent densities in relation with missingness.

File S3. Simulation framework.

File S4. Imputation methods.

File S5. Simulation evaluation results.

File S6. Metabolite-SNP associations–beeswarm plots.

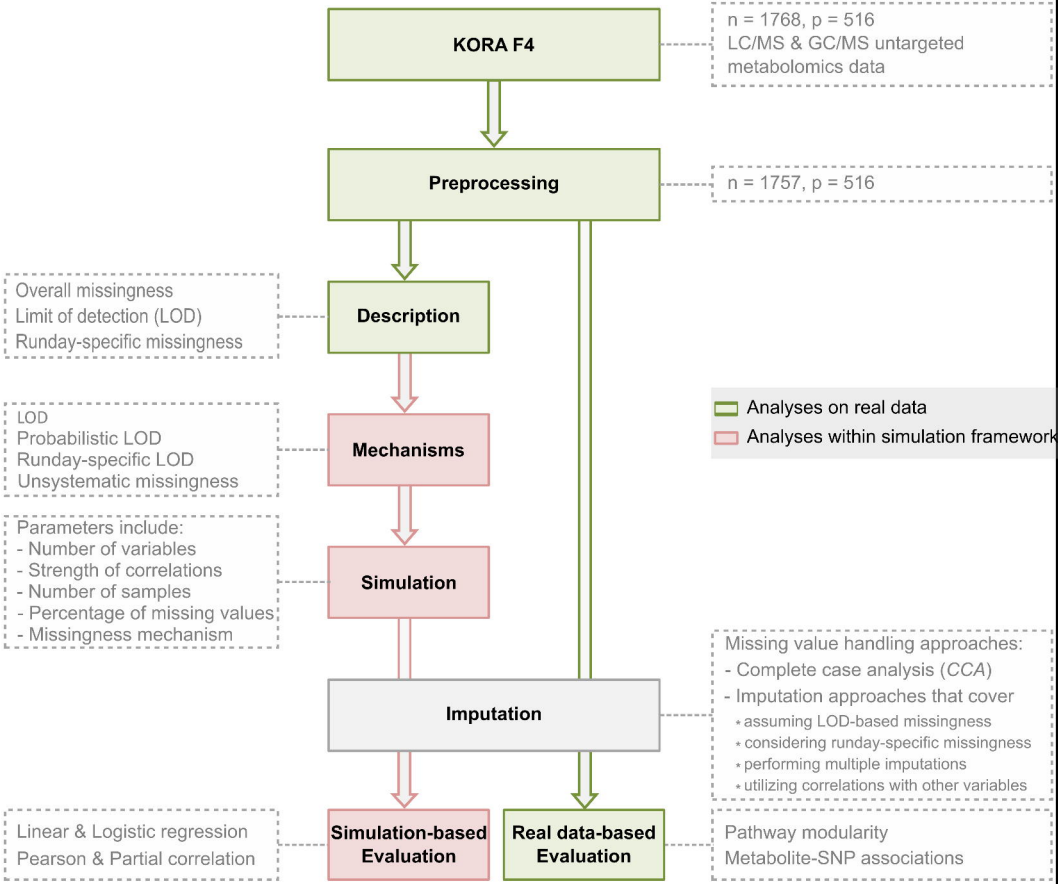
File S7. Metabolite-SNP associations compared with EPIC-Norfolk.

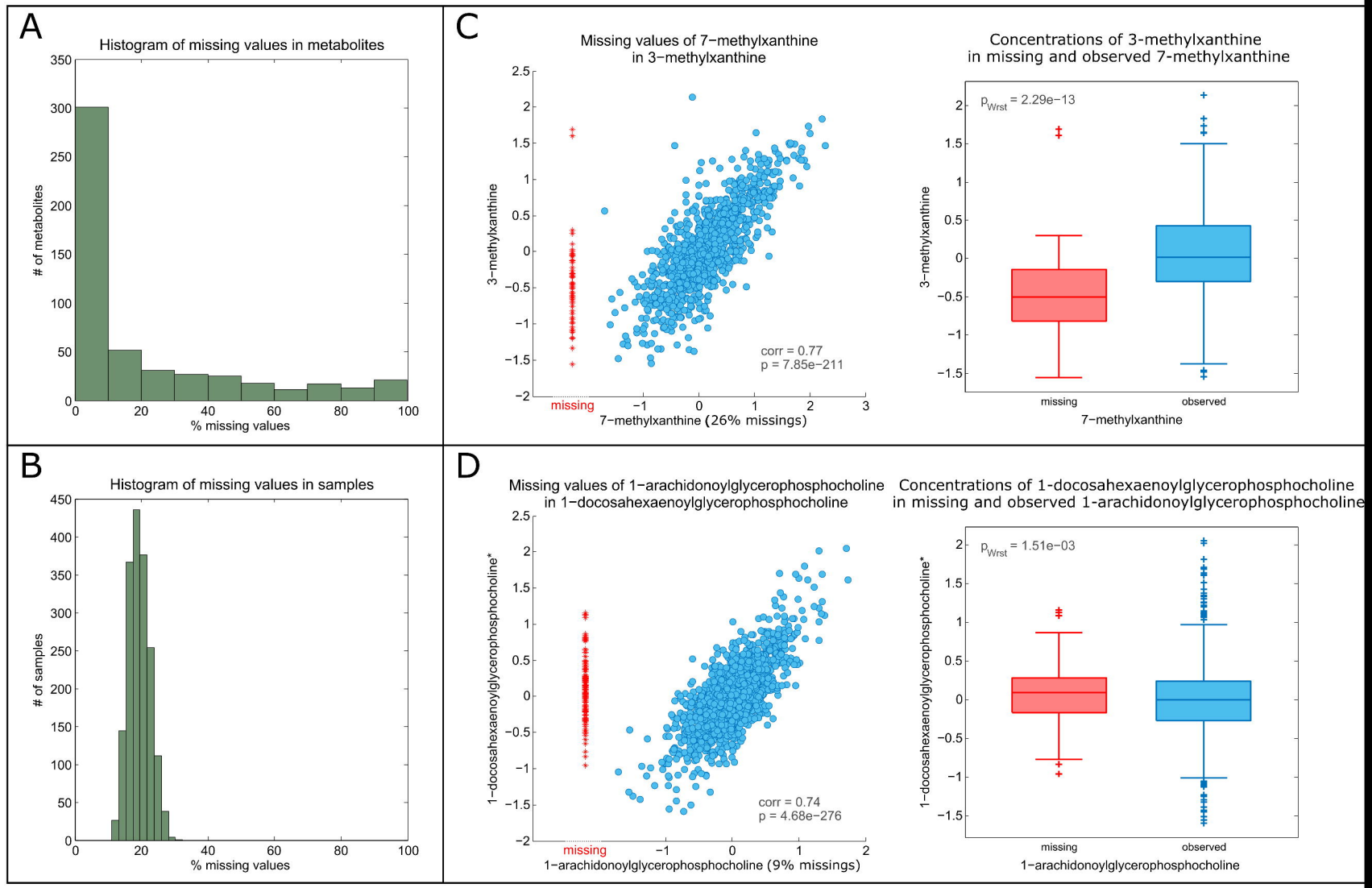
Table S8. Metabolite-SNP associations–linear regression results.

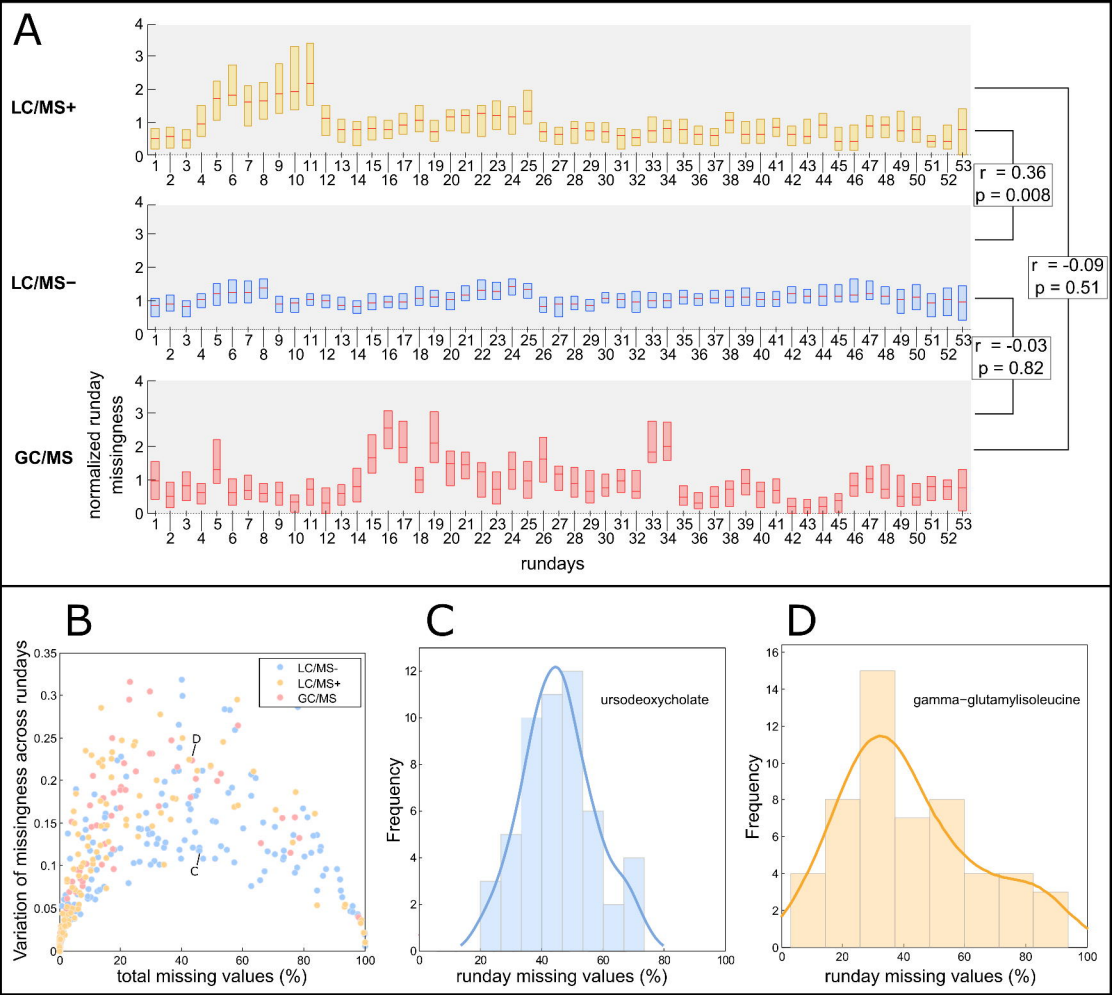
Table S9. KORA F4 annotations.

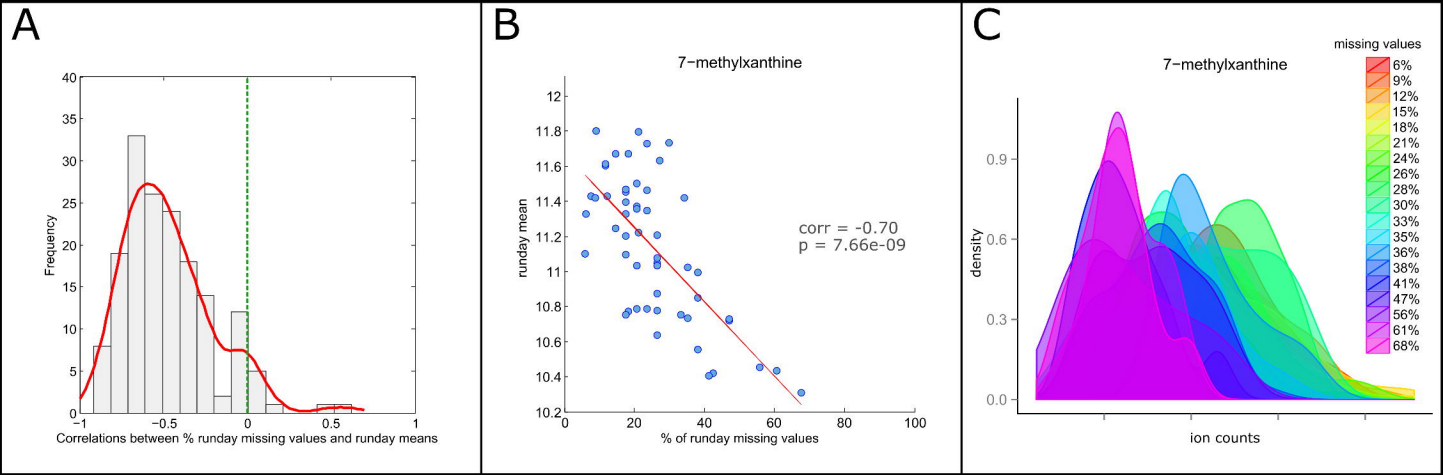
File S10. KORA F4 experimental setup.

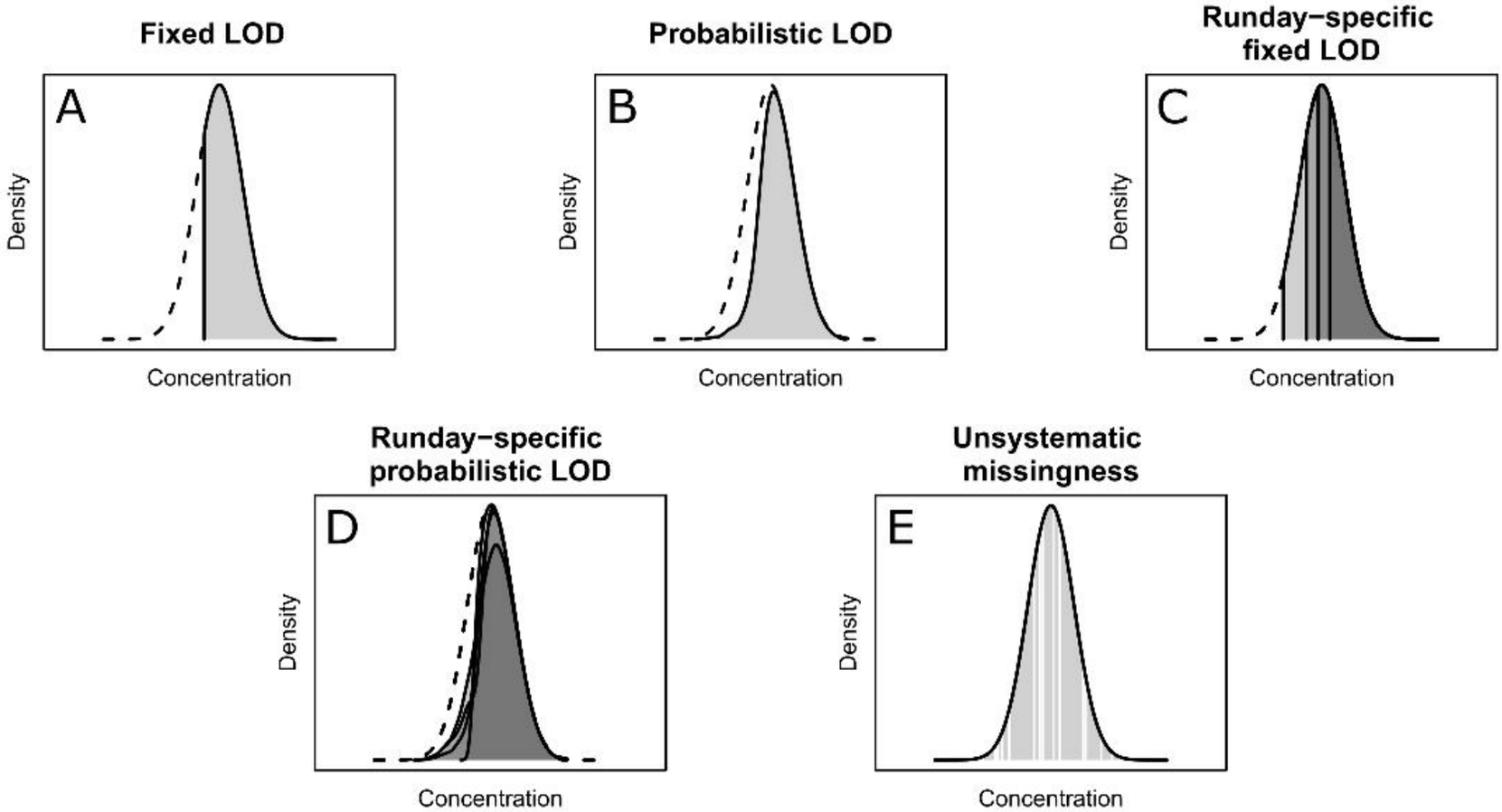
File S11. KNN-obs-sel and MICE imputation code.

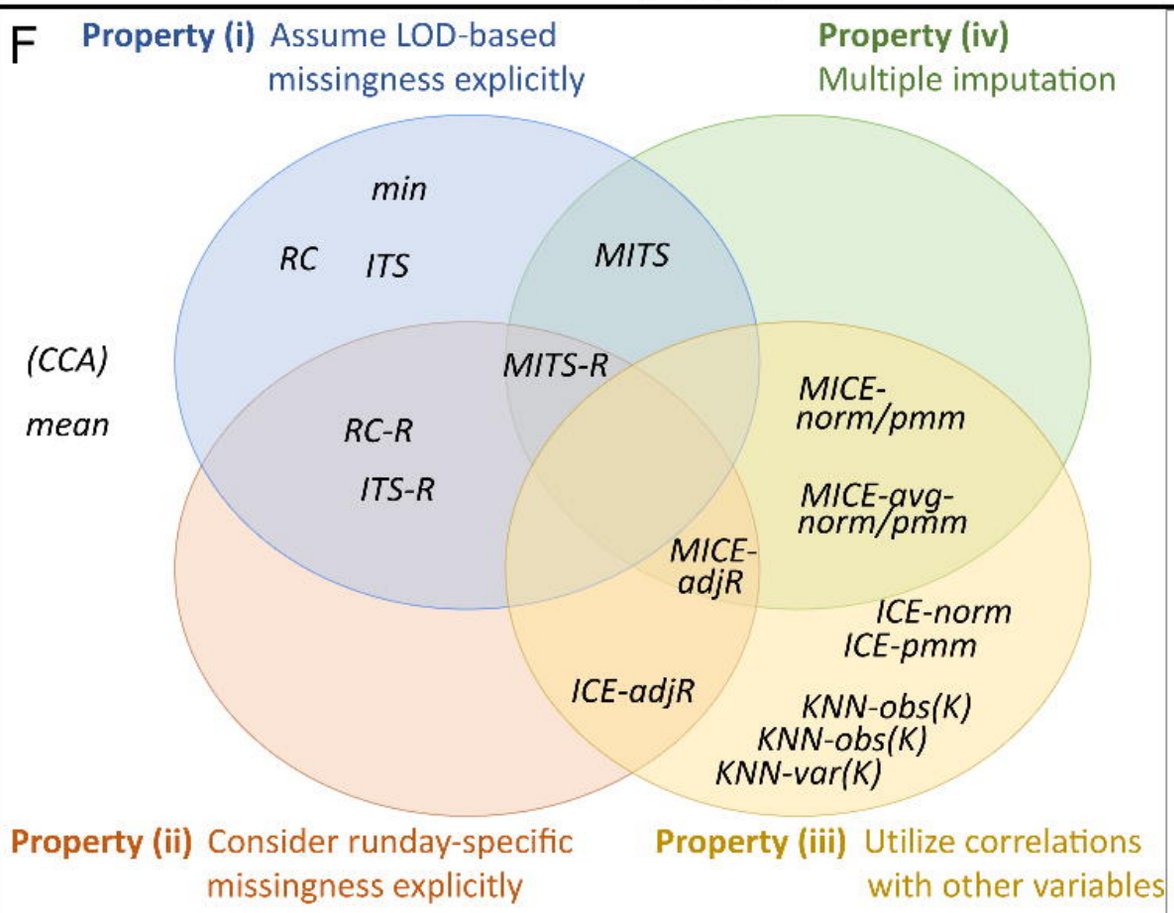

KORA F4
n = 1768, p = 516
LC/MS & GC/MS untargeted metabolomics data
Preprocessing
n = 1757, p = 516
Description
Overall missingness
Limit of detection (LOD)
Runday-specific missingness
Mechanisms
LOD
Probabilistic LOD
Runday-specific LOD
Unsystematic missingness
Simulation
Parameters include:
- Number of variables
- Strength of correlations
- Number of samples
- Percentage of missing values
- Missingness mechanism
Analyses on real data
Analyses within simulation framework
Imputation
Missing value handling approaches:
- Complete case analysis (CCA)
- Imputation approaches that cover
* assuming LOD-based missingness
* considering runday-specific missingness
* performing multiple imputations
* utilizing correlations with other variables
Simulation-based Evaluation
Linear & Logistic regression
Pearson & Partial correlation
Real data-based Evaluation
Pathway modularity
Metabolite-SNP associations


A
Histogram of missing values in metabolites
# of metabolites
% missing values
B
Histogram of missing values in samples
# of samples
% missing values
C
Missing values of 7−methylxanthine
in 3−methylxanthine
3−methylxanthine
7−methylxanthine (26% missings)
missing
corr = 0.77
p = 7.85e−211
Concentrations of 3-methylxanthine
in missing and observed 7-methylxanthine
p_Wrst = 2.29e−13
3−methylxanthine
missing
observed
7−methylxanthine
D
Missing values of 1−arachidonoylglycerophosphocholine
in 1−docosahexaenoylglycerophosphocholine
1−docosahexaenoylglycerophosphocholine*
1−arachidonoylglycerophosphocholine (9% missings)
missing
corr = 0.74
p = 4.68e−276
Concentrations of 1-docosahexaenoylglycerophosphocholine
in missing and observed 1-arachidonoylglycerophosphocholine
p_Wrst = 1.51e−03
1−docosahexaenoylglycerophosphocholine*
missing
observed
1−arachidonoylglycerophosphocholine

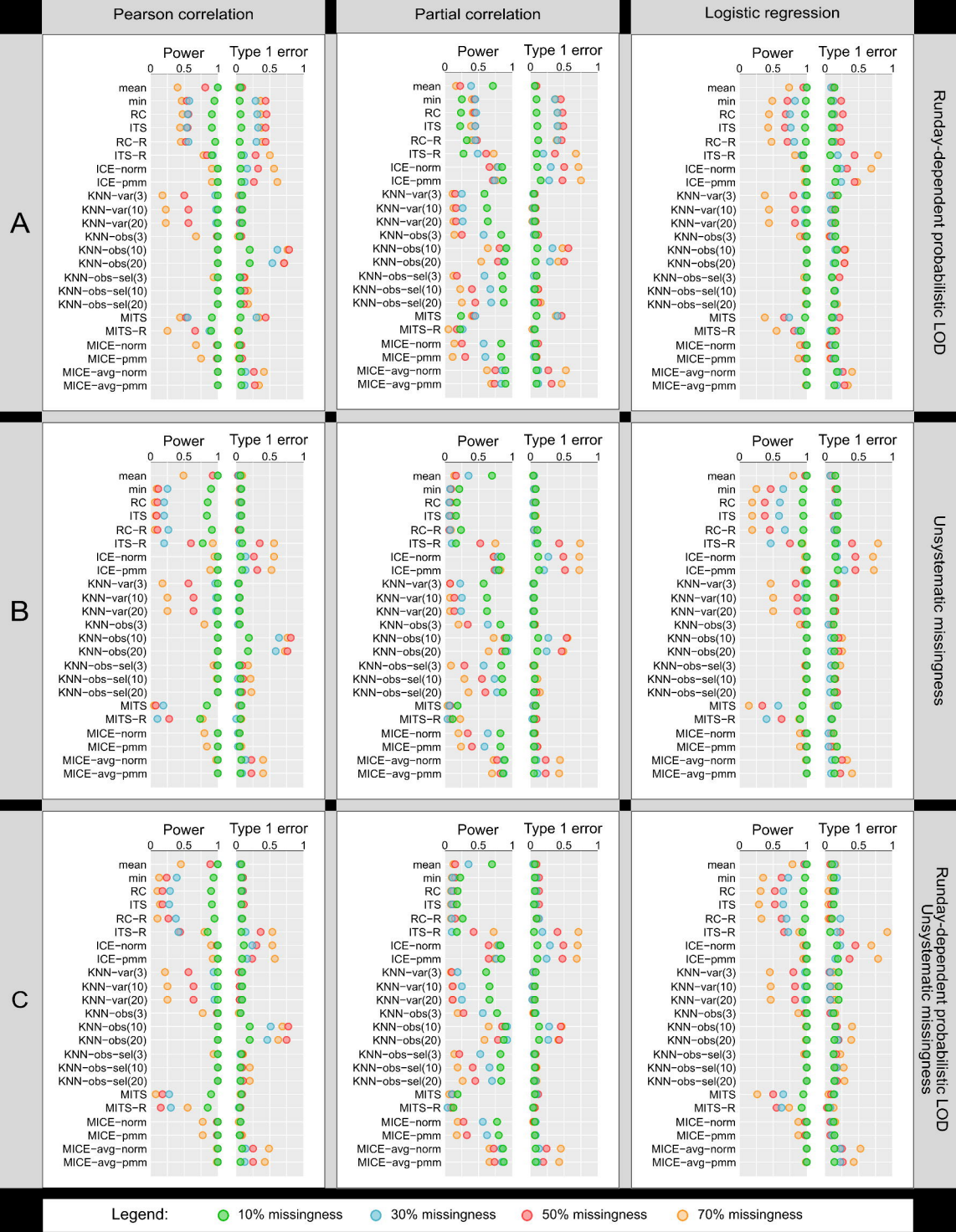

A
LC/MS+
LC/MS−
GC/MS
normalized runday missingness
rundays
r = 0.36
p = 0.008
r = -0.09
p = 0.51
r = -0.03
p = 0.82
B
Variation of missingness across rundays
total missing values (%)
LC/MS-
LC/MS+
GC/MS
D
C
C
ursodeoxycholate
Frequency
runday missing values (%)
D
gamma−glutamylisoleucine
Frequency
runday missing values (%)

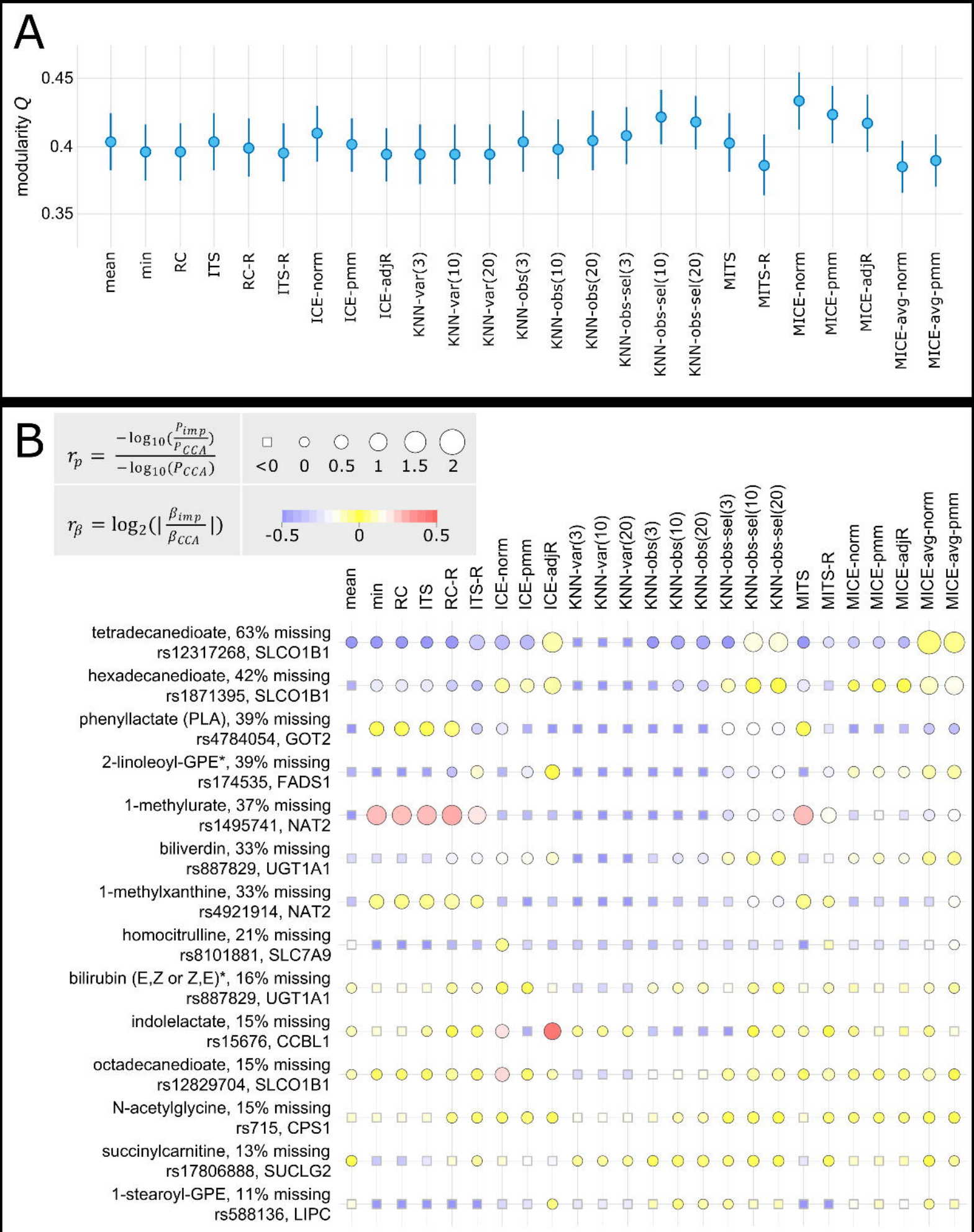

A
Frequency
Correlations between % runday missing values and runday means
B
7−methylxanthine
runday mean
% of runday missing values
corr = -0.70
p = 7.66e-09
C
7−methylxanthine
missing values
6%
9%
12%
15%
18%
21%
24%
26%
28%
30%
33%
35%
36%
38%
41%
47%
56%
61%
68%
density
ion counts





CCA	Complete case analysis
min	Minimum imputation
RC	Richardson & Ciampi
ITS	Imputation by truncated sampling
MITS	Muliple ITS
RC-R	RC within rundays
ITS-R	ITS within rundays
MITS-R	MITS within rundays
mean	Mean imputation
ICE-norm	Imputation by chained equations using Bayesian regression imputation
ICE-pmm	ICE using predictive mean matching
ICE-adjR	ICE with random runday intercept
MICE-norm	Multiple ICE-norm, pooling statistics
MICE-pmm	Multiple ICE-pmm, pooling statistics
MICE-avg-norm	Multiple ICE-norm, pooling data
MICE-avg-pmm	Multiple ICE-pmm, pooling data
MICE-adjR	Multiple ICE-adjR
KNN-var(K)	*K*-nearest neighbor imputation per variable
KNN-obs(K)	KNN per observation
KNN-obs-sel(K)	KNN per observation using selected variables


Pearson correlation
Partial correlation
Logistic regression
A
B
C
Runday-dependent probabilistic LOD
Unsystematic missingness
Runday-dependent probabilistic LOD
Unsystematic missingness
Power
Type 1 error
0
0.5
1
mean
min
RC
ITS
RC−R
ITS−R
ICE−norm
ICE−pmm
KNN−var(3)
KNN−var(10)
KNN−var(20)
KNN−obs(3)
KNN−obs(10)
KNN−obs(20)
KNN−obs−sel(3)
KNN−obs−sel(10)
KNN−obs−sel(20)
MITS
MITS−R
MICE−norm
MICE−pmm
MICE−avg−norm
MICE−avg−pmm
Legend:
10% missingness
30% missingness
50% missingness
70% missingness


A
modularity Q
0.45
0.4
0.35
mean
min
RC
ITS
RC-R
ITS-R
ICE-norm
ICE-pmm
ICE-adjR
KNN-var(3)
KNN-var(10)
KNN-var(20)
KNN-obs(3)
KNN-obs(10)
KNN-obs(20)
KNN-obs-sel(3)
KNN-obs-sel(10)
KNN-obs-sel(20)
MITS
MITS-R
MICE-norm
MICE-pmm
MICE-adjR
MICE-avg-norm
MICE-avg-pmm
B
$r_p = \frac{-\log_{10}(\frac{p_{imp}}{p_{CCA}})}{-\log_{10}(p_{CCA})}$
<0
0
0.5
1
1.5
2
$r_\beta = \log_2(|\frac{\beta_{imp}}{\beta_{CCA}}|)$
-0.5
0
0.5
tetradecanedioate, 63% missing
rs12317268, SLCO1B1
hexadecanedioate, 42% missing
rs1871395, SLCO1B1
phenyllactate (PLA), 39% missing
rs4784054, GOT2
2-linoleoyl-GPE*, 39% missing
rs174535, FADS1
1-methylurate, 37% missing
rs1495741, NAT2
biliverdin, 33% missing
rs887829, UGT1A1
1-methylxanthine, 33% missing
rs4921914, NAT2
homocitrulline, 21% missing
rs8101881, SLC7A9
bilirubin (E,Z or Z,E)*, 16% missing
rs887829, UGT1A1
indolelactate, 15% missing
rs15676, CCBL1
octadecanedioate, 15% missing
rs12829704, SLCO1B1
N-acetylglycine, 15% missing
rs715, CPS1
succinylcarnitine, 13% missing
rs17806888, SUCLG2
1-stearoyl-GPE, 11% missing
rs588136, LIPC