


REVIEW

Open Access



Statistical methods and resources for biomarker discovery using metabolomics

Najeha R. Anwardeen¹, Ilhame Diboun², Younes Mokrab², Asma A. Althani^{1,3} and Mohamed A. Elrayess^{1,3*} 

*Correspondence:
m.elrayess@qu.edu.qa

¹ Research and Graduate Studies,
Biomedical Research Center,
Qatar University, P.O. Box 2713,
Doha, Qatar

² Department of Human
Genetics, Sidra Medicine, Doha,
Qatar

³ QU Health, Qatar University,
Doha, Qatar

Abstract

Metabolomics is a dynamic tool for elucidating biochemical changes in human health and disease. Metabolic profiles provide a close insight into physiological states and are highly volatile to genetic and environmental perturbations. Variation in metabolic profiles can inform mechanisms of pathology, providing potential biomarkers for diagnosis and assessment of the risk of contracting a disease. With the advancement of high-throughput technologies, large-scale metabolomics data sources have become abundant. As such, careful statistical analysis of intricate metabolomics data is essential for deriving relevant and robust results that can be deployed in real-life clinical settings. Multiple tools have been developed for both data analysis and interpretations. In this review, we survey statistical approaches and corresponding statistical tools that are available for discovery of biomarkers using metabolomics.

Keywords: Metabolomics, Metabolomics tools, Statistical methods, Analytical workflow, Univariate, Multivariate

Overview of metabolomics

The term metabolome was first coined in 1998 [1] and became widely established in the early 2000 [2]. Metabolomics profiling is a high-throughput technique that quantifies the levels of endogenous metabolites in a sample (biological fluids, tissues, etc.). [3]. The study of metabolites or metabolite profiling has been gaining popularity in the past decade, thanks to the recent advances in analytical platforms such as Fourier-Transform Infrared spectrometry (FT-IR), Nuclear magnetic resonance (NMR), mass spectrometry (MS) coupled to separation techniques such as gas-chromatography (GC-MS), liquid chromatography (LC-MS), Fourier Transform mass spectrometry (FT-MS), Ultra-high performance liquid chromatography (UPLC-MS), Capillary electrophoresis (CE-MS), Inductively coupled plasma (IPC-MS), Ion chromatography (IC-MS) [4] etc. Metabolites are key molecules in cellular functions. Many biological disturbances involve a cascade of metabolic changes, making metabolites close descriptors for the phenotype. There are two main analytical techniques that are used in the quantification of metabolites (in a cell, tissue, or body fluids): NMR and MS [5–7] through a process that can be untargeted or targeted. The former is a comprehensive technique measuring



all metabolites in a sample without bias, including unknown chemical compounds. It is best suited for hypothesis-generating studies and leads to novel biomarker discovery, although the identification and categorisation of unknown compounds remains a great challenge. On the other hand, targeted metabolomics quantifies chemically known and annotated metabolites. Typically, the measured metabolites are labelled by comparing their masses to known compounds from spectral databases, which in addition to characteristic MS or NMR properties, also contain various information about nomenclature, compound concentrations, biological locations, enzyme and mutation data (see Table 1).

Since its introduction, metabolomics has been used in a wide range of applications such as health and disease biomarker and enzyme discoveries, food and nutrition, and plant biotechnology to name a few [10]. Metabolomics has proven to be a valuable tool in biomedical research, enabling the assessment of disturbances in biological systems caused by environmental factors, aiding in the diagnosis of diseases, and facilitating the identification of biomarkers. Biomarkers, short for, biological markers are objective indicators that provide information about cellular or organismal processes and can be used to characterize patients in a clinical setting [11]. Properties such as high specificity, sensitivity, repeatability, and clinical usefulness are necessary for a good biomarker. The process of biomarker validation entails in vitro and in vivo research followed by clinical trials in human cohorts. Biomarker discovery using metabolomics is considered to be a relatively improved method compared to traditional diagnostic approaches due to its sensitivity and specificity [12]. Metabolites have been found to be eligible molecular biomarkers in several studies; for instance, an untargeted metabolomics approach was used to show that non-alcoholic fatty liver diseases (NAFLD), featuring a range of severity levels from simple steatosis to complex hepatocellular carcinoma, are characterised each with a distinct metabolic profile [13, 14]. Furthermore, metabolomics have shown their

Table 1 Databases containing mass spectra data for metabolite annotation

Database	Comments	Source
Human Metabolome Database (HMDB 5.0)	217,920 known and 1,581,537 unknown compounds. Novel spectral data, physiological and pathological data, pathway data are available in a single platform [8]	https://hmdb.ca/
Golm Metabolome database	Dedicated to GC–MS technique. Contains custom libraries stored as mass spectra (MS) and retention time indices (RI) for metabolic profiling experiments and even observed mass spectral tags (MSTs) of unidentified metabolites	http://gmd.mpimp-golm.mpg.de/
Metlin	240,000 metabolite data is available as neutral or free acids, which enables single, batch, fragment, ion, neutral loss searches. High resolution of 72,000 MS/MS spectra is a key component of this database [9]	https://metlin.scripps.edu/landing_page.php?pgcontent
Massbank	MS database of high resolution spectral data with excellent structural searching methods. More than 41,000 spectra are available	https://massbank.eu/MassBank/
mzCloud	Freely accessible collection of mass spectra of endogenous and exogenous metabolites. Advanced searching capabilities enable finding metabolites that are not included in the library	https://www.mzcloud.org/

potential in diagnosis and management in early screening of oral cancer [15], pancreatic cancer [16], and breast cancer [17]. Additionally, it was shown that recurrence can be monitored using metabolite biomarkers in various cancer patients [18–20]. Further to cancer, metabolomic studies have investigated potential biomarkers associated with fitness [21], telomere length [22], cardiovascular demand [23], steroid profile [24], etc. in elite athletes. Other studies evaluated biomarkers of metabolic diseases such as polycystic ovary syndrome [25], insulin resistance [26–29], and diabetes [30] (See Table 2 for examples of biomarkers from the mentioned studies). With the recent outbreak of COVID-19, emerging metabolomics data have provided insights into COVID-19 pathogenesis in patients with pre-existing chronic conditions such as diabetes, hypertension, hypothyroidism, etc. and revealed biomarkers linked to mechanisms of disease progression, severity, and side-effects of COVID-19 in affected individuals [31–38].

Early biochemical investigations, in the field of metabolomics, featured a low number of measured analytes to ease the interpretation of results [43, 44]. Today, information systems have matured tremendously, and many tools have been developed to assist in analysing and interpreting high throughput metabolomics data. With the continuous advances in instrumental techniques, adopting the correct statistical approach remains critical for proper interpretation and optimal utilization of data. The purpose of this review is to provide an overview of metabolomics data analysis in current research, with special emphasis on methods available for biomarker discovery in human disease.

Metabolomics: analytical challenges and pre-processing

Like other omics fields, the workflow of metabolomics comprises of (i) Experimental design, (ii) Sample collection and preparation, (iii) Data retrieval/acquisition and pre-processing, and (iv) Data analysis and interpretation [45, 46]. Experimental design aids in tightening confidence intervals, minimising confounders and controlling the obvious sources of variation. Sample collection, preparation and data retrieval are the stages where systematic and random errors occur, although these can be controlled via strict work environment and protocol design to some extent [47]. It is during the pre-processing stage that the spectral data are converted to abundance of metabolites in each sample, a crucial link between raw data measurement and statistical analysis. Typical pre-processing steps include deconvolution, library-based identification, and alignment [48] which can be performed by a variety of analytical tools (refer to Table 3). For untargeted metabolomics, this step represents a major challenge due to the lack of spectra for the novel metabolites detected. However, methods to characterize the unknowns are being continuously explored. For example, Knowledge-guided multi-layer networks (KGMN), developed by Zhou et al., were used in untargeted metabolomics to enable global metabolite identification from knowns to unknowns by integrating knowledge-based metabolic reaction network, MS/MS similarity network as well as global peak correlation network [49]. Global network optimization approach, NetID, was recently developed by Chen et al. to annotate untargeted LC–MS data. NetID develops chemically meaningful peak-peak correlations, improves peak assignment accuracy, and creates a single network connecting most observed ion peaks, even for peaks missing MS spectra [50]. Statistical machine learning-based methods are geared towards the identification of unknowns based on feature similarity with the knowns: For instance, (MP-)

Table 2 Biomarker discoveries using metabolomics

Disease	Biomarker	Use	Ref
Esophageal squamous cell carcinoma (ESCC)	3'-UMP, palmitoleic acid, palmitaldehyde, and isobutyl decanoate	Disease recurrence	[39]
Hepatocellular carcinoma (HCC)	Leucine, valine, and tryptophan	Diagnostic biomarkers	[40]
Non- alcoholic fatty liver disease (NAFLD)	Glycocholic acid, Taurocholic acid, Phenylalanine, branched-chain amino-acids, Glutathione	Discrimination of steatosis, steatohepatitis and cirrhosis	[13]
Oral cancer	Pipecolate, Spermidine, Methionine, Tryptophan, Valine, Hypoxanthine, Trimethylamine N-oxide, Guanine, Guanosine, Taurine, Choline, Cadaverine, Threonine	Salivary biomarkers for oral cancer screening	[15]
Pancreatic cancer	1,5-Anhydo-d-glucitol	Diagnostic biomarker	[16]
Estrogen receptor negative breast cancer	Histidine, Glucose, Lactate, Tyrosine	Risk of disease recurrence	[17]
Colorectal cancer	Hexadecanedioic acid, 4-dodecylbenzenesulfonic acid, 2-pyrocatechuic acid, and Formylanthranilic acid	Screening and early detection using serum biomarkers	[41, 42]
Bladder cancer	N ϵ , N ϵ , N ϵ -trimethyllysine, N-acetyltryptophan, dopaquinone, leucine and hypoxanthine	Risk of disease recurrence	[20]
Sports related biomarkers	Glutamine, N-acetylglutamine, xanthine, beta-sitosterol, N2-acetyllysine, stearoyl-arachidonoyl-glycerol (18:0/20:4), N-acetylserine and 3-7-dimethylurate	Leukocyte telomere length prediction	[22]
	Arachidonic acid, branched-chain amino acids, plasmalogens, phosphatidylcholines, phosphatidylethanolamines, Gamma-glutamyl amino acids and glutathione	Potential biomarker signatures for assessing health, performance, and recovery of elite athletes	[23]
	5 α -androstane-3 α ,17 α -diol monosulfate, androstenediol (3 α , 17 α) monosulfate and cortisol	Steroid profile difference in elite female players and non-athletes	[24]
Polycystic ovary syndrome	Hexosylceramide (d18:2/24:0), ceramide (d18.0/24.1) and serine	Predicting low birth weight	[25]
Insulin resistance and diabetes	Androsterone glucuronide, phenylalanine derivative, carboxyethylphenylalanine	Biomarkers associated with insulin resistance in lean/overweight females	[26]
	Glycerophosphoethanolamine, glycerophosphorylcholine and choline	Increased risk of obesity-associated insulin resistance	[27]
	Glutamate	Predictor of gestational diabetes mellitus	[30]

Table 2 (continued)

Disease	Biomarker	Use	Ref
COVID-19	Tryptophan, kynurenine and 3-hydroxykynurenine	Prognostic markers	[33]
	A combination of d-fructose, citric acid and 2-palmitoyl-glycerol	Diagnostic biomarkers	[34]
	Palmitic (C16:0), docosapentaenoic (C22:5, DPA), and docosahexaenoic (C22:6, DHA) in diabetic patients palmitic, oleic (C18:1), and docosahexaenoic acids in hypertensive patients	Predicting disease progression	[36]
	Betaine and branched chain amino acids	Prognostic metabolic biomarkers of severity and mortality respectively	[37]

IOKR [51], MetFrag [52] and CSI:FingerID[53] employ fragmentation trees to learn rules for subclustering of metabolites[52]. Methods like MetFusion [54] were developed to allow access to large spectral databases such as MassBank [55] to allow for improved optimization of predictive models.

Statistics in metabolomics

In addition to analysis challenges encountered with omics data such as high variable dimensionality and intercorrelation, metabolomics data are particularly prone to noise and can be influenced by environment factors, diet, exercise as well as sample handling and batch measurement. In addition, metabolomics data are characterised by a greater extent of data missingness which can compound multivariate analysis and classification techniques. As a consequence, careful application of appropriate statistical methods is required; otherwise, crucial information may get lost or false trends/models may be identified.

The format of metabolomics data is typically a data matrix, with metabolite abundance and samples given in columns and rows or *vice-versa*. Even though metabolomics profiling is highly sought, there are no standard protocols established for the statistical analysis of the produced data. In this review, we discuss some of the widely adopted statistical approaches in recent studies. A simple schematic representation of the steps involved in metabolomics data analysis is depicted in Fig. 1.

Pre-analytical steps

The metabolomics data matrix is prone to elevated metabolite missingness due to several reasons, most notably the inability to measure when metabolite levels are below the detection level as well as technical errors such as peak misalignment or metabolite structural instability. General statistical techniques for multiple imputation have been traditionally applied on metabolomics data but more tailored approaches that acknowledge the frequent non-random pattern of missingness in metabolomics have recently been developed: MetabImpute, an R package which can assess the missingness as completely or partially missing due to randomness and non-randomness (MCAR—missing

Table 3 Quick view of sources for statistical analysis of metabolomics data

	MetaboAnalyst 5.0/ MetaboAnalystR	Mzmine 3	Metabolyzer	PhenoMeNal	SECIMTools	Umetrics SIMCA	XCMS online/ XCMS	MAIT	Omu (count data)	Specmine	pmartR	muma
Platforms	W/R	W	W	W	W	A	W/R	R	R	R	R	R
Ref	[56]	[57]	[58]	[59]	[60]	[61]	[62]	[63]	[64]	[65]	[66]	[67]
Pre-processing	✓	✓		✓	✓		✓	✓		✓	✓	✓
Imputation	✓			✓	✓			✓		✓	✓	✓
Filtering				✓	✓	✓		✓		✓	✓	✓
Normalization	✓		✓	✓	✓	✓		✓		✓	✓	✓
Univariate	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
Multivariate	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ROC analysis	✓		✓	✓	-	✓	✓	✓*		✓*	✓*	✓*

W: web-based, R: R programming language, A: Licensed application

*ROC analyses can be performed with base R functions while using these R packages

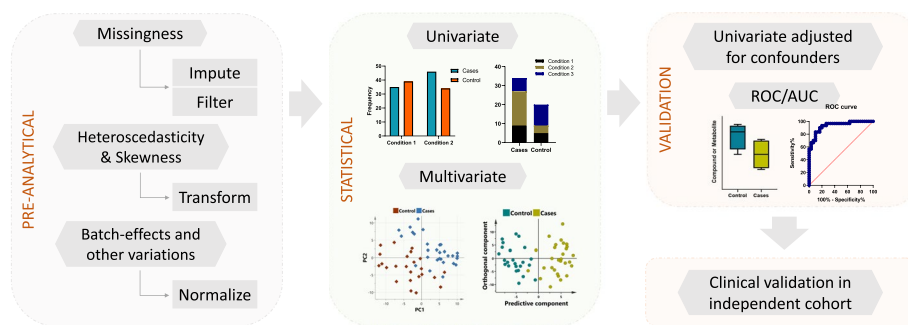


Fig. 1 Simplified workflow of the statistical steps in metabolomics

completely at-random, MAR—missing at-random and MNAR—missing not-at-random) [68]. Indeed, there is no general opinion on the right filter percentage, but cut-offs have been traditionally chosen within the range of 20–50% of metabolite missingness [69, 70]. Imputation is crucial when multivariate techniques, including classification, are applied on metabolomics data.

Complex metabolomics data is heteroscedastic and right skewed and requires normalization. The go-to method for correcting the skewness is log-transformation [71]. Furthermore, filtering of overly heterogenous or bad quality samples is a good practise to avoid the propagation of errors throughout the dataset and can be achieved by means of multivariate techniques such as principal component analysis and clustering. Data normalization, based on aligning the median or more generally quantiles, is crucial to eliminate between-sample variation. It should be noted that using a wrong pre-analytical method to normalize/transform the data will result in poor results and may impact the ranks of relevant metabolites. Additionally, data points should only be removed if there are valid biological justifications for considering them as outliers. It is possible to produce a model that seems to work well by excluding difficult-to-model data points, but that is not actually representational of the real biological system.

Statistical methods

Two main statistical approaches are available for metabolite differential level analysis: univariate and multivariate. Combination of both methodologies is common to metabolomic biomarker-based studies but this review shall focus on the advantages and increased power gained from multivariate analysis (MVA).

MVA is an essential part of metabolomics data analysis. Biological systems are not limited to single variable changes between healthy and diseased states. Investigation of system level changes is pivotal to deriving definitive conclusions about a certain condition and its potential biomarkers. MVA techniques incorporate all variables simultaneously and assess the relationships among them [72] as well as their joint contribution to the phenotype under study.

Unsupervised and supervised models of multivariate analysis are currently employed in metabolomics. One popular unsupervised technique is *Principal component analysis* (PCA) which identifies independent components in the data based on linear combinations of correlated features. Due to its unsupervised nature, PCA serves little purpose in

biomarker discovery. PCA components are often fed into the univariate tests as a means for correcting for hidden unmeasured confounder effects. Moreover, PCA is often used as a checkpoint during QC to screen for outlier data points [73]. For example, when Al-Khelaifi et al. conducted PCA to obtain a global perspective of the data, they noted that PC1 captured the extent of haemolysis between the samples, while PC2 suggested effect of exercise. Incorporation of these components in the regression model greatly improved the detection of marker metabolites in association with the biological groups of interest to their study.

PLS-DA or Partial Least Square Discriminant Analysis [74] is a supervised MVA technique, that has been incorporated in numerous metabolomics studies for the discovery of biomarkers in different health conditions [75–77]. PLS-DA attempts for optimal break-down of predictor variable X to best explain the response variable Y [73, 78, 79]. An upgraded version of PLS-DA called OPLS-DA (“orthogonal” PLS-DA) has also gained popularity [80–84]. This model recapitulates the variance into parts that are predictive of the experimental groups and parts that are purely due to noise, also referred to as ‘orthogonal’ [85–87]. Therefore OPLS-DA creates decipherable models with ease in comparison to its previous version [88, 89]. Once the PLS/OPLS-DA model is built, the VIP (variable influence of projection) measure can be obtained for the metabolites based on their association with the identified predictive components. Certain studies use $VIP > 1$ as a threshold and select the metabolites for further analysis using linear regression models to correct for measured confounders.

Support vector machines (SVM) is a supervised machine-learning algorithm that can be used for regression and classification of non-linear data. SVM can detect non-linear relationships in the data that do not comply with the assumptions of PCA and OPLS, making it versatile. It identifies support vectors or samples on the margin between two classes to search for a maximum margin hyperplane. The use of kernels simplifies separation of classes for difficult cases by providing non-linear solution in the original space. SVM has different extensions for classification with overlapping groups, multi-class classification, regression, and specializations. Importantly, once the hyper-plane partition is found, feature importance values can be derived which can aid in biomarker discovery. However, basic SVM algorithms are not time efficient to tune complex separating hyperplanes as they do not take into account prior knowledge about probability of class-member. Despite often producing good results, faster and more stable methods can outperform SVM [90, 91]. A limitation of SVM is its restriction to binary classification. Alternative methods have been proposed to extend the use of SVM to multi-class problems, the models are built on breaking down the dataset into units of binary groups which causes oversimplification and may lead to uninformative models[92].

Random forest (RF) is a supervised machine learning method which is based on a decision tree algorithm. It is considered an excellent classifier for its ease of implementation, speed, stability robustness against overfitting and most importantly its ability to handle datasets with biased number of classes/groups [93]. Developed by Breiman, RF is a combination of decision trees, with each tree trained using a random subset of the data and input features. The algorithm uses a bootstrap sampling technique to select the data subsets. A simple RF with random features is created by randomly selecting a small group of input variables (with a fixed size) at each node to split on [94]. RF algorithm

is highly adaptable to real-world datasets as it remains unaffected by scaling and normalization. However, a major challenge includes the requirement of excessive tuning of default parameters by the researcher to produce the best model, and eventual difficulty in visualizing the decision tree [95, 96]. Importantly, whilst classifying samples, RF performs a variable selection step which helps reduce the search space and aid in the process of pinpointing candidate metabolite biomarkers.

Variational autoencoders is an unsupervised deep learning method that operates by encoding input data into a non-linear, lower-dimensional latent space that can be used to reproduce the original data without loss of information. It has recently been advocated for use with metabolomics data to learn its transferable latent representations; which can help expose clusters of samples with specific metabolite levels [97].

The classification methods outlined in this review can be prioritised based on the research question and the characteristics of the data. SVM can handle binary and multinomial data with non-linear relationships between variables. RF, on the other hand, is compatible with continuous and categorical data and is used to create an ensemble of decision trees that can capture complex interaction between features, while being robust to outliers and normalization techniques. VAE is notably novel in the field of metabolomics and any added advantage to its use are yet to be shown.

OPLS/OPLS-DA is an excellent choice for small sized and highly correlated data with few groups of samples. It can explain most of the variation in the data by reducing the high dimensionality into predictive and orthogonal latent variables. It handles the missing values in the data efficiently and is robust to outliers [95]. One can argue that both RF and OPLS-DA methods are a good starting point for exploring metabolomics data due to their easiness of use and interpretability. Table 4 provides an overview of the methods, their strengths and weaknesses to be considered in metabolomics data analysis.

It is important to note that classification, prediction and biomarker discovery methods for metabolomics data extend to other models including logistic regression models, LASSO, CCPLS, ASCA + and APCA + (extension of ANOVA to multivariate classes) [98], multivariate curve resolution (MCR), neural networks, Gaussian mixture modeling etc. More details about these methods and how they have been deployed in the field of metabolomics can be found in [99–103].

Validation of model performance

Several metrics exist for assessment of model performance. With OPLS and PLS models, typical measures are R^2 which captures the goodness of fit, and the Q^2 that computes the predictive ability of the model, defined as the congruence of cross-validation of predicted data with the original data. OPLS further splits R^2X into R^2X_p and R^2X_o which respectively measure the explained sum of squared of the Y-predictive and Y-uncorrelated parts of X. [104]. $Q^2 > 0.4$ provides a satisfactory predictability of the model [105, 106]. Q^2 and R^2 values that are closer to 1 ensure a reliable model, while large discrepancy between the two scores depict an unreliable model [107]. Permutation tests are used to estimate Q^2 and provide a possibility of calculating significance (p -values) for these MV models [108–110].

Brier score is another CV procedure that measures the accuracy of binary outcome predictions by calculating the squared difference between the actual outcome and

Table 4 Synopsis of popular statistical methods for metabolomics studies

	Methods	Strengths	Limitations
Univariate	T test Mann Whitney Chi-square ANOVA Kruskal Wallis	Straightforward application Easy to interpret the results	Requires prior knowledge of data No information about inter-variable relationships that is crucial in a biological set-up Outliers cannot be determined
	Multiple linear regression with Bonferroni correction (with one explanatory variable)	Easy to apply and interpret	Significance level affected by sample size Does not account for intercorrelation
	Multiple linear regression with false discovery rate (with one explanatory variable)	Easy to use and interpret Preferred over Bonferroni method	Increases the number of false negatives
Multivariate	Principle component analysis	Effective in variable reduction Uses the complete collected data Easy to manage complex data Focuses on the inter-variable relationships Requires no prior knowledge of data	No clarity on how to rank the metabolites Biological interpretation may be challenging
	Partial least square discriminant analysis Orthogonal partial least square discriminant analysis	Dimensional reduction to comprehensible level No data wastage Shows relationship between variables, apt in a biological setting Handles large, complex data	Prior knowledge of data required Over-fitting issues No significance level of the most important metabolites Abundant variables mask the effect of lesser abundant variables Cross-validation steps required to predict accuracy of model
	Random Forest, SVM and other ML methods	Handles complex data Robust to outliers Finds complex relationships between metabolites and between metabolite and other factors	Excessive tuning may be required to retrieve best model Less efficient for truly linear data Does not provide metabolite selection

predicted probability. A perfect model has a score of 0 and a non-informative model has a score of 0.25[111]. Harrell's C-index is also a performance measure used with survival analyses. The index is driven by Kendall's tau statistic, depends on the censoring distribution, and considers the rankings of pairings of subjects in the data. The index ranges from 0 to 1 (indicating worst to best performance) and a value of 0.6 or higher is acceptable for clinical datasets [112].

The receiver operating characteristics (ROC) curve analysis assesses the specificity and sensitivity of a potential biomarker by plotting the true positive rate (y axis) as a function of the false positive rate (x axis). It produces the area under the curve (AUC) measure that indicates the ability of a biomarker to distinguish between two study groups. Multivariate receiver operating characteristic analysis (MultiROC) [113] is an extension of ROC analysis that allows for different combinations of biomarkers to be clinically explored [114, 115] and is compatible with the inherent nature of multivariate classifiers such as PLS/OPLS-DA models.

There are other cross-validation procedures employed in predictive analysis such as leave-n-out, Monte Carlo cross-validation (MCCV), corrected-MCCV (CMCCV) etc. For detailed information, readers are referred to Sammut et al. and Xu et al. [116, 117].

The metrics outlined above have been instrumental in assessing the performance of MV classification methods to ensure validity and reliability of the results. For example, a study by Chen et al. compared four classifiers, PCA, SVM, LDA and RF using several methods including cross-validation, R^2/Q^2 plot, ROC curve and Pearson correlation. RF was found to be associated with better performance with respect to sample classification and biomarker selection [118].

Tools available for the statistical analysis of metabolomics data

Several tools are available for data analysis in metabolomics. The tools required for highly intricate metabolomics data analysis should be able to handle the large data size, perform pre-processing steps adequately, conduct statistical methods to identify significantly different metabolites, and provide striking visualization techniques such as heatmaps, correlation and pathway networks. We intend to cover some of the widely used tools that provide data pre-processing, univariate and multivariate methodologies used for biomarker discovery. Table 3 provides a quick view of the methods available in the tools discussed below.

(i) **MetaboAnalyst**: Extensive web-based toolkit for complete data analysis of metabolomics data. It provides multiple statistical workflows for one-factor, two-factor/time-series, meta-analysis data formats, which include univariate (t-tests, ANOVA) and multivariate (PCA, PLS-DA, OPLS-DA). The latest version (MetaboAnalyst 5.0) is user-friendly compared to its predecessor. It contains a biomarker discovery option using ROC analyses with straightforward data input and user-defined options for pre-processing steps and normalization. This web-based platform has been utilized in various studies for biomarker identification due to its amenable nature [119–123].

(ii) **MZmine 3**:

Built on the success and popularity of MZmine 2, MZmine 3 is an open-source platform for data pre-processing and analyses with LC-MS in mind. The updated version has focused on improving the user-friendly graphics with the original eight modules [124].

(iii) **MetaboLyzr**:

It is a command line interface (CLI) providing general as well as metabolomics-suited statistical analysis and data visualization [125]. Integration with small-metabolite databases such as HMDB, KEGG, BioCyc and LipidMaps allows for ion identification and relevant data analysis. We would argue that it is more appropriate for expert-level bioinformatician in terms of user-friendliness.

(iv) **PhenoMeNal**:

To our knowledge, a comprehensive and unmatched tool that brings metabolomics to cloud computing after Galaxy. Ongoing immense data generation requires cloud-based tools to reduce the load on personal or workplace environment by storing the data onto cloud space. Data analysis tools are tested and stored as Docker containers [126]. PhenoMeNal has successfully developed sophisticated data analysis workflows, which reduces the burden on the researcher.

(v) **SECIMTools** (SouthEast Center for Integrated Metabolomics):

Designed to complement both the previous Galaxy metabolomics tools, Galaxy-M and Workflow4metabolomics, SECIMtools begins with features which follows quality control (QC) and advanced statistical assessment. It has four major functionalities: data

pre-processing, QC, data analysis and utilities [127]. A guide to use the galaxy interface of SECIMTools can be found here. [https://ctsi-secim.sites.medinfo.ufl.edu/files/2015/08/7_7_2015_Galaxy_UserGuide.pdf]

(vi) SIMCA®

By Sartorius AG, SIMCA is the tool of choice for multivariate analysis by many studies [108]. It is user-friendly, with multiple interactive visualization methods, has the ability to fit models that best suit the data at hand, perform ROC analysis, analyse multiple datasheets, to name a few. For metabolomics, investigation of metabolites with significantly different abundances, metabolite pathways (if present in the datasheet) associated with experimental groups, examining relationship between variables and quick identification of potential biomarkers are relatively easy for non-programmers. SIMCA contains in-built cross validation steps that provide the predictive ability of the model. Although this tool is not suited to univariate analysis and is not in an open-source format and requires license purchase prior to use.

(vii) R (R foundation for statistical computing, Vienna, Austria) [128] packages for metabolomics:

For statisticians who are well-versed in programming languages, R is the best option for metabolomics data analysis as it provides a more flexible work environment as opposed to rigid online tools with limited user-defined options. There are several packages for normalization, imputation, univariate hypothesis testing, multivariate exploratory analysis in R.

(a) XCMS.

R based powerful tool for processing of LC-MS data using retention time correction, peak identification and matching to derive necessary information. It can be combined with base R functions to perform all statistical methods for a comprehensive data analysis.

(b) MetaboAnalystR.

Corresponding R package of web-based MetaboAnalyst, with more adjustable programming feature to enable autonomy of metabolomics data analysis.

(c) MAIT (Metabolite Automatic Identification Toolkit):

Provides a comprehensive end-to-end analysis for LC-MS data. Although it is more suited to peak identification and annotation. Parametric and non-parametric univariate statistical tools and multivariate analyses such as PLS-DA are available with user defined grouping option [63].

(d) Omu.

Performs simple t-tests, ANOVA, PCA and combine functional information and the associated gene names of the metabolites in the dataset using KEGG. It was developed for inexperienced R users to analyze metabolite count data. The input format should contain KEGG IDs to process the data. The package contains multiple visualization techniques such boxplots, heatmaps, volcano plots etc. [64].

(e) Specmine.

Multi-level analysis is available in this package, which includes pre-processing, metabolite annotation, uni- and multivariate analyses, ML (machine learning) and selection of significant features [65].

(f) pmartR.

Quality control processing, statistical analysis of metabolomics, lipidomics and proteomics data can be performed using *pmartR*. Analyses such as transformation, normalization, simple univariate and summarising PCA and correlation analyses are available [66].

(g) *muma*.

Built with non-programmers in mind, *muma* provides user-friendly stepwise univariate and multivariate analysis via R program. Data pre-processing, imputation, data exploration through various visualizations and statistical analysis are available in this package [67].

Limitations of statistics in biomarker discovery

Biomarkers are measured indicators of biological and/or pathogenic processes, or response to therapies [11]. Metabolite biomarkers are quantified at a cheaper rate compared to other types of biomarkers [129]. There is certainly a rapid increase in the number of metabolite biomarkers discovered due to improvements in the analytical procedures but are not in practical use due to limitations in experimental design, statistical rigor, and efficacy [130, 131]. Biomarkers in clinical practice should be easy to quantify and should bring value in relation to early detection of disease, improvement in treatment outcomes, reduction in the reliance on expensive treatment options, or decrease in disease-related fatalities. Unfortunately, appropriate biomarkers with appreciable specificity and sensitivity are hard to come by. Using the combinatorial capacity of a variety of distinct biomarkers is one possibility to improve the overall specificity [132, 133]. Present-day metabolomics have substantially benefited from upgraded study design that contributed to the decrease in the demographic differences and sources of bias. This approach has been applied to all sorts of study designs such as interventional, observational, and with multi-tiers. Study enrolment with balanced demographic attributes under a multi-cohort setting should have sufficient sample size to comply with the requirements for adequate statistical power [134]. Improvised prospective trials are required to verify biomarkers' ability to detect physiological changes before onset of phenotype. Validation of biomarkers has been carried out in small, unbridled trials so far [135]. However, large scale validation remains inadequate leading to very few metabolomics biomarkers finding their way to clinical practice [136, 137]. More insights on ways in which metabolomics research can be advanced to meet the challenges of biomarker discovery can be found in Poste et al. [136].

Conclusion

This mini review has introduced the user to standard methodologies with easy-to-use tools for analysis of metabolomics datasets and biomarker discovery. Metabolite biomarkers are constantly growing interest in the omics field as they depict a phenotype as close to accurate as possible from the physiological or pathological state. In the future, we expect the evolution of existing statistical methods to provide even deeper insights into metabolite biomarkers from the larger perspective of systems' biology and precision medicine. In this context, biomarkers identified using multi-omics techniques can broaden the scope of individualized treatment plans by providing markers for patient stratification, early diagnosis, prediction, and progression monitoring, etc.

To this end, advanced statistical and machine learning methods are being developed to provide effective approaches for multi-omics data integration [138]. Aligning the biological information from multi-level omics analysis has the advantage of reducing noise and provides an extra level of biomarker validation. More importantly, integration with genotype data can help distinguish biomarkers associated with causal effects as opposed to those of secondary nature, that occur because of the disease or pathology of interest as well as those contributed by the environment. Methods for stratification of patients into homogeneous groups with unified analyte levels, such as supervised biclustering [132, 133], have been recently applied in the field of transcriptomics and offer an interesting opportunity for metabolomics to embark in the field of precision medicine.

In parallel to technological advancement, progress in computational and statistical analysis is also required to tackle some of the remaining limitations in the field of metabolomics; notably with regard to annotation/identification of unknown compounds with untargeted metabolomics. Machine learning approaches are of great value in this respect and can offer better performance with improved and more accurate information on compound masses, retention time, fragment mass spectra, and isotopic properties [134].

It should be noted that all statistical methods incorporated in the field of omics are simply hypothesis creators, essentially shortening a seemingly limitless list of metabolites to a manageable set whose properties and merits should be evaluated by downstream experimental work. Standardization of validation protocols including replication and experimental validation in animal models is essential for metabolite biomarkers to make their way to pre-clinical settings.

Abbreviations

AUC	Area under the curve
CE-MS	Capillary electrophoresis–mass spectrometry
CMCCV	Corrected- Monte Carlo cross-validation
FT-IR	Fourier-transform infrared
FT-MS	Fourier transform–mass spectrometry
GC-MS	Gas chromatography–mass spectrometry
IPC-MS	Inductively coupled plasma mass spectrometry
IC-MS	Ion chromatography–mass spectrometry
KGMMN	Knowledge-guided multi-layer networks
KEGG	Kyoto encyclopedia of genes and genomes
LC-MS	Liquid chromatography–mass spectrometry
LDA	Linear dimension analysis
MCCV	Monte Carlo cross-validation (MCCV)
NMR	Nuclear magnetic resonance
ROC	Receiver operating characteristics
UPLC-MS	Ultra-high performance liquid chromatography–mass spectrometry

Acknowledgements

Authors would like to thank Qatar National Research Fund (QNRF) for funding this project and open access funding is provided by the Qatar National Library.

Author contributions

NRA conducted the literature review and drafted the manuscript. ID and MAE conceived the idea of the study and helped to draft the manuscript. YM reviewed the concepts and performed scientific editing of the manuscript. NRA and MAE finalized the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding provided by the Qatar National Library. This research was funded by the Qatar National Research Fund (QNRF), grant number NPRP135-1230-190008.

Availability of data and materials

Not applicable.

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent to publish

Not applicable.

Competing interests

Authors declare no competing interests.

Received: 30 October 2022 Accepted: 9 June 2023

Published online: 15 June 2023

References

1. Oliver SG, et al. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 1998;16(9):373–8.
2. Griffin JL. The Cinderella story of metabolic profiling: Does metabolomics get to go to the functional genomics ball? *Philos Trans R Soc Lond B Biol Sci.* 2006;361(1465):147–61.
3. Clish CB. Metabolomics: an emerging but powerful tool for precision medicine. *Cold Spring Harb Mol Case Stud.* 2015;1(1): a000588.
4. Macedo AN, et al. Analytical platforms for mass spectrometry-based metabolomics of polar and ionizable metabolites. *Adv Exp Med Biol.* 2021;1336:215–42.
5. Schrimpe-Rutledge AC, et al. Untargeted metabolomics strategies-challenges and emerging directions. *J Am Soc Mass Spectrom.* 2016;27(12):1897–905.
6. Wang JH, Byun J, Pennathur S. Analytical approaches to metabolomics and applications to systems biology. *Semin Nephrol.* 2010;30(5):500–11.
7. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol.* 2016;17(7):451–9.
8. Wishart DS, et al. HMDB 5.0: the human metabolome database for 2022. *Nucl Acids Res.* 2021;50(D1):D622–31.
9. Guijas C, et al. METLIN: a technology platform for identifying knowns and unknowns. *Anal Chem.* 2018;90(5):3156–64.
10. Gomez-Casati DF, Zanor MI, Busi MV. Metabolomics in plants and humans: applications in the prevention and diagnosis of diseases. *Biomed Res Int.* 2013;2013: 792527.
11. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS.* 2010;5(6):463–6.
12. Kotlowska A, Szefer P. Recent advances and challenges in steroid metabolomics for biomarker discovery. *Curr Med Chem.* 2019;26(1):29–45.
13. Masarone M, et al. Untargeted metabolomics as a diagnostic tool in NAFLD: discrimination of steatosis, steatohepatitis and cirrhosis. *Metabolomics.* 2021;17(2):12.
14. Masoodi M, et al. Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests. *Nat Rev Gastroenterol Hepatol.* 2021;18(12):835–56.
15. Ishikawa S, et al. Identification of salivary metabolomic biomarkers for oral cancer screening. *Sci Rep.* 2016;6:31520.
16. Kobayashi T, et al. A novel serum metabolomics-based diagnostic approach to pancreatic cancer. *Cancer Epidemiol Biomarkers Prev.* 2013;22(4):571–9.
17. Tenori L, et al. Serum metabolomic profiles evaluated after surgery may identify patients with oestrogen receptor negative early breast cancer at increased risk of disease recurrence. Results from a retrospective study. *Mol Oncol.* 2015;9(1):128–39.
18. Loras A, et al. Bladder cancer recurrence surveillance by urine metabolomics analysis. *Sci Rep.* 2018;8(1):9172.
19. Zhang F, et al. Metabolomics for biomarker discovery in the diagnosis, prognosis, survival and recurrence of colorectal cancer: a systematic review. *Oncotarget.* 2017;8(21):35460–72.
20. Alberice JV, et al. Searching for urine biomarkers of bladder cancer recurrence using a liquid chromatography-mass spectrometry and capillary electrophoresis-mass spectrometry metabolomics approach. *J Chromatogr A.* 2013;1318:163–70.
21. AlMuraikhy S, et al. Comparing the metabolic profiles associated with fitness status between insulin-sensitive and insulin-resistant non-obese individuals. *Int J Environ Res Public Health.* 2022. **19**(19).
22. Al-Muraikhy S, et al. Metabolic signature of leukocyte telomere length in elite male soccer players. *Front Mol Biosci.* 2021;8: 727144.
23. Al-Khelaifi F, et al. Metabolic profiling of elite athletes with different cardiovascular demand. *Scand J Med Sci Sports.* 2019;29(7):933–43.
24. Tarkhan AH, et al. Comparing metabolic profiles between female endurance athletes and non-athletes reveals differences in androgen and corticosteroid levels. *J Steroid Biochem Mol Biol.* 2022;219: 106081.
25. Diboun I, et al. Metabolomic profiling of pregnancies with polycystic ovary syndrome identifies a unique metabolic signature and potential predictive biomarkers of low birth weight. *Front Endocrinol (Lausanne).* 2021;12: 638727.
26. Diboun I, et al. Metabolomics of lean/overweight insulin-resistant females reveals alterations in steroids and fatty acids. *J Clin Endocrinol Metab.* 2021;106(2):e638–49.

27. Al-Sulaiti H, et al. Metabolic signature of obesity-associated insulin resistance and type 2 diabetes. *J Transl Med.* 2019;17(1):348.
28. Al-Sulaiti H, et al. Triglyceride profiling in adipose tissues from obese insulin sensitive, insulin resistant and type 2 diabetes mellitus individuals. *J Transl Med.* 2018;16(1):175.
29. Helaleh M, et al. Association of polybrominated diphenyl ethers in two fat compartments with increased risk of insulin resistance in obese individuals. *Chemosphere.* 2018;209:268–76.
30. Diboun I, et al. Metabolic profiling of pre-gestational and gestational diabetes mellitus identifies novel predictors of pre-term delivery. *J Transl Med.* 2020;18(1):366.
31. Song JW, et al. Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. *Cell Metab.* 2020;32(2):188–202 e5.
32. Shen B, et al. Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell.* 2020;182(1):59–72 e15.
33. Hasan MR, Suleiman M, Pérez-López A. Metabolomics in the diagnosis and prognosis of COVID-19. *Front Genet.* 2021;12: 721556.
34. Shi D, et al. The serum metabolome of COVID-19 patients is distinctive and predictive. *Metabolism.* 2021;118: 154739.
35. Dogan HO, et al. Understanding the pathophysiological changes via untargeted metabolomics in COVID-19 patients. *J Med Virol.* 2021;93(4):2340–9.
36. Elrayess MA, et al. Metabolic signatures of type 2 diabetes mellitus and hypertension in COVID-19 patients with different disease severity. *Front Med (Lausanne).* 2021;8: 788687.
37. Diboun I, et al. Identification of prognostic metabolomic biomarkers at the interface of mortality and morbidity in pre-existing TB cases infected With SARS-CoV-2. *Front Cell Infect Microbiol.* 2022;12: 929689.
38. Taleb S, et al. Predictive biomarkers of intensive care unit and mechanical ventilation duration in critically-ill coronavirus disease 2019 patients. *Front Med (Lausanne).* 2021;8: 733657.
39. Zhu Q, et al. Metabolomic analysis of exosomal-markers in esophageal squamous cell carcinoma. *Nanoscale.* 2021;13(39):16457–64.
40. Morine Y, et al. Essential amino acids as diagnostic biomarkers of hepatocellular carcinoma based on metabolic analysis. *Oncotarget.* 2022;13(1):1286.
41. Liesenfeld DB, et al. Metabolomics and transcriptomics identify pathway differences between visceral and subcutaneous adipose tissue in colorectal cancer patients: the ColoCare study. *Am J Clin Nutr.* 2015;102(2):433–43.
42. Zhang C, et al. Metabolomic profiling identified serum metabolite biomarkers and related metabolic pathways of colorectal cancer. *Dis Markers.* 2021;2021:6858809.
43. Bhattacharya M, et al. Single-run separation and detection of multiple metabolic intermediates by anion-exchange high-performance liquid chromatography and application to cell pool extracts prepared from *Escherichia coli*. *Anal Biochem.* 1995;232(1):98–106.
44. Tweeddale H, Notley-McRobb L, Ferenci T. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool (“Metabolome”) analysis. *J Bacteriol.* 1998;180(19):5109–16.
45. Manchester M, Anand A. Metabolomics: Strategies to define the role of metabolism in virus infection and pathogenesis. *Adv Virus Res.* 2017;98:57–81.
46. Nalbantoglu, S. (2019) Metabolomics: basic principles and strategies. *Molecular Medicine, IntechOpen*
47. Korman A, et al. Statistical methods in metabolomics. *Methods Mol Biol.* 2012;856:381–413.
48. Mastrangelo A, et al. From sample treatment to biomarker discovery: a tutorial for untargeted metabolomics based on GC-(EI)-Q-MS. *Anal Chim Acta.* 2015;900:21–35.
49. Zhou Z, et al. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat Commun.* 2022;13(1):6656.
50. Chen L, et al. Metabolite discovery through global annotation of untargeted metabolomics data. *Nat Methods.* 2021;18(11):1377–85.
51. Brouard C, et al. Magnitude-preserving ranking for structured outputs, in Proceedings of the Ninth Asian Conference on Machine Learning, Z. Min-Ling and N. Yung-Kyun, Editors. 2017, PMLR: Proceedings of Machine Learning Research. p. 407–422.
52. Ruttkies C, Neumann S, Posch S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics.* 2019;20(1):376.
53. Dührkop K, et al. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A.* 2015;112(41):12580–5.
54. Gerlich M, Neumann S. MetFusion: integration of compound identification strategies. *J Mass Spectrom.* 2013;48(3):291–8.
55. Horai H, et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom.* 2010;45(7):703–14.
56. *MetaboAnalyst 5.0.* Available from: <https://www.metaboanalyst.ca/>.
57. *Mzmine 3.* Available from: <http://mzmine.github.io/>.
58. *MetaboLyzer.* Available from: <https://sites.google.com/a/georgetown.edu/fornace-lab-informatics/home/metabolyzer>.
59. *PhenoMeNa.* Available from: <https://phenomenal-h2020.eu/home/>.
60. *SECIMTools.* Available from: <http://secim.ufl.edu/secim-tools/secim-galaxy-tools/>.
61. *SIMCA.* Available from: <https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca>.
62. XCMS online. <https://xcmsonline.scripps.edu/>.
63. Fernández-Albert F, et al. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). *Bioinformatics.* 2014;30(13):1937–9.
64. Tiffany CR, Bäumlér AJ. omu, a metabolomics count data analysis tool for intuitive figures and convenient meta-data collection. *Microbiol Resour Announc.* 2019;8(15):e00129–e219.

65. Costa C, Maraschin M, Rocha M. An R package for the integrated analysis of metabolomics and spectral data. *Comput Methods Progr Biomed*. 2016;129:117–24.
66. Stratton KG, et al. pmartR: quality control and statistics for mass spectrometry-based biological data. *J Proteome Res*. 2019;18(3):1418–25.
67. Gaude E, et al. muma, an R package for metabolomics univariate and multivariate statistical analysis. *Curr Metabol*. 2013;1(2):180–9.
68. Davis TJ, et al. Addressing missing data in GC x GC metabolomics: Identifying missingness type and evaluating the impact of imputation methods on experimental replication. *Anal Chem*. 2022;94(31):10912–20.
69. Payne TG, et al. A signal filtering method for improved quantification and noise discrimination in fourier transform ion cyclotron resonance mass spectrometry-based metabolomics data. *J Am Soc Mass Spectrom*. 2009;20(6):1087–95.
70. Bijlsma S, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal Chem*. 2006;78(2):567–74.
71. Antonelli J, et al. Statistical workflow for feature selection in human metabolomics data. *Metabolites*. 2019;9(7):143.
72. Dillon WR, Goldstein M. *Multivariate analysis: methods and applications*. New York: Wiley; 1984.
73. Chen Y, Li EM, Xu LY. Guide to metabolomics analysis: a bioinformatics workflow. *Metabolites*. 2022;12(4):357.
74. Barker M, Rayens W. Partial least squares for discrimination. *J Chemom*. 2003;17(3):166–73.
75. Broughton-Neiswanger LE, et al. Urinary chemical fingerprint left behind by repeated NSAID administration: discovery of putative biomarkers using artificial intelligence. *PLoS ONE*. 2020;15(2):e0228989.
76. Lopez-Hernandez Y, et al. Targeted metabolomics identifies high performing diagnostic and prognostic biomarkers for COVID-19. *Sci Rep*. 2021;11(1):14732.
77. Kelly RS, et al. Partial least squares discriminant analysis and Bayesian networks for metabolomic prediction of childhood asthma. *Metabolites*. 2018;8(4):68.
78. Worley B, Powers R. PCA as a practical indicator of OPLS-DA model reliability. *Curr Metabolomics*. 2016;4(2):97–103.
79. Brereton RG, Llyod GR. Partial least squares discriminant analysis: taking the magic away. *J Chemom*. 2014;28(4):213–25.
80. Tonoyan NM, et al. Alterations in lipid profile upon uterine fibroids and its recurrence. *Sci Rep*. 2021;11(1):11447.
81. Minale G, et al. Characterization of metabolites in plasma, urine and feces of healthy participants after taking brahmi essence for twelve weeks using LC-ESI-QTOF-MS metabolomic approach. *Molecules*. 2021;26(10):2944.
82. Liu H, et al. UHPLC-Q-Orbitrap-HRMS-based global metabolomics reveal metabolome modifications in plasma of young women after cranberry juice consumption. *J Nutr Biochem*. 2017;45:67–76.
83. Pang Z, et al. Serum metabolomics analysis of asthma in different inflammatory phenotypes: a cross-sectional study in Northeast China. *Biomed Res Int*. 2018;2018:2860521.
84. Do E, et al. Metabolomic analysis of healthy human urine following administration of glimepiride using a liquid chromatography-tandem mass spectrometry. *Transl Clin Pharmacol*. 2017;25:67.
85. Gromski PS, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*. 2014;4(2):433–52.
86. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*. 2006;2(4):171–96.
87. Steuer AE, Brockbals L, Kraemer T. Metabolomic strategies in biomarker research—new approach for indirect identification of drug consumption and sample manipulation in clinical and forensic toxicology? *Front Chem*. 2019;7:319.
88. Wiklund S, et al. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal Chem*. 2008;80(1):115–22.
89. Kim K, et al. Urine metabolomics analysis for kidney cancer detection and biomarker discovery. *Mol Cell Proteomics*. 2009;8(3):558–70.
90. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, Jia W, Zhao A. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evid-Based Complementary Altern Med*. 2013. <https://doi.org/10.1155/2013/298183>.
91. Liland KH. Multivariate methods in metabolomics: from pre-processing to dimension reduction and statistical analysis. *TrAC Trends Anal Chem*. 2011;30(6):827–41.
92. Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw*. 2002;13(2):415–25.
93. Amaratunga D, Cabrera J, Lee YS. Enriched random forests. *Bioinformatics*. 2008;24(18):2010–4.
94. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
95. Gromski PS, et al. A tutorial review: metabolomics and partial least squares-discriminant analysis: a marriage of convenience or a shotgun wedding. *Anal Chim Acta*. 2015;879:10–23.
96. Riekeberg E, Powers R. New frontiers in metabolomics: from measurement to insight. *F1000Res*. 2017;6:1148.
97. Gomari DP, et al. Variational autoencoders learn transferrable representations of metabolomics data. *Commun Biol*. 2022;5(1):645.
98. Thiel M, Féraud B, Govaerts B. ASCA+ and APCA+: extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J Chemom*. 2017;31(6):e2895.
99. Tian X, et al. Towards enhanced metabolomic data analysis of mass spectrometry image: multivariate curve resolution and machine learning. *Anal Chim Acta*. 2018;1037:211–9.
100. Olsson M, et al. Metabolomics analysis for diagnosis and biomarker discovery of transthyretin amyloidosis. *Amyloid*. 2021;28(4):234–42.
101. Efimenko M, Ignatev A, Koshechkin K. Review of medical image recognition technologies to detect melanomas using neural networks. *BMC Bioinform*. 2020;21(11):270.
102. Perng W, et al. Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. *Pediatr Res*. 2019;85(3):262–8.

103. Vasquez MM, et al. Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. *BMC Med Res Methodol*. 2016;16(1):154.
104. Worley B, Powers R. Multivariate analysis in metabolomics. *Curr Metabol*. 2013;1(1):92–107.
105. Zheng X, et al. Metabolic signature of pregnant women with neural tube defects in offspring. *J Proteome Res*. 2011;10(10):4845–54.
106. Cai H-L, et al. Metabolomic analysis of biochemical changes in the plasma and urine of first-episode neuroleptic-naïve schizophrenia patients after treatment with risperidone. *J Proteome Res*. 2012;11(8):4338–50.
107. Bevilacqua M, Bro R. Can we trust score plots? *Metabolites*. 2020;10(7):278.
108. Triba MN, et al. PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Mol BioSyst*. 2015;11(1):13–9.
109. Szymanska E, et al. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*. 2012;8(Suppl 1):3–16.
110. Eriksson L, Trygg J, Wold S. CV-ANOVA for significance testing of PLS and OPLS® models. *J Chemom*. 2008;22(11–12):594–600.
111. Pepe MS, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882–90.
112. Zhang Y et al. SurvBenchmark: comprehensive benchmarking study of survival analysis methods using both omics data and clinical data. *bioRxiv*, 2021: p. 2021.07.11.451967.
113. Shultz EK. Multivariate receiver-operating characteristic curve analysis: prostate cancer screening as an example. *Clin Chem*. 1995;41(8 Pt 2):1248–55.
114. Rahman MA, et al. LC-HRMS based non-targeted metabolomic profiling of wheat (*Triticum aestivum* L.) under post-anthesis drought stress. *Am J Plant Sci*. 2017;08:3024–61.
115. Tyagi R, et al. Urine metabolomics based prediction model approach for radiation exposure. *Sci Rep*. 2020;10(1):16063.
116. Leave-one-out cross-validation, In: C. Sammut and G.I. Webb (Eds.) *Encyclopedia of Machine Learning*, 2010, Springer US: Boston, MA. p. 600–601.
117. Xu Q-S, Liang Y-Z, Du Y-P. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J Chemom*. 2004;18(2):112–20.
118. Chen T, et al. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evid Based Complement Alternat Med*. 2013;2013: 298183.
119. Sun Y, et al. Metabolomics signatures in type 2 diabetes: a systematic review and integrative analysis. *J Clin Endocrinol Metab*. 2020;105(4):1000.
120. Schmidt JC, et al. Metabolomics as a truly translational tool for precision medicine. *Int J Toxicol*. 2021;40(5):413–26.
121. Yao M, et al. Identification of biomarkers for preeclampsia based on metabolomics. *Clin Epidemiol*. 2022;14:337–60.
122. Lai W, Du D, Chen L. Metabolomics provides novel insights into epilepsy diagnosis and treatment: a review. *Neurochem Res*. 2022;47(4):844–59.
123. Luo J, et al. Human plasma metabolomics identify 9-cis-retinoic acid and dehydrophytylphosphingosine levels as novel biomarkers for early ventricular fibrillation after ST-elevated myocardial infarction. *Bioengineered*. 2022;13(2):3334–50.
124. Pluskal T, et al. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform*. 2010;11:395.
125. Mak TD, et al. MetaboLyzer: a novel statistical workflow for analyzing postprocessed LC–MS metabolomics data. *Anal Chem*. 2014;86(1):506–13.
126. Peters K, et al. PhenoMeNal: processing and analysis of metabolomics data in the cloud. *GigaScience*. 2018;8(2):giy149.
127. Kirpich AS, et al. SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinform*. 2018;19(1):151.
128. R Core Team (R Foundation for Statistical Computing, A., R: A Language and Environment for Statistical Computing. 2013.
129. Goldansaz SA, et al. Livestock metabolomics and the livestock metabolome: a systematic review. *PLoS ONE*. 2017;12(5): e0177675.
130. Trivedi DK, Hollywood KA, Goodacre R. Metabolomics for the masses: the future of metabolomics in a personalized world. *New Horiz Transl Med*. 2017;3(6):294–305.
131. Broadhurst D, et al. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*. 2018;14(6):72.
132. Nezhad MZ et al. SUBIC: A supervised bi-clustering approach for precision medicine. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2017.
133. Yi H, et al. COBRAC: a fast implementation of convex biclustering with compression. *Bioinformatics*. 2021;37(20):3667–9.
134. Tolstikov V, et al. Current status of metabolomic biomarker discovery: impact of study design and demographic characteristics. *Metabolites*. 2020;10(6):224.
135. Munafo MR, et al. A manifesto for reproducible science. *Nat Hum Behav*. 2017;1:0021.
136. Poste G. Bring on the biomarkers. *Nature*. 2011;469(7329):156–7.
137. Kohler I, et al. Integrating clinical metabolomics-based biomarker discovery and clinical pharmacology to enable precision medicine. *Eur J Pharm Sci*. 2017;109:S15–21.
138. Pedersen HK, et al. A computational framework to integrate high-throughput “-omics” datasets for the identification of potential mechanistic links. *Nat Protoc*. 2018;13(12):2781–800.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.