







Article

# Metabolomics Analytics Workflow for Epidemiological Research: Perspectives from the Consortium of Metabolomics Studies (COMETS) †

Mary C. Playdon <sup>1,2,\*</sup>, Amit D. Joshi <sup>3,4,5</sup>, Fred K. Tabung <sup>6,7,8</sup> , Susan Cheng <sup>9</sup>, Mir Henglin <sup>10</sup>, Andy Kim <sup>10</sup>, Tengda Lin <sup>2,11</sup>, Eline H. van Roekel <sup>12</sup> , Jiaqi Huang <sup>13</sup>, Jan Krumsiek <sup>14</sup>, Ying Wang <sup>15</sup>, Ewy Mathé <sup>16</sup> , Marinella Temprosa <sup>17</sup>, Steven Moore <sup>13</sup>, Bo Chawes <sup>18</sup> , A. Heather Eliassen <sup>19,20</sup>, Andrea Gsur <sup>21</sup>, Marc J. Gunter <sup>22</sup>, Sei Harada <sup>23</sup>, Claudia Langenberg <sup>24,25</sup>, Matej Oresic <sup>26,27</sup> , Wei Perng <sup>28,29</sup>, Wei Jie Seow <sup>30,31</sup> and Oana A. Zeleznik <sup>19</sup> 

- <sup>1</sup> Department of Nutrition and Integrative Physiology, College of Health, University of Utah, Salt Lake City, UT 84112, USA
- <sup>2</sup> Division of Cancer Population Sciences, Huntsman Cancer Institute, Salt Lake City, UT 84112, USA
- <sup>3</sup> Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, MA 02114, USA
- <sup>4</sup> Division of Gastroenterology, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
- <sup>5</sup> Program in Genetic Epidemiology and Statistical Genetics, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA
- <sup>6</sup> Division of Medical Oncology, Department of Internal Medicine, The Ohio State University College of Medicine, Columbus, OH 43210, USA
- <sup>7</sup> The Ohio State University Comprehensive Cancer Center, Arthur G. James Cancer Hospital and Richard J. Solove Research Institute, Columbus, OH 43210, USA
- <sup>8</sup> Division of Epidemiology, The Ohio State University College of Public Health, Columbus, OH 43210, USA
- <sup>9</sup> Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
- <sup>10</sup> Cardiovascular Division, Brigham and Women's Hospital, Boston, MA 02115, USA
- <sup>11</sup> Department of Population Health Sciences, School of Medicine, University of Utah, Salt Lake City, UT 84112, USA
- <sup>12</sup> Department of Epidemiology, GROW School for Oncology and Developmental Biology, Maastricht University, 6200 MD Maastricht, The Netherlands
- <sup>13</sup> Division of Cancer Epidemiology and Genetics, Metabolic Epidemiology Branch, National Cancer Institute, Rockville, MD 20850, USA
- <sup>14</sup> Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10021, USA
- <sup>15</sup> Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA 30303, USA
- <sup>16</sup> College of Medicine, Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA
- <sup>17</sup> Department of Epidemiology and Biostatistics, Milken Institute School of Public Health, George Washington University, Washington, DC 20052, USA
- <sup>18</sup> COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, 1165 Copenhagen, Denmark
- <sup>19</sup> Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
- <sup>20</sup> Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
- <sup>21</sup> Institute of Cancer Research, Department of Medicine, Medical University of Vienna, 1090 Vienna, Austria
- <sup>22</sup> Section of Nutrition and Metabolism, International Agency for Research on Cancer, World Health Organization, 69008 Lyon, France
- <sup>23</sup> Department of Preventive Medicine and Public Health, Keio University School of Medicine, Tokyo 160-8582, Japan
- <sup>24</sup> MRC Epidemiology Unit, Public Health, University of Cambridge, Cambridge CB2 1 TN, UK
- <sup>25</sup> The Francis Crick Institute, London NW1 1ST, UK

- <sup>26</sup> Turku Centre for Biotechnology, University of Turku, 20500 Turku, Finland
- <sup>27</sup> School of Medical Sciences, Örebro University, 702 81 Örebro, Sweden
- <sup>28</sup> Department of Epidemiology, Colorado School of Public Health, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO 80045, USA
- <sup>29</sup> Life course epidemiology of adiposity and diabetes (LEAD) Center, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO 80045, USA
- <sup>30</sup> Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore 117549, Singapore
- <sup>31</sup> Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore 119228, Singapore
- \* Correspondence: mary.playdon@hci.utah.edu; Tel.: +801-213-6264
- † Disclaimer: Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Received: 10 June 2019; Accepted: 4 July 2019; Published: 17 July 2019



**Abstract:** The application of metabolomics technology to epidemiological studies is emerging as a new approach to elucidate disease etiology and for biomarker discovery. However, analysis of metabolomics data is complex and there is an urgent need for the standardization of analysis workflow and reporting of study findings. To inform the development of such guidelines, we conducted a survey of 47 cohort representatives from the Consortium of Metabolomics Studies (COMETS) to gain insights into the current strategies and procedures used for analyzing metabolomics data in epidemiological studies worldwide. The results indicated a variety of applied analytical strategies, from biospecimen and data pre-processing and quality control to statistical analysis and reporting of study findings. These strategies included methods commonly used within the metabolomics community and applied in epidemiological research, as well as novel approaches to pre-processing pipelines and data analysis. To help with these discrepancies, we propose use of open-source initiatives such as the online web-based tool COMETS Analytics, which includes helpful tools to guide analytical workflow and the standardized reporting of findings from metabolomics analyses within epidemiological studies. Ultimately, this will improve the quality of statistical analyses, research findings, and study reproducibility.

**Keywords:** metabolomics; epidemiology; statistical analysis; reporting; analytical methods; data analysis; pre-processing

---

## 1. Introduction

Recent advances in high-throughput methods to characterize the human metabolome present an unprecedented opportunity to strengthen epidemiological research and broaden its scope. Metabolomics is being utilized to shed light on disease etiology through objective biomarkers of exposures that are otherwise fraught with measurement error; to refine or complement our current methods of phenotypic assessment; to understand biological pathways linking exposures to health outcomes; identify early onset disease; and subtype diseases with heterogeneous etiologies [1]. Metabolomics is the comprehensive characterization of small molecules present in biospecimens such as plasma, urine, and stool. Since these small molecules reflect influences from environmental factors, as well as endogenous factors such as genetics, epigenetics, transcription, protein structure and function, and gut microbiota, metabolomics has the potential to provide a more nuanced assessment of physiology (or pathophysiology) that is often unachievable with traditional epidemiological approaches such as evaluation of single biomarkers, or self-reported data collected via questionnaires. Diverse research

fields including epidemiology, systems biology, biochemistry, microbiology, pharmacology, toxicology, clinical science, and biostatistics converge through metabolomics to advance a multidisciplinary understanding of health and disease [2].

To date, metabolomics has shown success in screening newborns for inborn errors of metabolism, for identifying candidate biomarkers for early disease detection, particularly for some diseases like diabetes and cancer [3,4], for understanding disease mechanisms [5], and has been used to develop better measures of disease risk factors like smoking, diet, and obesity [6–8]. As this emerging technology is increasingly incorporated into disease research, including epidemiological studies, a bottleneck to advancing the field is the complexity and lack of standard protocols or best practices for analyzing metabolomics data [9]. Indeed, metabolomics has a unique data structure that depends on the platform (e.g., *targeted* quantification of defined groups of chemically characterized and biochemically annotated metabolites or *untargeted* semi-quantified analyses of all measurable analytes), which determines the analytical strategy to be taken. Challenges in the analysis of metabolomics data are multi-fold [10], including workflow choices for data harmonization, pre-processing (alignment, filtering), metabolite identification/annotation, data preparation (centering, scaling, and transformation) [11], imputation, and statistical approaches [12]. Moreover, since there is a high degree of collinearity between metabolites according to biochemical pathway, considering the pattern of metabolite values in addition to individual metabolites can create additional statistical obstacles. Metabolomics data management and data analysis consist of a series of complex steps that can be performed in many ways with no defined order, and some of these are optional depending on the study aims. This complexity is further compounded by the lack of adherence to a set of standard reporting guidelines [13], which makes it difficult to determine common or best practices, and leads to problems in replicating results, comparing findings, and conducting systematic reviews and meta-analyses. The purpose of this study was to summarize current practices of investigators participating in the international Consortium of Metabolomics Studies (COMETS) [14] in the analysis of metabolomics data from epidemiological studies.

## 2. Results

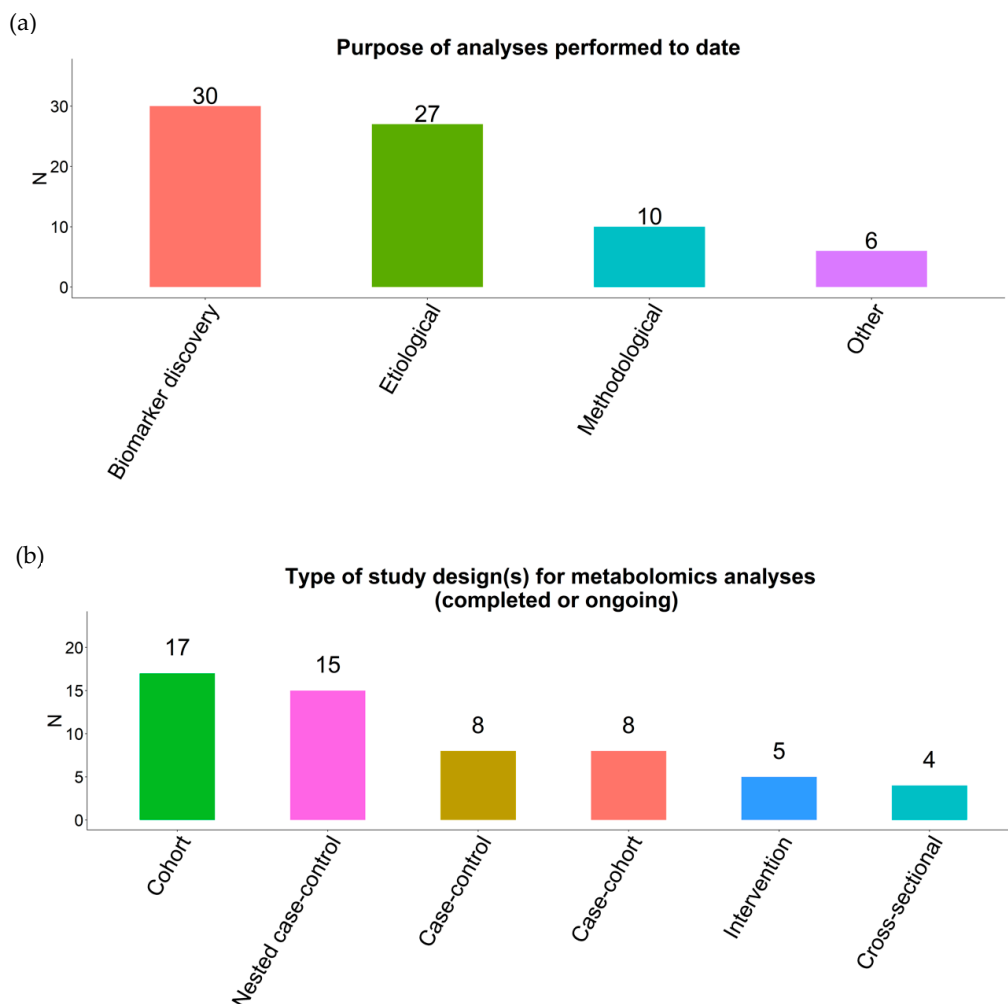
### 2.1. Response Rate

Thirty-three out of 47 (70%) of the participating COMETS cohorts responded to an online questionnaire up to October 2018 (See Supplementary Table S1 for a summary of all results). The questionnaire inquired about current practices in the preparation, analysis, and reporting of metabolomics data. The total number of respondents was used as the denominator in calculating response rate (%). Respondents could check multiple options for each question (all that apply), and follow-up questions were asked based on some responses. When questions were unanswered, the response was denoted as “missing”. Most respondents were Principal Investigators (42%) followed by postdoctoral fellows (16%), research analysts (13%), research scientists (8%), biostatisticians (5%), and PhD students (3%). Respondents reported having conducted a median of 6 (range 1–30) analyses of metabolomics data. Multiple analyses were conducted on the same datasets, with different analysis goals.

### 2.2. Datasets

Metabolomics data were derived predominantly from untargeted metabolomics platforms (55%), with 27% derived from combining untargeted and targeted platforms and 18% from targeted only. The analysis goals were largely for biomarker discovery (91%) and/or to investigate disease etiology (82%), methodology development (31%), and other purposes (18%; e.g., metabolome-wide association study) (Figure 1a). Different study designs were used to generate metabolomics data, including: 17 cohorts, 15 nested case-control studies, 8 case-control studies, 8 case-cohorts, 5 randomized trials, and 4 cross-sectional analyses within prospective studies, with an average of 3302 participants (standard

deviation 3972) (Figure 1b). Reported outcomes being analyzed included cancer (39%), cardiovascular disease (CVD) (30%), diabetes (21%), and pregnancy outcomes (6%). Human-immunodeficiency virus (HIV) infection, cardiometabolic measures, asthma, and amyotrophic lateral sclerosis (ALS) were each reported as outcomes by one respondent.



**Figure 1.** Description of study purpose (a) and study design (b) of participating Consortium of Metabolomics Studies (COMETS) cohorts.

### 2.3. Power Calculations

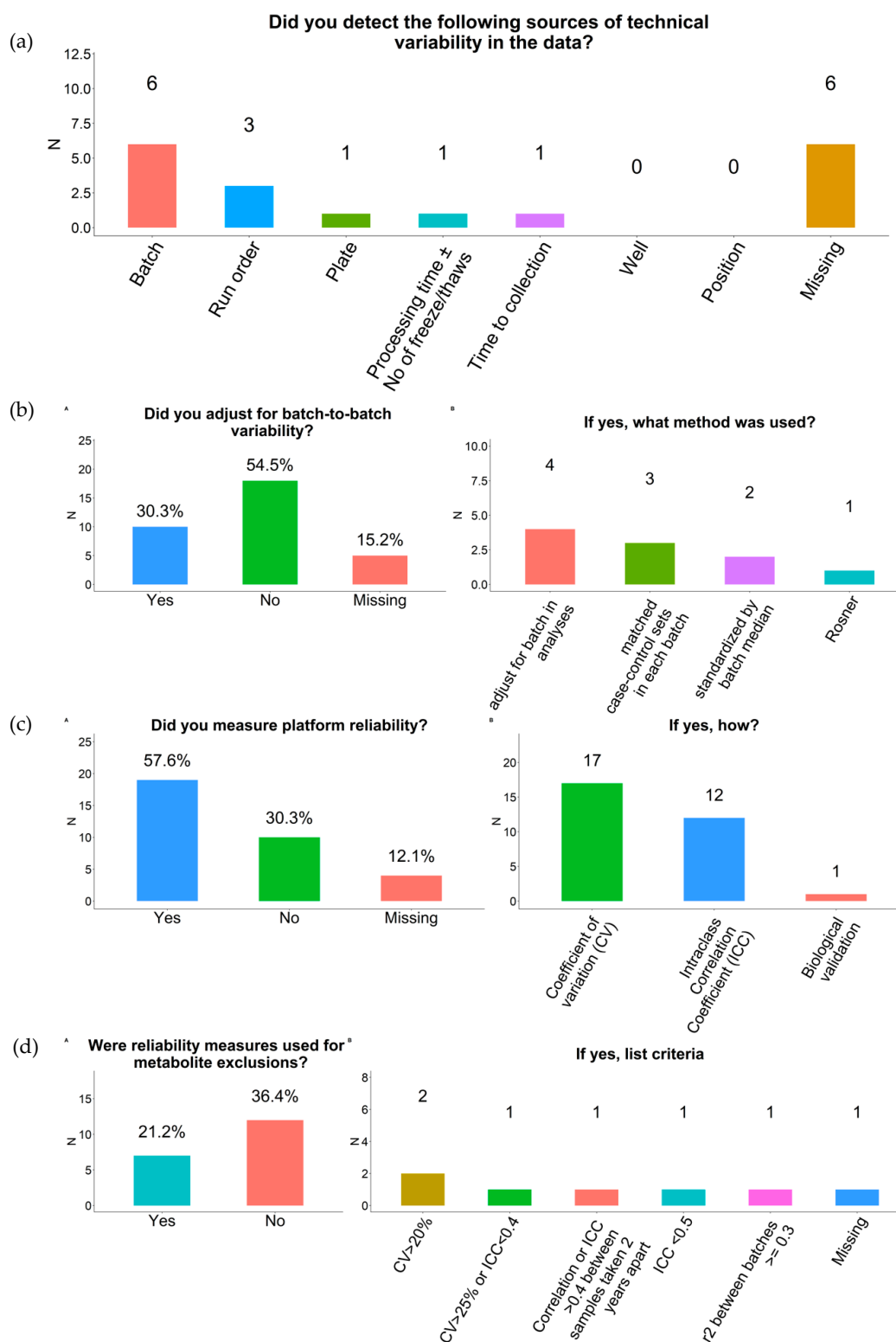
Approximately 45% of respondents anticipated power or sample size prior to analysis. Of those that did sample size or power calculations, most were performed in Quanto ( $n = 6$ ) or R ( $n = 4$ ). Other resources used were Power V3 ( $n = 2$ ), PASS ( $n = 1$ ), and GPower ( $n = 1$ ).

### 2.4. Outliers and Technical Variability

Extreme metabolite values (i.e., outliers) were evaluated by 39% of respondents, predominantly by using principal component analysis (PCA;  $n = 13$ ). A subset used principal component partial R-square (PC-PR2) ( $n = 2$ ) or analysis of variance (ANOVA) ( $n = 2$ ) to identify outliers.

Of those that evaluated sources of metabolite variability, reported sources included batch effects ( $n = 6$ ), run order ( $n = 3$ ), plate, time to sample collection, and time from sample collection to freezing (all  $n = 1$ ) (Figure 2a). However, most respondents did not exclude metabolites based on these sources of variability. Ten respondents reported adjusting for batch-to-batch variability (Figure 2b).

Methods included adding case-control sets to each batch (n = 3), adjusting for batch in analysis (n = 4), standardizing metabolites to the batch median (n = 2), or using the Rosner approach [15] (n = 1).



**Figure 2.** Reliability measures among participating COMETS cohorts. (a) Sources of technical variability; (b) Batch-to-batch variability; (c) Platform reliability; (d) Metabolite exclusion criteria. Missing refers to the proportion of respondents that did not answer the question.

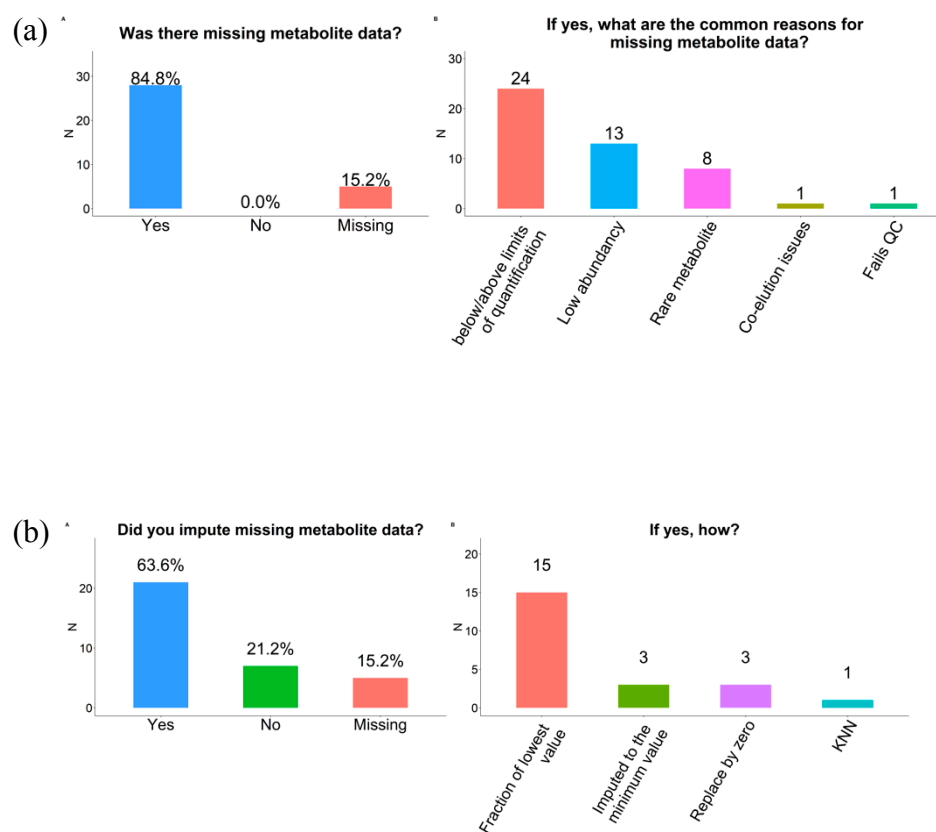
Nineteen respondents measured platform reliability using coefficient of variation (CV) ( $n = 17$ ) and/or intraclass correlation coefficient (ICC) ( $n = 12$ ) (Figure 2c). The range of CVs reported for completed studies was 0–78% (most up to 20%), while the range of ICCs reported was up to 1.0 (most  $> 0.90$ ). Seven respondents reported using this information to exclude metabolites from analysis, but criteria for exclusion were variable (e.g.,  $CV > 20\%$  or  $ICC < 0.40$ ) (Figure 2d).

### 2.5. Data Preparation

Centering, scaling, and data transformation are data preparation methods used in metabolomics studies [11]. Thirty-nine percent of respondents reported centering individual metabolite values while 42% did not (15% missing). The most common method was centering to the mean ( $n = 10$ ). Scaling methods included Pareto-scaling ( $n = 2$ ), auto-scaling/z-transformation/standard-deviation scaling ( $n = 7$ ), probit-score scaling ( $n = 1$ ), and median absolute deviation (MAD) ( $n = 1$ ). Most respondents reported transforming metabolite data (85%; 15% missing), including by log-transformation ( $n = 21$ ).

### 2.6. Missing Data

Most respondents (85%) reported that their metabolomics data had missing values. Missingness was due to the limit of detection (LOD)/quantification of the platform ( $n = 24$ ), low abundance ( $n = 13$ ), and rare metabolites ( $n = 8$ ). Co-elution issues and failed quality control (QC) were also reported ( $n = 1$  each) (Figure 3a).



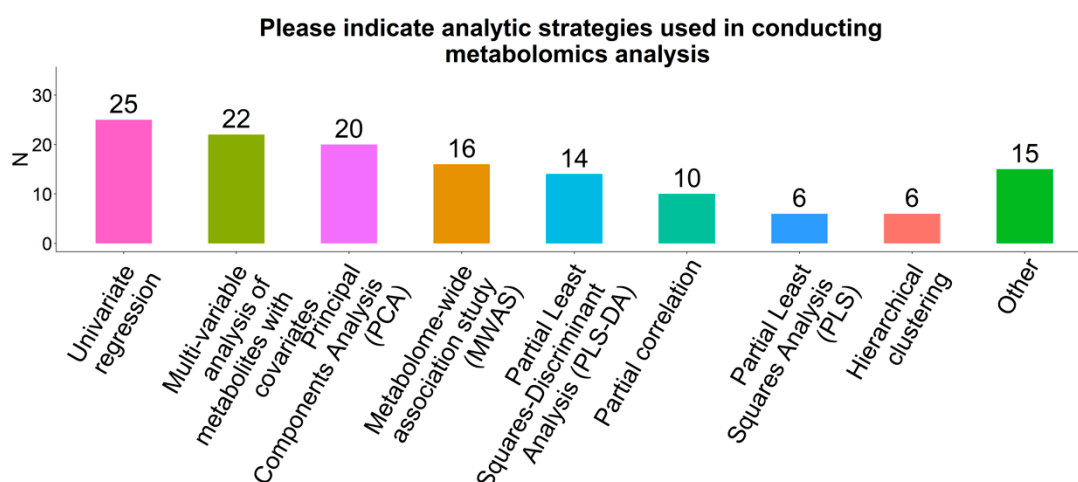
**Figure 3.** Data pre-processing steps conducted among participating COMETS cohorts. (a) Missingness; (b) Imputation of missing values.

Twenty-one respondents reported imputing the missing values while seven did not ( $n = 5$  did not answer). The most common approach to treat missing values was to replace these by a fraction of the lowest value ( $n = 15$ ). Replacing missing values by zero or by the minimum value were each reported by three respondents. K-nearest neighbor imputation (KNN) [16] was used by one respondent while

none used multiple imputation by chained equations (MICE) or Markov chain Monte Carlo (MCMC) (Figure 3b). Most respondents ( $n = 15$ ) reported excluding metabolites with a percent of missingness above a certain threshold (median 50%; range 5% to 90%). Dichotomization, categorization as missing or min-median or median-max, imputation to the mean, flagging, and complete exclusion were each reported by one respondent.

### 2.7. Statistical Analysis Methods

Respondents used multiple statistical analysis strategies to analyze metabolomics data (Figure 4). The most common strategy was univariate regression (e.g., linear regression of a single exposure on a single metabolite) (76%) followed by multiple/multivariable analysis of metabolites on the exposure of interest, with adjustment for covariates (67%), principal component analysis (PCA, 61%), metabolome-wide association study (MWAS, 48%), partial least squares-discriminant analysis (PLS-DA, 42%), partial correlation (30%), partial least squares analysis (PLS, 18%), and hierarchical clustering (18%). The following analysis techniques were each reported by one respondent: canonical correspondence analysis (CCA), treelet transform, K-means clustering, least absolute shrinkage and selection operator regression (LASSO), supervised gradient descent, random forest, support vector machines (SVM), weighted gene co-expression network analysis (WGCNA), metabolite set enrichment analysis, over-representation analysis, differential networks, hierarchical cluster analysis, Bayesian non-parametric methods, orthogonal projections to latent structures discriminant analysis (OPLS-DA) and generalized linear mixed models (GLM). A third of respondents reported having used variable selection methods incorporating penalization, including LASSO, SVM, and sparse seemingly unrelated regression (SUR). Mediation analysis was conducted by 15% of respondents.



**Figure 4.** Analytic strategies employed for metabolomics data among participating COMETS cohorts.

Almost half of the respondents (48%) assessed the performance of biomarker classification using area under the receiver operator characteristic curve (AUC) ( $n = 16$ ), net reclassification improvement ( $n = 2$ ), sensitivity/specificity/positive predictive value/ negative predictive value ( $n = 1$ ) and PLS-DA ( $n = 1$ ).

Approximately 40% of respondents measured metabolite intercorrelations. The most common method for assessing metabolite intercorrelations was partial correlation ( $n = 11$ ). Gaussian graphical modeling (GGM;  $n = 2$ ) and WGCNA ( $n = 1$ ) were also used for this purpose.

A quarter of respondents conducted network analyses. WGCNA ( $n = 3$ ) and unspecified methods incorporated into programs within the analytic resource MetaboAnalyst ( $n = 2$ ) were the most common approaches followed by GGM with linkage to biological pathway, BayesNet, Gene-Set Enrichment Analysis (GSEA), over representation analysis (ORA), Metscape, and yED graphical networking software ( $n = 1$  each).

## 2.8. Cross Validation and External Validation

Seven respondents (21%) performed cross-validation (CV) analysis. Six respondents used k-fold CV and one reported simulation/permutation of data. The proportion of training data varied (60% to 90%) as did the proportion of testing data (10% to 40%). Five respondents used bootstrapping, and 11 externally validated findings in another cohort.

## 2.9. Visualization

Respondents visualized results using heat maps (n = 17), volcano plots (n = 6), Manhattan plots (n = 5), forest plots (n = 2), and individual approaches (n = 1).

## 2.10. Multiple Testing Correction

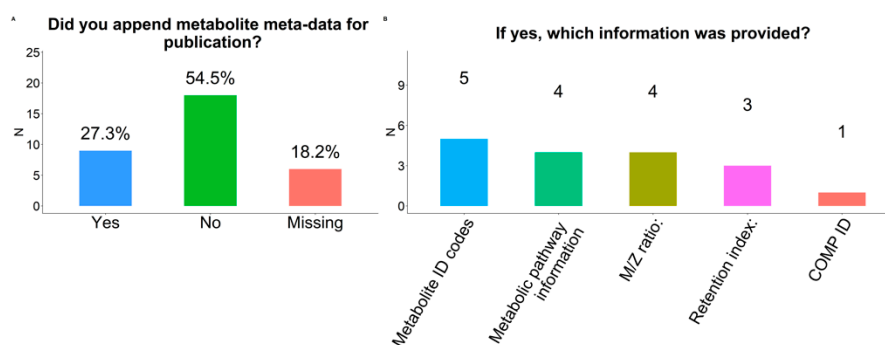
Correcting for testing multiple hypotheses following analysis of metabolomics data (e.g., when regressing multiple metabolites on an exposure, separately) was done by 79% of respondents (Figure 5). The Benjamini-Hochberg false discovery rate (FDR) was the most common approach (n = 22), followed by Bonferroni correction (n = 12), Bonferroni-Holm, Dunn-Sidak, and permutation tests (n = 1 each).



**Figure 5.** Strategies for correcting for multiple hypothesis testing among participating COMETS cohorts. (a) Use of multiple testing correction (yes/no); (b) Methods for correcting for multiple hypothesis tests.

## 2.11. Meta-Data

In total, 9 of 33 respondents (27%) appended metabolite meta-data for publication (Figure 6). Unique identifiers such as those from publicly available metabolomics databases such as Human Metabolome Database (HMDB) and PUBCHEM were most often appended to metabolomics study results (n = 5). The addition of pathway information and mass-to-charge ratio (m/z ratio) were also reported (n = 4) along with retention time (n = 3) and internal database compound tracking number in the platform's chemical library (COMP ID) (n = 1).

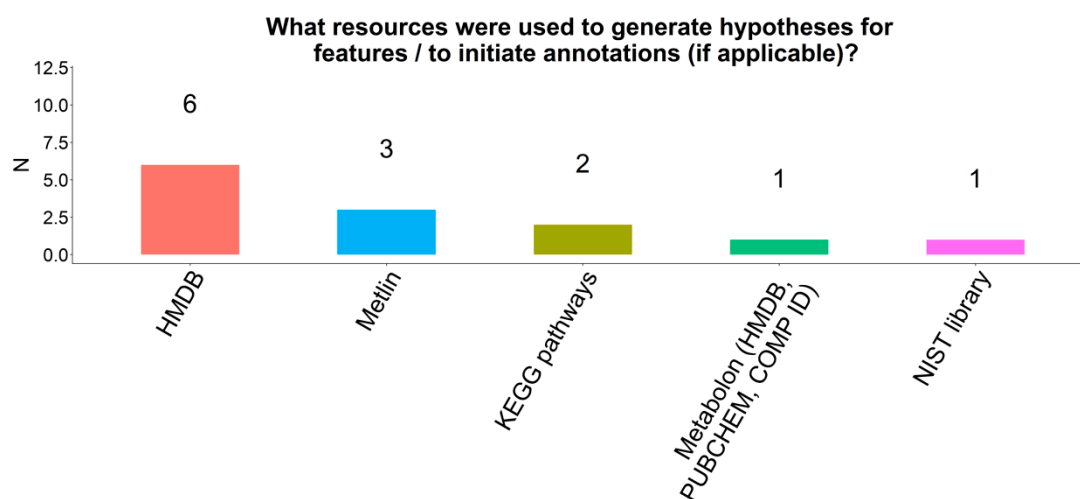


**Figure 6.** Appending metabolite meta-data in publications of findings from participating COMETS cohorts. (a) Include meta-data for publication (yes/no); (b) Information provided in appended meta-data.



### 2.12. Annotations

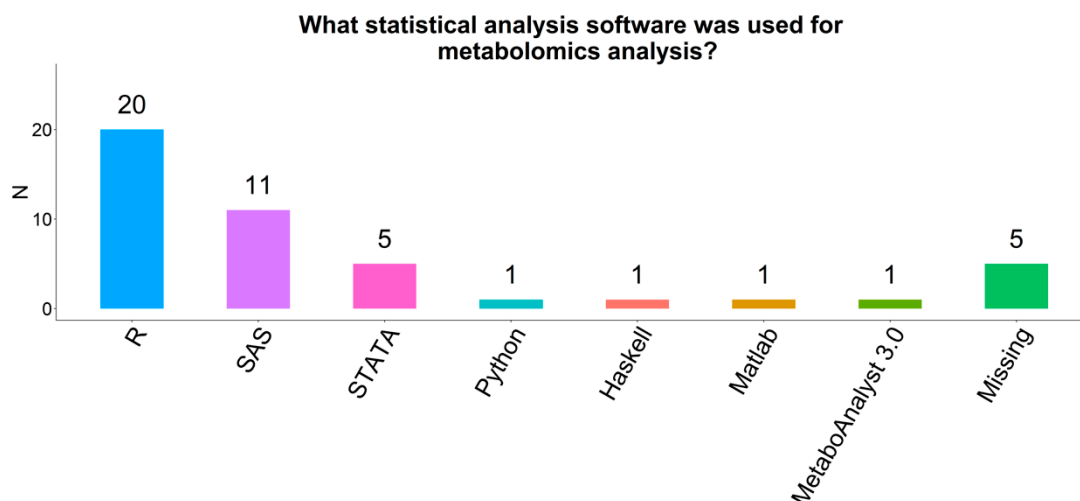
Eight respondents (24%) annotated metabolites. Of those, databases that were used for annotation included HMDB (n = 6), Metlin (n = 3), KEGG pathways (n = 2), Metabolon proprietary system (n = 1), and the National Institute for Standards and Technology (NIST) library (n = 1) (Figure 7).



**Figure 7.** Strategies for metabolite annotation among participating COMETS cohorts.

### 2.13. Coding Language

Most respondents (61%) wrote their statistical code in R (Figure 8). Other popular coding languages included SAS (33%) and STATA (15%). Fewer respondents used other languages (e.g., Python, Haskell, Matlab; n = 1 each). Four respondents (19%) used a statistical coding style guide (e.g., tidyverse, Google's R style guide) in the design of their code.



**Figure 8.** Statistical analysis software used by participating COMETS cohorts.

### 2.14. Software

Of the numerous open source software packages available online (for examples, see Table 1), none were leveraged by respondents to this survey. Rather, they reported writing original code or using R packages for analysis of metabolomics data.

### 2.15. Minimum Reporting Standards

There were many suggested minimum reporting standards for analysis of metabolomics data, including reporting: study aim and objectives; study hypothesis; statistical assumptions; overall analytical strategy; metabolomics data standard operating procedures; pre-analytical processing measures; analytical platform; quality control measures used; statistical packages and software used; strategy for adjustment for multiple comparisons; meta-data; effect estimates and confidence (betas, confidence intervals, P-values and Q-values for both nominal and statistically adjusted results); confounder selection; cross-validation strategy; external validation strategy; and providing statistical code for replication.

## 3. Discussion

The application of metabolomics in epidemiological studies has increased dramatically in recent years [17]. The COMETS consortium is currently collaborating on large-scale replication studies and meta-analyses of metabolomics data [14]. With an overall study population in excess of 130,000 participants, median age 51 years (range 0–100 years) representing European, Asian, African, Hispanic, native Hawaiian and other mixed populations, COMETS is a rich resource for addressing research questions that leverage blood metabolomics data. Given that application of metabolomics technology to epidemiological studies is an emerging field, we conducted a survey of participating COMETS cohort representatives to gain insights into the strategies and procedures used for analyzing metabolomics data, including data pre-processing, analysis, and reporting of results. With a range of experience levels analyzing metabolomics data, participating COMETS cohorts have analyzed and published on both targeted and untargeted metabolomics data from a variety of study designs in relation to disease risk factors [8,18–27] and many disease outcomes, predominantly cancer, CVD, and diabetes [28–40]. However, there was little consensus on approaches to data pre-processing, statistical analysis or reporting of results, which is echoed in the broader metabolomics community [13].

Metabolomics studies have predominantly investigated disease etiology or biological mechanisms underlying progression and for biomarker discovery, using metabolomics data generated on both targeted or untargeted platforms in the context of a variety of epidemiological study designs [23,40–43]. The type of study design selected is typically related to the research question of interest in addition to the availability of metabolomics data and the frequency of the outcomes of interest in that same study. Importantly, statistical analytical considerations will differ depending on study design. For instance, a nested case-control study may require propensity-score matching to avoid ascertainment bias, by controlling for the probability to be included in the metabolomics data based on eligibility criteria [38]. By contrast, large cohort studies that investigate rarer outcomes may utilize multivariable regression or tree-based analysis. Furthermore, data reduction approaches may be needed for high-dimensional untargeted metabolomics data.

### 3.1. Data Pre-Processing

A first step in analysis of metabolomics data includes data pre-processing. Extreme metabolite values are frequently observed even after applying any one of a variety of normalization methods to large metabolomics datasets [10], yet less than half of respondents evaluated them. A variety of sources drive extreme metabolite values including pre-analytical conditions (e.g., processing delay), batch variability (due to a variety of technical factors), misalignment (between or within batches), chemical instability (typically manifesting as within-batch variation), true biological variation that may arise from rare genetic determinants or rare exposures, other random effects that are not easily classifiable, or some combination of the above [44]. Filtering or censoring extreme values can reduce the skewness of a distribution and stabilize metabolite variance, improving reliability and interpretability of statistical analysis results. However, this approach could lead to misclassification or loss of information, particularly for metabolites that may represent rare exposures that could be associated with rare

outcomes (e.g., certain drugs, chemicals, or foods) or extreme manifestations of common outcomes. Alternately, transformations based on rank only, such as the probit transformation, represent an elegant solution and avoid such exclusions. Log scaling and winsorization are other strategies to reduce the influence of outliers.

Missingness among metabolite data was also a common occurrence. Missing values may be due to biological factors, such as metabolites being absent (e.g., drug metabolites), and various technical limitations in computational detection, including separation of metabolite signal to noise, low signal intensity (e.g., lower value for detection), and measurement error [45]. Various analytical approaches for imputation were reported, including replacement with zero, half (or another proportion) of the minimal detected value, or more complex statistical approaches such as KNN [16], PCA, or random forest imputation [45]. Missing values are often assumed to be due to technical limitations including being below the metabolomics platform's LOD; however, truly absent values must be considered. It is challenging to distinguish between the two as this task requires extensive biochemical knowledge. However, imputing missing values when the metabolite is absent (e.g., for a drug metabolite), is not meaningful and may result in spurious results. Another important consideration is the percentage of missing values per metabolite. Imputation for metabolites with a large proportion of missing values may result in a metabolite with low information content but increasing the multiple testing burden. In this case, exclusion of the metabolite or dichotomization to missing/not-missing values may represent more suitable alternatives. Prior to removing such metabolites, evaluating their relationship with the experimental condition or exposure of interest should still be considered to ensure that valuable information is not discarded. Therefore, analysts must consider the chemical nature or source of extreme or missing metabolite values in determining how to deal with them.

Batch variability or signal drift adjustments were not commonly conducted. These are considered a standard part of the workflow in the field. It is likely that the lack of reporting batch adjustments was due to the use of commercial platforms (e.g., Metabolon Inc. [46] and the Broad Institute [47]) that conduct batch and other laboratory adjustments as part of their standard operating procedures. In some cases when data pre-processing is done by the metabolomics laboratory (or bioinformaticians who work closely with the laboratory), the steps used are not always available to the end user. Additionally, depending on the type of analytical instrument used (i.e., NMR versus MS-based), pre-processing steps could be drastically different [48]. We found that centering, scaling, and transformation of metabolomics data were common, such as adjustment of feature/metabolite intensity by the median across samples, or standard normalization approaches like  $\log_2$  or  $\log_{10}$  transformation [11,49–51]. The most appropriate normalization method for any given large cohort experiment depends on the type of mass spectrometry method used and the size of the cohort experiment; ongoing work is being done to investigate the relative performance of different normalization methods applied to large cohort metabolomics data [52,53].

Most respondents (58%) calculated reliability measures such as metabolite CVs and ICCs, although they generally did not exclude metabolites using these criteria. Encouragingly, of those that measured ICCs, reliability was excellent, on average ( $ICC > 0.9$ ). The variability of metabolite levels in population studies is an important consideration when estimating study power and the true compared with observed study effect estimates. Three main sources of variability include (1) between-subject variability or usual level in the population, (2) within-subject variability representing the usual level within an individual (e.g., year-to-year variability), and (3) technical or laboratory reproducibility or variance expected from identical samples. These components can be integrated into the technical ICC or the proportion of the total variation that is attributable to biologic variance versus random laboratory error [54]. Studies with higher biologic variance and lower technical metabolite variance (e.g., cohorts enriched with certain disease risk traits and measures of primarily highly abundant metabolites) may have higher study power to detect epidemiological associations. Repeated samples can also reduce within-individual variability and improve study power [54,55].

### 3.2. Data Analysis

#### 3.2.1. Analytic Approaches

Following multivariate dimension reduction and/or identification of relevant metabolites, epidemiological methods such as multivariable regression analysis were commonly used for data analysis, but novel and more complex approaches such as adaptations of penalized regression and network analysis are also emerging [56–58]. While the data pre-processing pipeline should be consistent, the data analysis techniques used are driven by the goals of the study. For instance, predictive models like LASSO may be useful for biomarker discovery, but methods applied to the study of mechanisms or metabolomic profiles associated with exposures will depend on the directionality of the underlying biology. Our findings are consistent with a recent survey of the broader metabolomics community that surveyed metabolomics workflow and computation strategies [59]. As an emerging field, metabolomics databases are still incomplete and thus interpreting results from metabolomics datasets in a biological context is challenging. Data-driven network-based approaches support a better understanding of the biological processes driving exposure-disease associations [60], and can provide biological information independent of background databases as well as incorporating unknown metabolites [61]. Examples include Gaussian graphical modeling [62], weighted gene co-expression network analysis [63], sparse network modeling [64], Bayesian approaches [65], and machine learning methods such as random forests [66].

Automated text mining is a bioinformatics approach and queries databases to provide biological context based on a metabolite list [61]. As metabolomics technologies continue to evolve and expand to include larger numbers of novel (i.e., unknown) molecules, the ability for existing databases to provide structure for network analyses becomes more limited. For chemically characterized metabolites, chemical pathway analysis [67] can identify biologically meaningful metabolite groups (i.e., representing biochemical pathways) using information from biochemical databases, and may also serve to strengthen power to detect associations compared with evaluating single metabolites.

To facilitate metabolomics analyses, which are considerably more complex than traditional epidemiological studies, online platforms that aid in identification of relevant biochemical pathways, such as Metaboanalyst [68], the Metabolomics Workbench [69], and others [70] were developed. Moreover, an increasing number of studies are measuring more than one ‘omics data type [71,72] and methods that integrate multiple ‘omics datasets are also under development. Examples of useful tools available for processing and analysis of metabolomics data are presented in Table 1.

**Table 1.** Resources available for analysis and interpretation of metabolomics data. <sup>a</sup>

Resource	Name	Description	Website
Consortia and Societies	Consortium of METabolomics Studies (COMETS)	Consortium of prospective studies with blood metabolomics data.	<a href="https://epi.grants.cancer.gov/comets/">https://epi.grants.cancer.gov/comets/</a> [14]
	Metabolomics Society	Summary of metabolomics databases.	<a href="http://metabolomicssociety.org/">http://metabolomicssociety.org/</a>
	COordination of Standards in MetabOloomicsS (COSMOS)	Standards for data dissemination.	<a href="http://cosmos-fp7.eu/">http://cosmos-fp7.eu/</a> [73]
	Metabolomics Workbench	Metabolomics resource sponsored by the Common Fund of the National Institutes of Health.	<a href="http://www.metabolomicsworkbench.org/">http://www.metabolomicsworkbench.org/</a> [69]
Statistical Analysis Tools; Meta-Data and Other Resources	MetaboAnalyst	Program for statistical, functional and integrative analysis of metabolomics data.	<a href="https://www.metaboanalyst.ca/MetaboAnalyst/faces/home.xhtml">https://www.metaboanalyst.ca/MetaboAnalyst/faces/home.xhtml</a> [68]
	Metabox	A toolbox for metabolomic data analysis, interpretation, and integrative exploration.	<a href="http://kwanjeeraw.github.io/metabox/">http://kwanjeeraw.github.io/metabox/</a> [74]
	MZmine	A modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.	<a href="http://mzmine.github.io/">http://mzmine.github.io/</a> [75]
	XCMSOnline	Metabolomics data processing and analysis platform.	<a href="https://xcmsonline.scripps.edu/landing_page.php?pgcontent=mainPage">https://xcmsonline.scripps.edu/landing_page.php?pgcontent=mainPage</a> [76]
	Workflow4Metabolomics	Collaborative research infrastructure for computational metabolomics.	<a href="https://workflow4metabolomics.org/">https://workflow4metabolomics.org/</a> [77,78]
	PhenoMeNal	Cloud-based platform for metabolomics processing and analysis.	<a href="http://phenomenal-h2020.eu/home/">http://phenomenal-h2020.eu/home/</a> [79]
	Metabolomics Tools Wiki	Classified and searchable list of metabolomics software and tools.	<a href="https://raspicer.github.io/MetabolomicsTools/">https://raspicer.github.io/MetabolomicsTools/</a>
	MetaboLights	Database for metabolomics experiments and derived information.	<a href="https://www.ebi.ac.uk/metabolights/">https://www.ebi.ac.uk/metabolights/</a> [80]
	MetabolomeXchange	An international data aggregation and notification service for metabolomics.	<a href="http://www.metabolomexchange.org/site/">http://www.metabolomexchange.org/site/</a>

<sup>a</sup> The metabolomics resources cited here are provided as a summary of existing tools rather than an endorsement of specific tools.

### 3.2.2. Correction for Multiple Statistical Testing

One of the most important differences between conducting a single biomarker versus a metabolomics analysis within an epidemiological study is the number of tested hypotheses. In order to account for the high number of study hypotheses in many metabolomics studies (particularly for untargeted metabolomics), several methods are available to reduce the rate of Type I errors. Respondents predominantly used the false discovery rate (FDR), which is considered a less stringent approach, followed by Bonferroni correction to account for testing multiple hypotheses. These reflect the most widely used methods currently, although FDR approaches that account for highly correlated data are lacking. A detailed description of these approaches together with proposed alternatives, such as resampling-based strategies, have been reported elsewhere [81]. There is a need to determine the most appropriate method for correcting for multiple statistical testing given the correlated nature of metabolomics data.

### 3.2.3. Classification Performance

Reporting of classification performance is an important step in translating risk factor or disease biomarkers to a clinical setting. Approximately half of respondents noted conducting such analyses. Inconsistent reporting of biomarker classification performance and poor transparency in reporting prediction algorithms have been observed among the broader metabolomics community [82]. Biomarker discovery includes selecting biomarkers that maximally discriminate cases from controls, validating the biomarker panel, and deriving a final model with a fixed mathematical algorithm for predicting the clinical outcome [82]. Measures of biomarker sensitivity, specificity, and receiver operator characteristic (ROC) curves are used to assess the performance of biomarkers for classifying disease diagnosis, prognosis, and risk factor or prediction biomarkers [83]. For disease biomarkers, reporting of ROC curves for disease classification would support biomarker comparison across studies.

### 3.2.4. Meta-Data

Metabolite metadata includes information on metabolomics platform and procedures such as software used, reliability (CV, ICC), chemical identification, mass-to-charge ratio and retention time, chemical pathway, and biological information, among others. This information was not commonly presented in publications. Metadata is crucial for linkage across metabolite databases to retrieve metabolite information, conducting between-study comparisons, metabolite annotation, and informing replication efforts. Databases such as the Human Metabolome database (HMDB) [84] and PubChem [85] assign unique identifiers and compile useful accompanying metadata from previous studies. Quality of metabolomics metadata has been reviewed previously [86]. In an epidemiological setting, appending metabolite metadata will support future replication and meta-analysis efforts, such as those proposed within COMETS.

### 3.2.5. Validation

Cross- and external validation of metabolomics analyses were uncommon among COMETS respondents. External validation represents the gold-standard approach to show generalizability of results. The lack of independent validation is a major challenge in metabolomics biomarker discovery [87]. Validation is particularly important when not all metabolites are stable over long periods of time and across sample collection methods [88]. Extensive costs of acquiring metabolomics data and difficulties in obtaining suitable samples (e.g., in case of a rare disease) may complicate external validation. In that case, as an alternative one may choose to apply cross-validation [89] or double cross-validation [90] by conducting the main analysis on a subset of the study participants and validating the results on the remaining subset. This may often represent the only way to validate results, but the investigators must keep in mind that this approach leads to lower power in both the

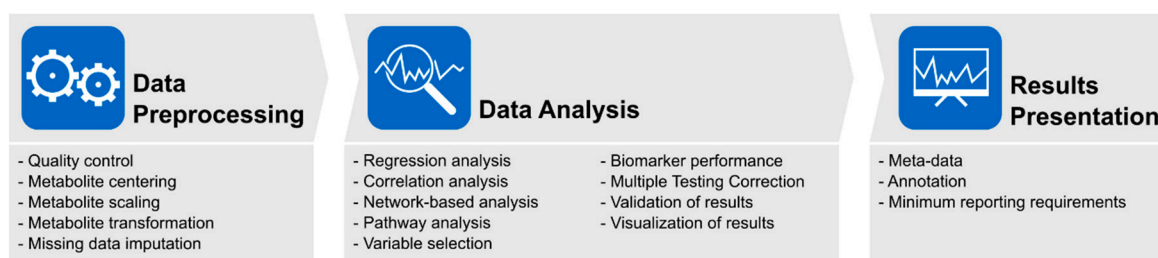
discovery and validation datasets. COMETS provides a unique opportunity for increased biomarker validation in cohorts with diverse participant demographics and clinical features.

### 3.2.6. Coding Language

Most statistical programming was conducted in R. There are many freely available software packages for metabolomics analysis, summarized by the Metabolomics Society [91] and elsewhere [92]. Moreover, platforms such as Galaxy, originally designed for developing genomics research workflow pipelines, have been applied to metabolomics [59]. The recognized advantages of using R and other open source coding languages include the open source framework, which allow investigators and programmers to fully access each line of code and edit or adapt code as needed for a given data management or analysis purpose. For this reason, flexible and open tools are likely to continue being developed in R, with more user-friendly adaptations of code being developed as ready-to-use R packages. Standardizing statistical analysis approaches in workflows developed as R packages will also help to augment potential for replication of analyses across large cohort study designs.

### 3.3. Reporting of Data Analysis Workflow

Pre-analytical and analytical strategies are often poorly reported within scientific manuscripts [93]. A recent review of 27 studies published between 2008 and 2014 assessed the standard of reporting of data management and analysis steps in metabolomics biomarker discovery studies and investigated whether the level of detail reported allows basic understanding of the steps employed and/or reuse of the protocol [13]. The authors concluded that there is unclear and incomplete reporting of these procedures in metabolomics studies that preclude replication in another study. Standardized reporting of observational studies in epidemiology is outlined by the STROBE statement [94] and CONSORT statement [95], with application to genetic epidemiology studies through the Standardized Reporting of Genetic Association Studies (STEGA) [96], among others. Recommendations for standardizing reporting of epidemiological studies with metabolomic analyses and infrastructure to support it have also been proposed [59,77,97], including: experimental design; analytical dataset format; sample handling and data acquisition parameters; post-instrument data processing; multivariate statistical procedures; data modeling; and model validation. Based on our findings, a summary of observed workflow is presented in Figure 9.



**Figure 9.** Suggested metabolomics analysis workflow.

In summary, there is a need to develop standardized analytical workflows, reporting standards, and tools for analysis of metabolomics data in epidemiological studies. Our survey was conducted among a small number of respondents, but they were representatives of their respective prospective cohorts and therefore represent the views of a larger population of analysts. Nonetheless, conducting similar surveys in a larger sample would strengthen the current findings. To coordinate and streamline consortium-based data analyses, COMETS developed COMETS Analytics, a secure online statistical analysis platform for metabolomics data analysis [98]. COMETS Analytics processes summary data generated by participating cohorts, as opposed to individual-level data. The web-based application performs three main tasks: it harmonizes metabolite identifiers across cohorts that utilized different metabolomics platforms, conducts statistical analyses in large batches of user-defined models (including

correlation analysis and multivariable regression), and produces standardized, meta-analysis ready output. The data pre-processing steps are completed by each cohort according to their workflow prior to analysis in COMETS Analytics. The source code is publicly available through GitHub (<https://rdrr.io/github/CBIIT/R-cometsAnalytics/>). The platform aims to accelerate data analysis and lower error rates compared with more conventional approaches [14]. Educational tools are under development to guide analytical workflow and reporting of findings. These resources are intended to be open-source and freely available to the public to support rigorous research efforts and training in the analysis of metabolomics data.

## 4. Materials and Methods

### 4.1. Study Population

The study population included representatives of 47 prospective cohorts participating in the Consortium of Metabolomics Studies (COMETS) [14]. Representatives were cohort Principal Investigators and those with hands-on experience conducting analyses of metabolomics data from their respective cohorts (i.e., research analysts, biostatisticians, research scientists, postdocs, and graduate students).

### 4.2. Questionnaire

Participants were invited to complete an online questionnaire designed to collect information on the workflow and analytical strategies used for past and current metabolomics analyses of epidemiological data within their cohort. The questionnaire was conducted using Survey Monkey (<https://www.surveymonkey.com>) between 8 June and 5 October 2018. Questions were informed by common reporting of metabolomics analysis methods in the literature and were either multiple choice or multiple choice with the option of open-ended responses. Topics included: (1) Study information (purpose of the analysis, study design, type of analysis [targeted, untargeted, or both]); (2) Exploratory analyses and pre-processing (power calculations, data normalization, dealing with missing data, assessing technical platform reliability); (3) statistical analysis (analytic strategies, visualizations, cross-validation); (4) metabolite annotations; (5) other (statistical software, biostatistician input, open-source packages used, statistical coding language, minimal reporting standards). Responses were collated and summarized as frequency of responses (N/%) based on total number of respondents. Open-ended responses were summarized. The questionnaire can be found in the Supplementary Materials.

## 5. Conclusions

We conducted a survey of 47 participating COMETS cohort representatives to gain insights into the strategies and procedures used for analyzing metabolomics data in epidemiological studies worldwide. Our results indicate a large variety of analytical strategies being applied, from data pre-processing and quality control to statistical analysis and reporting of the findings. These methods are both common epidemiological approaches and emerging novel methods. We found that there was consensus on several aspects of metabolomics analysis workflow, including data transformation/normalization, dealing with missing values, multiple testing correction, and choice of statistical software. However, more thought is merited on what would be most appropriate for metabolomics data, such as the optimal multiple testing correction given its highly correlated nature. Moreover, there is a clear need to establish benchmarks in relation to other data pre-processing steps, use of cross and external validation, and minimum reporting standards including reporting metabolite reliability estimates and appending meta-data to study results. Although there was a wide range of analytic approaches applied to metabolomics data, it is likely that analytical choices will continue to depend on the study question and the nature of the data (i.e., targeted or untargeted). Altogether, our results indicate the need for standardized analytical workflows, reporting standards, and openly shared tools for analysis of metabolomics data in large-scale epidemiological studies—an approach that has catalyzed scientific progress in other similarly expansive fields [99]. Accordingly, the open-source COMETS Analytics

initiative [98] is currently developing a set of educational modules and analytic code using common statistical coding language to support the analysis and interpretation of large-scale metabolomics data derived from epidemiological studies. Our current findings can be leveraged to inform the development of minimum reporting standards for metabolomics data analysis to support best practices and study reproducibility. Ultimately, we anticipate this will improve the quality of metabolomics data analysis and results and enable a better comparison and interpretation of the results across metabolomics studies.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2218-1989/9/7/145/s1>, Table S1: Summary of Questionnaire Responses from COMETS Cohort Representatives (n = 33).

**Author Contributions:** Conceptualization, M.C.P. and O.A.Z.; Formal analysis, M.C.P., A.D.J., F.K.T. and O.A.Z.; Methodology, M.C.P., A.D.J., F.K.T. and O.A.Z.; Visualization, A.D.J. and O.A.Z.; Writing – original draft, M.C.P., A.D.J., F.K.T., S.C., M.H., A.K., T.L. and O.A.Z.; Writing – review & editing, M.C.P., A.D.J., F.K.T., S.C., M.H., A.K., T.L., E.H.v.R., J.H., J.K., Y.W., E.M., M.T., S.M., B.C., H.E., A.G., M.J.G., S.H., C.L., M.O., W.P., W.J.S. and O.A.Z.

**Funding:** This research received no external funding. Mary C. Playdon was supported by the National Cancer Institute grant number 5R00CA218694-03 and Huntsman Cancer Institute Cancer Center Support Grant number P30CA040214. E.H. van Roekel was financially supported by Wereld Kanker Onderzoek Fonds (WKOF), as part of the World Cancer Research Fund International grant programme (grant number 2016/1620); A.D.J. was supported by NIDDK grant number K01-DK110267. Fred K. Tabung was supported by National Cancer Institute grant number R00CA207736. Oana A. Zeleznik was supported by the National Cancer Institute grant numbers P01CA087969 and R01CA050385.

**Acknowledgments:** The authors acknowledge the Consortium of Metabolomics Studies (COMETS) Steering Committee in supporting the COMETS Code Repository Working Group and COMETS Analytics.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Su, L.J.; Fiehn, O.; Maruvada, P.; Moore, S.C.; O’Keefe, S.J.; Wishart, D.S.; Zanetti, K.A. The use of metabolomics in population-based research. *Adv. Nutr.* **2014**, *5*, 785–788. [[CrossRef](#)] [[PubMed](#)]
2. Beger, R.D.; Dunn, W.; Schmidt, M.A.; Gross, S.S.; Kirwan, J.A.; Cascante, M.; Brennan, L.; Wishart, D.S.; Oresic, M.; Hankemeier, T.; et al. Metabolomics enables precision medicine: “A White Paper, Community Perspective”. *Metabolomics* **2016**, *12*, 149. [[CrossRef](#)] [[PubMed](#)]
3. Liesenfeld, D.B.; Habermann, N.; Owen, R.W.; Scalbert, A.; Ulrich, C.M. Review of mass spectrometry-based metabolomics in cancer research. *Cancer Epidemiol. Biomark. Prev.* **2013**, *22*, 2182–2201. [[CrossRef](#)] [[PubMed](#)]
4. Guasch-Ferre, M.; Hruby, A.; Toledo, E.; Clish, C.B.; Martinez-Gonzalez, M.A.; Salas-Salvado, J.; Hu, F.B. Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care* **2016**, *39*, 833–846. [[CrossRef](#)] [[PubMed](#)]
5. Johnson, C.H.; Ivanisevic, J.; Siuzdak, G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 451–459. [[CrossRef](#)]
6. Gu, F.; Derkach, A.; Freedman, N.D.; Landi, M.T.; Albanes, D.; Weinstein, S.J.; Mondul, A.M.; Matthews, C.E.; Guertin, K.A.; Xiao, Q.; et al. Cigarette smoking behaviour and blood metabolomics. *Int. J. Epidemiol.* **2016**, *45*, 1421–1432. [[CrossRef](#)] [[PubMed](#)]
7. Guasch-Ferre, M.; Bhupathiraju, S.N.; Hu, F.B. Use of Metabolomics in Improving Assessment of Dietary Intake. *Clin. Chem.* **2018**, *64*, 82–98. [[CrossRef](#)]
8. Moore, S.C.; Matthews, C.E.; Sampson, J.N.; Stolzenberg-Solomon, R.Z.; Zheng, W.; Cai, Q.; Tan, Y.T.; Chow, W.H.; Ji, B.T.; Liu, D.K.; et al. Human metabolic correlates of body mass index. *Metabolomics* **2014**, *10*, 259–269. [[CrossRef](#)]
9. Hivert, M.F.; Perng, W.; Watkins, S.M.; Newgard, C.S.; Kenny, L.C.; Kristal, B.S.; Patti, M.E.; Isganaitis, E.; DeMeo, D.L.; Oken, E.; et al. Metabolomics in the developmental origins of obesity and its cardiometabolic consequences. *J. Dev. Orig. Health Dis.* **2015**, *6*, 65–78. [[CrossRef](#)]
10. Tzoulaki, I.; Ebbels, T.M.; Valdes, A.; Elliott, P.; Ioannidis, J.P. Design and analysis of metabolomics studies in epidemiologic research: A primer on -omic technologies. *Am. J. Epidemiol.* **2014**, *180*, 129–139. [[CrossRef](#)]



11. van den Berg, R.A.; Hoefsloot, H.C.; Westerhuis, J.A.; Smilde, A.K.; van der Werf, M.J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genom.* **2006**, *7*, 142. [[CrossRef](#)] [[PubMed](#)]
12. Misra, B.B.; Langefeld, C.D.; Olivier, M.; Cox, L.A. Integrated Omics: Tools, Advances, and Future Approaches. *J. Mol. Endocrinol.* **2018**, *62*, R21–R45. [[CrossRef](#)] [[PubMed](#)]
13. Considine, E.C.; Thomas, G.; Boulesteix, A.L.; Khashan, A.S.; Kenny, L.C. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics* **2017**, *14*, 7. [[CrossRef](#)] [[PubMed](#)]
14. Yu, B.; Zanetti, K.A.; Temprosa, M.; Albanes, D.; Appel, N.; Barrios Barrera, C.; Ben-Shlomo, Y.; Boerwinkle, E.; Casas, J.P.; Clish, C.; et al. The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies. *Am. J. Epidemiol.* **2019**, *188*, 991–1012. [[CrossRef](#)] [[PubMed](#)]
15. Rosner, B. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics* **1983**, *25*, 165–172. [[CrossRef](#)]
16. 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. [[CrossRef](#)] [[PubMed](#)]
17. van Roekel, E.H.; Loftfield, E.; Kelly, R.S.; Zeleznik, O.A.; Zanetti, K.A. Metabolomics in epidemiologic research: Challenges and opportunities for early-career epidemiologists. *Metabolomics* **2019**, *15*, 9. [[CrossRef](#)]
18. Cheng, S.; Rhee, E.P.; Larson, M.G.; Lewis, G.D.; McCabe, E.L.; Shen, D.; Palma, M.J.; Roberts, L.D.; Dejam, A.; Souza, A.L.; et al. Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation* **2012**, *125*, 2222–2231. [[CrossRef](#)]
19. Guertin, K.; Moore, S.C.; Sampson, J.N.; Huang, W.Y.; Xiao, Q.; Stolzenberg-Solomon, R.Z.; Sinha, R.; Cross, A.J. Metabolomics in nutritional epidemiology: Identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations. *Am. J. Clin. Nutr.* **2014**, *100*, 208–217. [[CrossRef](#)]
20. Mondul, A.M.; Sampson, J.N.; Moore, S.C.; Weinstein, S.J.; Evans, A.M.; Karoly, E.D.; Virtamo, J.; Albanes, D. Metabolomic profile of response to supplementation with  $\beta$ -carotene in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study. *Am. J. Clin. Nutr.* **2013**, *98*, 488–493. [[CrossRef](#)]
21. Newgard, C.B.; An, J.; Bain, J.R.; Muehlbauer, M.J.; Stevens, R.D.; Lien, L.F.; Haqq, A.M.; Shah, S.H.; Arlotto, M.; Slentz, C.A.; et al. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell Metab.* **2009**, *9*, 311–326. [[CrossRef](#)] [[PubMed](#)]
22. Pallister, T.; Jennings, A.; Mohny, R.P.; Yarand, D.; Mangino, M.; Cassidy, A.; MacGregor, A.; Spector, T.D.; Menni, C. Characterizing Blood Metabolomics Profiles Associated with Self-Reported Food Intakes in Female Twins. *PLoS ONE* **2016**, *11*, e0158568. [[CrossRef](#)] [[PubMed](#)]
23. Playdon, M.C.; Ziegler, R.G.; Sampson, J.N.; Stolzenberg-Solomon, R.; Thompson, H.J.; Irwin, M.L.; Mayne, S.T.; Hoover, R.N.; Moore, S.C. Nutritional metabolomics and breast cancer risk in a prospective study. *Am. J. Clin. Nutr.* **2017**, *106*, 637–649. [[CrossRef](#)]
24. Scalbert, A.; Brennan, L.; Manach, C.; Andres-Lacueva, C.; Dragsted, L.O.; Draper, J.; Rappaport, S.M.; van der Hoof, J.J.J.; Wishart, D.S. The food metabolome: A window over dietary exposure. *Am. J. Clin. Nutr.* **2014**, *99*, 1286–1308. [[CrossRef](#)] [[PubMed](#)]
25. Schmidt, J.A.; Rinaldi, S.; Ferrari, P.; Carayol, M.; Achaintre, D.; Scalbert, A.; Cross, A.J.; Gunter, M.J.; Fensom, G.K.; Appleby, P.N.; et al. Metabolic profiles of male meat eaters, fish eaters, vegetarians, and vegans from the EPIC-Oxford cohort. *Am. J. Clin. Nutr.* **2015**, *102*, 1518–1526. [[CrossRef](#)] [[PubMed](#)]
26. Schmidt, J.A.; Rinaldi, S.; Scalbert, A.; Ferrari, P.; Achaintre, D.; Gunter, M.J.; Appleby, P.N.; Key, T.J.; Travis, R.C. Plasma concentrations and intakes of amino acids in male meat-eaters, fish-eaters, vegetarians and vegans: A cross-sectional analysis in the EPIC-Oxford cohort. *Eur. J. Clin. Nutr.* **2016**, *70*, 306–312. [[CrossRef](#)]
27. Zheng, Y.; Yu, B.; Alexander, D.; Steffen, L.M.; Boerwinkle, E. Human metabolome associates with dietary intake habits among African Americans in the atherosclerosis risk in communities study. *Am. J. Epidemiol.* **2014**, *179*, 1424–1433. [[CrossRef](#)] [[PubMed](#)]
28. Floegel, A.; Stefan, N.; Yu, Z.; Mühlenbruch, K.; Drogan, D.; Joost, H.-G.; Fritsche, A.; Häring, H.-U.; Hrabě de Angelis, M.; Peters, A.; et al. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **2013**, *62*, 639–648. [[CrossRef](#)]

29. Huang, J.; Weinstein, S.J.; Kitahara, C.M.; Karoly, E.D.; Sampson, J.N.; Albanes, D. A prospective study of serum metabolites and glioma risk. *Oncotarget* **2017**, *8*, 70366–70377. [[CrossRef](#)]
30. Jiang, X.; Zeleznik, O.A.; Lindstrom, S.; Lasky-Su, J.; Hagan, K.; Clish, C.B.; Eliassen, A.H.; Kraft, P.; Kabrhel, C. Metabolites Associated With the Risk of Incident Venous Thromboembolism: A Metabolomic Analysis. *J. Am. Heart Assoc.* **2018**, *7*, e010317. [[CrossRef](#)]
31. Kraus, W.E.; Muoio, D.M.; Stevens, R.; Craig, D.; Bain, J.R.; Grass, E.; Haynes, C.; Kwee, L.; Qin, X.; Slentz, D.H.; et al. Metabolomic Quantitative Trait Loci (mQTL) Mapping Implicates the Ubiquitin Proteasome System in Cardiovascular Disease Pathogenesis. *PLoS Genet.* **2015**, *11*, e1005553. [[CrossRef](#)] [[PubMed](#)]
32. Kühn, T.; Floegel, A.; Sookthai, D.; Johnson, T.; Rolle-Kampczyk, U.; Otto, W.; von Bergen, M.; Boeing, H.; Kaaks, R. Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study. *BMC Med.* **2016**, *14*, 13. [[CrossRef](#)] [[PubMed](#)]
33. Mayers, J.R.; Wu, C.; Clish, C.B.; Kraft, P.; Torrence, M.E.; Fiske, B.P.; Yuan, C.; Bao, Y.; Townsend, M.K.; Tworoger, S.S.; et al. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nat. Med.* **2014**, *20*, 1193–1198. [[CrossRef](#)] [[PubMed](#)]
34. Menni, C.; Fauman, E.; Erte, I.; Perry, J.R.B.; Kastenmüller, G.; Shin, S.-Y.; Petersen, A.-K.; Hyde, C.; Psatha, M.; Ward, K.J.; et al. Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. *Diabetes* **2013**, *62*, 4270–4276. [[CrossRef](#)] [[PubMed](#)]
35. Mondul, A.M.; Moore, S.C.; Weinstein, S.J.; Karoly, E.D.; Sampson, J.N.; Albanes, D. Metabolomic analysis of prostate cancer risk in a prospective cohort: The alpha-tocolpherol, beta-carotene cancer prevention (ATBC) study. *Int. J. Cancer* **2015**, *137*, 2124–2132. [[CrossRef](#)] [[PubMed](#)]
36. Shah, S.H.; Bain, J.R.; Muehlbauer, M.J.; Stevens, R.D.; Crosslin, D.R.; Haynes, C.; Dungan, J.; Newby, L.K.; Hauser, E.R.; Ginsburg, G.S.; et al. Association of a Peripheral Blood Metabolic Profile With Coronary Artery Disease and Risk of Subsequent Cardiovascular Events. *Circ. Cardiovasc. Genet.* **2010**, *3*, 207–214. [[CrossRef](#)] [[PubMed](#)]
37. Tang, W.H.W.; Wang, Z.; Levison, B.S.; Koeth, R.A.; Britt, E.B.; Fu, X.; Wu, Y.; Hazen, S.L. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* **2013**, *368*, 1575–1584. [[CrossRef](#)]
38. Wang, T.J.; Larson, M.G.; Vasan, R.S.; Cheng, S.; Rhee, E.P.; McCabe, E.; Lewis, G.D.; Fox, C.S.; Jacques, P.F.; Fernandez, C. Metabolite Profiles and the Risk of Developing Diabetes. *Nat. Med.* **2011**, *17*, 448–453. [[CrossRef](#)]
39. Yu, D.; Moore, S.C.; Matthews, C.E.; Xiang, Y.-B.; Zhang, X.; Gao, Y.-T.; Zheng, W.; Shu, X.-O. Plasma metabolomic profiles in association with type 2 diabetes risk and prevalence in Chinese adults. *Metabolomics* **2016**, *12*, 3. [[CrossRef](#)]
40. Zeleznik, O.; Clish, C.B.; Kraft, P.; Avila-Pancheco, J.; Eliassen, A.; Tworoger, S.S. Circulating Lysophosphatidylcholines, Phosphatidylcholines, Ceramides, and Sphingomyelins and Ovarian Cancer Risk: A 23-year Prospective Study. *BioRxiv* **2019**. [[CrossRef](#)]
41. Geijsen, A.; Brezina, S.; Keski-Rahkonen, P.; Baierl, A.; Bachleitner-Hofmann, T.; Bergmann, M.M.; Boehm, J.; Brenner, H.; Chang-Claude, J.; van Duijnhoven, F.J.B.; et al. Plasma metabolites associated with colorectal cancer: A discovery-replication strategy. *Int. J. Cancer* **2019**, *10*. [[CrossRef](#)] [[PubMed](#)]
42. Moore, S.C.; Playdon, M.C.; Sampson, J.N.; Hoover, R.N.; Trabert, B.; Matthews, C.E.; Ziegler, R.G. A Metabolomics Analysis of Body Mass Index and Postmenopausal Breast Cancer Risk. *J. Natl Cancer Inst.* **2018**, *110*, 588–597. [[CrossRef](#)] [[PubMed](#)]
43. Hada, M.; Edin, M.L.; Hartge, P.; Lih, F.B.; Wentzensen, N.; Zeldin, D.C.; Trabert, B. Prediagnostic Serum Levels of Fatty Acid Metabolites and Risk of Ovarian Cancer in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. *Cancer Epidemiol. Biomark. Prev.* **2019**, *28*, 189–197. [[CrossRef](#)] [[PubMed](#)]
44. Watrous, J.D.; Henglin, M.; Claggett, B.; Lehmann, K.A.; Larson, M.G.; Cheng, S.; Jain, M. Visualization, Quantification, and Alignment of Spectral Drift in Population Scale Untargeted Metabolomics Data. *Anal. Chem.* **2017**, *89*, 1399–1404. [[CrossRef](#)] [[PubMed](#)]
45. 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. [[CrossRef](#)] [[PubMed](#)]
46. Evans, A.M.; Bridgewater, B.; Liu, Q.; Mitchell, M.W.; Robinson, R.J.; Dai, H.; Stewart, S.J.; DeHaven, C.D.; Miller, L. High resolution mass spectrometry improves data quality and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. *Metabolomics* **2014**, *2014*, 132.

47. Roberts, L.D.; Souza, A.L.; Gerszten, R.E.; Clish, C.B. Targeted metabolomics. *Curr. Protoc. Mol. Biol.* **2012**, *98*, 30.2.1–30.2.24. [[CrossRef](#)]
48. Issaq, H.J.; Van, Q.N.; Waybright, T.J.; Muschik, G.M.; Veenstra, T.D. Analytical and statistical approaches to metabolomics research. *J. Sep. Sci.* **2009**, *32*, 2183–2199. [[CrossRef](#)]
49. Li, B.; Tang, J.; Yang, Q.; Cui, X.; Li, S.; Chen, S.; Cao, Q.; Xue, W.; Chen, N.; Zhu, F. Performance Evaluation and Online Realization of Data-driven Normalization Methods Used in LC/MS based Untargeted Metabolomics Analysis. *Sci. Rep.* **2016**, *6*, 38881. [[CrossRef](#)]
50. Ejigu, B.A.; Valkenburg, D.; Baggerman, G.; Vanaerschot, M.; Witters, E.; Dujardin, J.C.; Burzykowski, T.; Berg, M. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *OMICS* **2013**, *17*, 473–485. [[CrossRef](#)]
51. Wulff, J.E.; Mitchell, M.W. A Comparison of Various Normalization Methods for LC/MS Metabolomics Data. *Adv. Biosci. Biotechnol.* **2018**, *9*, 339. [[CrossRef](#)]
52. Reisetter, A.C.; Muehlbauer, M.J.; Bain, J.R.; Nodzenski, M.; Stevens, R.D.; Ilkayeva, O.; Metzger, B.E.; Newgard, C.B.; Lowe, W.L., Jr.; Scholtens, D.M. Mixture model normalization for non-targeted gas chromatography/mass spectrometry metabolomics data. *BMC Bioinform.* **2017**, *18*, 84. [[CrossRef](#)] [[PubMed](#)]
53. Wen, B.; Mei, Z.; Zeng, C.; Liu, S. metaX: A flexible and comprehensive software for processing metabolomics data. *BMC Bioinform.* **2017**, *18*, 183. [[CrossRef](#)] [[PubMed](#)]
54. Sampson, J.N.; Boca, S.M.; Shu, X.O.; Stolzenberg-Solomon, R.Z.; Matthews, C.E.; Hsing, A.W.; Tan, Y.T.; Ji, B.T.; Chow, W.H.; Cai, Q.; et al. Metabolomics in epidemiology: Sources of variability in metabolite measurements and implications. *Cancer Epidemiol. Biomark. Prev.* **2013**, *22*, 631–640. [[CrossRef](#)]
55. Xiao, Q.; Moore, S.C.; Boca, S.M.; Matthews, C.E.; Rothman, N.; Stolzenberg-Solomon, R.Z.; Sinha, R.; Cross, A.J.; Sampson, J.N. Sources of variability in metabolite measurements from urinary samples. *PLoS ONE* **2014**, *9*, e95749. [[CrossRef](#)]
56. Czysz, A.H.; South, C.; Gadad, B.S.; Arning, E.; Soyombo, A.; Bottiglieri, T.; Trivedi, M.H. Can targeted metabolomics predict depression recovery? Results from the CO-MED trial. *Transl. Psychiatry* **2019**, *9*, 11. [[CrossRef](#)]
57. Do, K.T.; Kastenmuller, G.; Mook-Kanamori, D.O.; Yousri, N.A.; Theis, F.J.; Suhre, K.; Krumsiek, J. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *J. Proteome Res.* **2015**, *14*, 1183–1194. [[CrossRef](#)]
58. Iqbal, K.; Dietrich, S.; Wittenbecher, C.; Krumsiek, J.; Kuhn, T.; Lacruz, M.E.; Kluttig, A.; Prehn, C.; Adamski, J.; von Bergen, M.; et al. Comparison of metabolite networks from four German population-based studies. *Int. J. Epidemiol.* **2018**, *47*, 2070–2081. [[CrossRef](#)]
59. Weber, R.J.M.; Lawson, T.N.; Salek, R.M.; Ebbels, T.M.D.; Glen, R.C.; Goodacre, R.; Griffin, J.L.; Haug, K.; Koulman, A.; Moreno, P.; et al. Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. *Metabolomics* **2017**, *13*, 12. [[CrossRef](#)]
60. Paynter, N.P.; Balasubramanian, R.; Giulianini, F.; Wang, D.D.; Tinker, L.F.; Gopal, S.; Deik, A.A.; Bullock, K.; Pierce, K.A.; Scott, J.; et al. Metabolic Predictors of Incident Coronary Heart Disease in Women. *Circulation* **2018**, *137*, 841–853. [[CrossRef](#)]
61. Barupal, D.K.; Fan, S.; Fiehn, O. Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. *Curr. Opin. Biotechnol.* **2018**, *54*, 1–9. [[CrossRef](#)] [[PubMed](#)]
62. Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **2011**, *5*, 21. [[CrossRef](#)] [[PubMed](#)]
63. Zhang, B.; Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*. [[CrossRef](#)] [[PubMed](#)]
64. Basu, S.; Duren, W.; Evans, C.R.; Burant, C.F.; Michailidis, G.; Karnovsky, A. Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. *Bioinformatics* **2017**, *33*, 1545–1553. [[CrossRef](#)] [[PubMed](#)]
65. McGeachie, M.J.; Chang, H.H.; Weiss, S.T. CGBayesNets: Conditional Gaussian Bayesian network learning and inference with mixed discrete and continuous data. *PLoS Comput. Biol.* **2014**, *10*, e1003676. [[CrossRef](#)] [[PubMed](#)]
66. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]

67. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [[CrossRef](#)] [[PubMed](#)]
68. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D.S.; Xia, J. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **2018**, *46*, W486–W494. [[CrossRef](#)]
69. Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K.S.; et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **2016**, *44*, D463–D470. [[CrossRef](#)]
70. Do, K.T.; Rasp, D.J.N.; Kastenmuller, G.; Suhre, K.; Krumsiek, J. MoDentify: Phenotype-driven module identification in metabolomics networks at different resolutions. *Bioinformatics* **2019**, *35*, 532–534. [[CrossRef](#)]
71. Shin, S.Y.; Fauman, E.B.; Petersen, A.K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.P.; et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **2014**, *46*, 543–550. [[CrossRef](#)] [[PubMed](#)]
72. Zhang, B.; Hu, S.; Baskin, E.; Patt, A.; Siddiqui, J.K.; Mathe, E.A. RaMP: A Comprehensive Relational Database of Metabolomics Pathways for Pathway Enrichment Analysis of Genes and Metabolites. *Metabolites* **2018**, *8*, 16. [[CrossRef](#)] [[PubMed](#)]
73. Salek, R.M.; Neumann, S.; Schober, D.; Hummel, J.; Billiau, K.; Kopka, J.; Correa, E.; Reijmers, T.; Rosato, A.; Tenori, L.; et al. Coordination of Standards in MetabOmicS (COSMOS): Facilitating integrated metabolomics data access. *Metabolomics* **2015**, *11*, 1587–1597. [[CrossRef](#)] [[PubMed](#)]
74. Wanichthanarak, K.; Fan, S.; Grapov, D.; Barupal, D.K.; Fiehn, O. Metabox: A Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration. *PLoS ONE* **2017**, *12*, e0171046. [[CrossRef](#)] [[PubMed](#)]
75. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **2010**, *11*, 395. [[CrossRef](#)]
76. Tautenhahn, R.; Patti, G.J.; Rinehart, D.; Siuzdak, G. XCMS Online: A web-based platform to process untargeted metabolomic data. *Anal. Chem.* **2012**, *84*, 5035–5039. [[CrossRef](#)] [[PubMed](#)]
77. Giacomoni, F.; Le Corguille, G.; Monsoor, M.; Landi, M.; Pericard, P.; Petera, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.F.; Jacob, D.; et al. Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics* **2015**, *31*, 1493–1495. [[CrossRef](#)]
78. Afgan, E.; Baker, D.; van den Beek, M.; Blankenberg, D.; Bouvier, D.; Cech, M.; Chilton, J.; Clements, D.; Coraor, N.; Eberhard, C.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **2016**, *44*, W3–W10. [[CrossRef](#)]
79. Peters, K.; Bradbury, J.; Bergmann, S.; Capuccini, M.; Cascante, M.; de Atauri, P.; Ebbels, T.M.D.; Foguet, C.; Glen, R.; Gonzalez-Beltran, A.; et al. PhenoMeNal: Processing and analysis of metabolomics data in the cloud. *Gigascience* **2019**, *8*, giy149. [[CrossRef](#)]
80. Haug, K.; Salek, R.M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendrakar, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. MetaboLights—An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013**, *41*, D781–D786. [[CrossRef](#)]
81. Dudoit, S.; Popper Shaffer, J.; Boldrick, J.C. Multiple Hypothesis Testing in Microarray Experiments. *Stat. Sci.* **2003**, *18*, 71–103. [[CrossRef](#)]
82. Xia, J.; Broadhurst, D.I.; Wilson, M.; Wishart, D.S. Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics* **2013**, *9*, 280–299. [[CrossRef](#)] [[PubMed](#)]
83. Wang, Y.; Gapstur, S.M.; Carter, B.D.; Hartman, T.J.; Stevens, V.L.; Gaudet, M.M.; McCullough, M.L. Untargeted Metabolomics Identifies Novel Potential Biomarkers of Habitual Food Intake in a Cross-Sectional Study of Postmenopausal Women. *J. Nutr.* **2018**, *148*, 932–943. [[CrossRef](#)] [[PubMed](#)]
84. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vazquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [[CrossRef](#)] [[PubMed](#)]
85. Dashti, H.; Wedell, J.R.; Westler, W.M.; Markley, J.L.; Eghbalnia, H.R. Automated evaluation of consistency within the PubChem Compound database. *Sci. Data* **2019**, *6*, 190023. [[CrossRef](#)] [[PubMed](#)]

86. Ferreira, J.D.; Inacio, B.; Salek, R.M.; Couto, F.M. Assessing Public Metabolomics Metadata, Towards Improving Quality. *J. Integr. Bioinform.* **2017**, *14*. [CrossRef]
87. Marchand, C.R.; Farshidfar, F.; Rattner, J.; Bathe, O.F. A Framework for Development of Useful Metabolomic Biomarkers and Their Effective Knowledge Translation. *Metabolites* **2018**, *8*, 59. [CrossRef]
88. Townsend, M.K.; Clish, C.B.; Kraft, P.; Wu, C.; Souza, A.L.; Deik, A.A.; Tworoger, S.S.; Wolpin, B.M. Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clin. Chem.* **2013**, *59*, 1657–1667. [CrossRef]
89. Refaeilzadeh, P.; Tang, L.; Liu, H. *Cross-Validation*; Ling, L.M., ÖZSU, T., Eds.; Springer: Boston, MA, USA, 2009; Volume 1.
90. Westerhuis, J.A.; Hoefsloot, H.C.J.; Smit, S.; Vis, D.J.; Smilde, A.K.; van Velzen, E.J.J.; van Duijnhoven, J.P.M.; van Dorsten, F.A. Assessment of PLS-DA cross validation. *Metabolomics* **2008**, *4*, 81–89. [CrossRef]
91. Metabolomics Society. Freely Available Software Tools. Available online: [http://wiki.metabolomicsociety.org/index.php/Freely\\_available\\_software\\_tools](http://wiki.metabolomicsociety.org/index.php/Freely_available_software_tools) (accessed on 25 March 2019).
92. Spicer, R.; Salek, R.M.; Moreno, P.; Canueto, D.; Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics* **2017**, *13*, 106. [CrossRef]
93. Goodacre, R.; Broadhurst, D.; Smilde, A.K.; Kristal, B.S.; Baker, J.D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A.; et al. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **2007**, *3*, 231–241. [CrossRef]
94. Strengthening the Reporting Of Observational Studies in Epidemiology. Available online: <https://www.strobe-statement.org/index.php?id=strobe-publications> (accessed on 25 April 2019).
95. CONSORT Transparent Reporting of Trials. Available online: <http://www.consort-statement.org/> (accessed on 25 April 2019).
96. Oxford, U.O. Enhancing the QUALity and Transparency of health Research. Available online: <http://www.equator-network.org/reporting-guidelines/strobe-strega/> (accessed on 25 April 2019).
97. Lindon, J.C.; Nicholson, J.K.; Holmes, E.; Keun, H.C.; Craig, A.; Pearce, J.T.; Bruce, S.J.; Hardy, N.; Sansone, S.A.; Antti, H.; et al. Summary recommendations for standardization and reporting of metabolic analyses. *Nat. Biotechnol.* **2005**, *23*, 833–838. [CrossRef] [PubMed]
98. Temprosa, E.; Mathe, E. CBIIT/R-cometsAnalytics: Comets Analytics for Consortium Based Metabolomic Analyses. Available online: <https://rdrr.io/github/CBIIT/R-cometsAnalytics/> (accessed on 24 March 2019).
99. Lowndes, J.S.S.; Best, B.D.; Scarborough, C.; Afflerbach, J.C.; Frazier, M.R.; O'Hara, C.C.; Jiang, N.; Halpern, B.S. Our path to better science in less time using open data science tools. *Nat. Ecol. Evol.* **2017**, *1*, 160. [CrossRef] [PubMed]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).