Trends in Analytical Chemistry 178 (2024) 117819

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

journal homepage: www.elsevier.com/locate/trac

ML-based clinical decision support models based on metabolomics data

Michał Burdukiewicz [a,b], Jarosław Chilimoniuk [b], Krystyna Grzesiak [b,c], Adam Krętowski [b], Michał Ciborowski [b,*]

[a] Institute of Biotechnology and Biomedicine, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain
[b] Metabolomics and Proteomics Laboratory, Clinical Research Centre, Medical University of Bialystok, Poland
[c] Mathematical Institute, Faculty of Mathematics and Computer Science, University of Wrocław, Poland

ARTICLE INFO

Keywords:
Bioinformatics
Biomarker discovery
Clinical decision support systems
Clinical utility
Nuclear magnetic resonance
Mass spectrometry
Pathway analysis

ABSTRACT

Machine learning-based clinical decision support models in healthcare emulate clinicians' cognitive processes, leveraging artificial intelligence to analyze intricate medical data. In the era of evidence-based and precision medicine, they are slowly becoming a cornerstone of clinical analysis. In this review, we describe the utilization of machine learning models trained on metabolomics data within the clinical context. We overview the recent advancements in the field, highlight existing challenges and scrutinize intricacies stemming from the unique complexity of the metabolite information.

1. Introduction

One of the hallmarks of modern medicine is its ability to generate large quantities of data. Patients, previously described using mostly subjective assessment, now are represented by humongous datasets like high-resolution imagining or '-omics' information. Here, '-omics' is a comprehensive assessment of a given biological system at a specific organizational level, such as genome, transcriptome, proteome or metabolome [1]. Each organizational level offers unique insight into a patient's well-being and health. The hierarchical nature of these levels is followed by two underlying factors: their reaction speed to external stimuli and their chemical complexity [2]. Metabolomics, defined as a comprehensive study of small molecules with a molecular weight <1500 Da [3], occupies an extreme position in this hierarchy. It reacts swiftly to external stimuli and, at the same time, presents the greatest chemical complexity, containing a diverse set of molecules like nucleic acids, amino acids, peptides, carbohydrates or lipids (Fig. 1A.). This immediacy makes metabolomics essential in medicine for assessing a patient's current state. However, the complexity of metabolomics data significantly impedes its collection and analysis [4,5].

In metabolomics, two primary techniques stand out: mass spectrometry (MS) and nuclear magnetic resonance (NMR). NMR offers a reproducible and high-throughput insight into a limited set of abundant metabolites (roughly 50–150) while ensuring minimal sample destruction and a relatively short sample preparation time [6]. On the other hand, MS, though less suitable for high-throughput studies, lower reproducibility and more extensive sample preparation, theoretically detects thousands of metabolites even in very low concentrations [7,8].

These methods apply both in untargeted and targeted analyses (Fig. 1B.). Untargeted studies capture the broad metabolomic variability of a given biological system and are ideal for hypothesis-generating studies without exhaustive metabolite identification as the primary objective [9]. Given the comprehensive nature of non-targeted metabolomics, one of its main challenges is the assignment of identified metabolites to specific pathways [10]. In addition, these metabolites may also be annotated through analytical chemistry [11].

On the contrary, targeted studies aim to measure the concentrations of predefined metabolites [12]. This method requires prior knowledge of the metabolites of interest (e.g. potential biomarkers), relies on signals specific to those metabolites, and does not aim for global coverage [13]. Despite the smaller scope, metabolite identification still poses a challenge in targeted studies. For example, metabolite interference is a typical case of misannotation [14].

Concluding, non-targeted metabolomics typically provides more comprehensive information, while targeted metabolomics focuses more on quantitative measurements. All combinations of abovementioned approaches and techniques (targeted and untargeted metabolomics utilizing either MS or NMR, Supplementary Table 1) complement each other in clinical applications. Firstly, the untargeted analysis identifies metabolomic biomarkers that could be used to diagnose medical conditions, monitor disease progression, or assess treatment response.

* Corresponding author.
E-mail address: michal.ciborowski@umb.edu.pl (M. Ciborowski).

https://doi.org/10.1016/j.trac.2024.117819
Received 12 November 2023; Received in revised form 19 April 2024; Accepted 13 June 2024
Available online 18 June 2024

Abbreviations

AI	Artificial Intelligence
CD	Continuous Delivery
CDSSs	Clinical Decision Support Systems
CI	Continuous Integration
CVD	Cardiovascular Disease
FCS	Functional Class Scoring
GSEA	Gene Set Enrichment Analysis
LDA	Linear Discriminant Analysis
MAR	Missing at Random
MCAR	Missing Completely at Random
ML	Machine Learning
MNAR	Missing Not at Random
NB	Net Benefit
NNB	Number Needed to Benefit
NPV	Negative Predictive Value
OLS	Ordinary Least Squares
ORA	Over-Representation Analysis
PA	Pathway Analysis
PCA	Principal Component Analysis
PPV	Positive Predictive Value
SOFA	Sequential Organ Failure Assessment
TBI	Traumatic Brain Injury
TPA	Topological Pathway Analysis
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
WGCNA	Weighted Gene Co-expression Network Analysis
XAI	Explainable Artificial Intelligence

Secondly, identified biomarkers are utilized in targeted diagnostic approaches to validate their usefulness further. However, metabolite patterns are notoriously elusive. It is much harder to identify metabolite-based biomarkers because they carry more noise than biomarkers identified by other '-omics' (Fig. 3). Thus, both targeted and untargeted studies require thousands of samples from diverse groups of patients [15].

Analysis of such enormous datasets requires dedicated analytical pipelines and algorithms to detect associations between metabolite patterns and clinically relevant features. Although plenty of existing

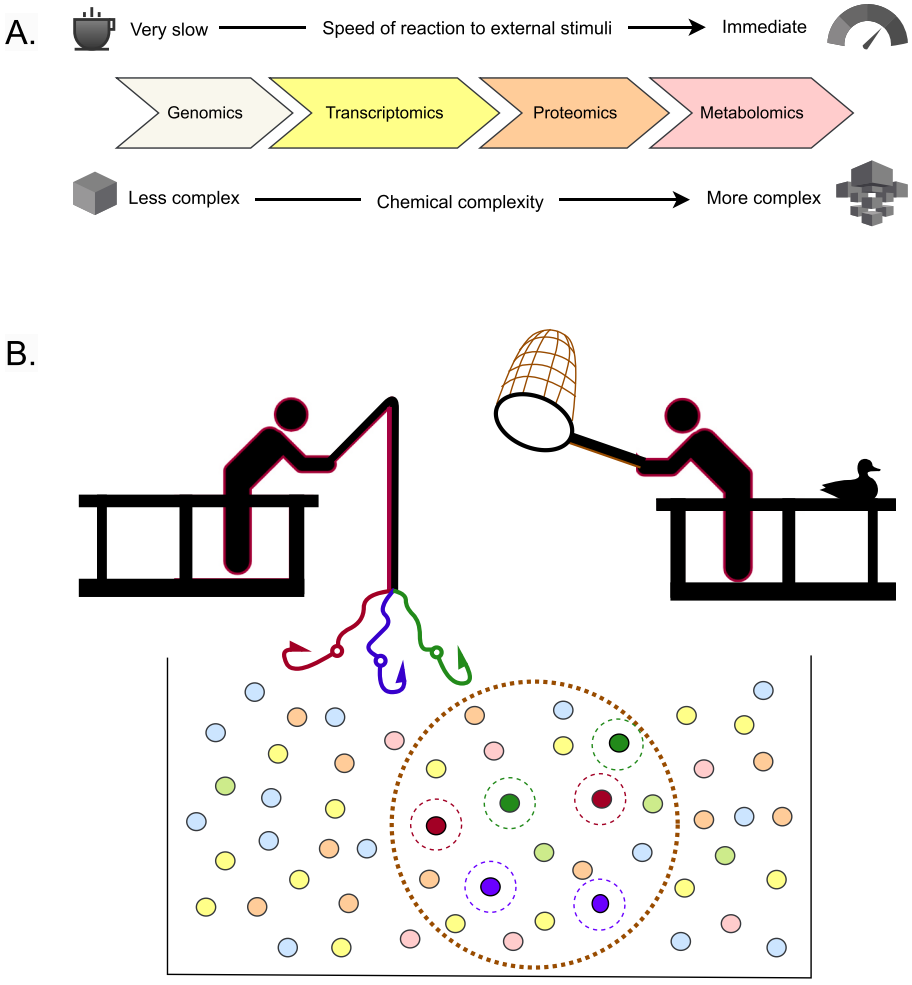

Fig. 1. A. The scheme of reaction speed to external stimuli and the chemical complexity of genomics, transcriptomics, proteomics and metabolomics. Metabolomics has the most immediate reactions to external stimuli while being the most complex to interpret. B. The schematic representation of untargeted and targeted studies. The untargeted studies, here represented as a net, provide an overview of the given biological system (marked by the large brown circle). Targeted studies, represented as a fishing rod, provide insight only into the level of specific metabolites (marked by the smaller red, green and purple circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

solutions belong to classical statistics, the advent of artificial intelligence (AI) altered the landscape of metabolomics by providing access to more intricate models. Some of these models are incorporated into clinical decision support systems (CDSSs), where automated tools for processing medical information enhance a clinician's decision-making [16]. This review aims to summarize machine learning applications, a subfield of AI, in clinic-oriented metabolomics and how they contribute to the novel CDSSs.

2. ML-based clinical decision support models

2.1. Machine learning models

In the most straightforward approach, machine learning (ML) algorithms use sample data to construct intricate mathematical models not only able to identify patterns, but also extrapolate predictive trends, allowing us to draw conclusions. In the case of omics data they are especially used as diagnostic, prognostic or predictive tools revealing the intricate complexities inherent in biological information [17–20].

In particular, machine learning algorithms are employed in clinical metabolomics to analyze complex metabolic data and identify patterns associated with diseases, aiding in early diagnosis and personalized treatment approaches. Here, the data would consist of metabolites measured in biological materials derived from patients, such as blood, urine, or tissue samples (Fig. 2, Samples) and clinical information, including information regarding e.g., known mutation (Fig. 2, Medical interview). Regardless of the input data, we divide machine learning into several techniques depending on the exact learning process, such as unsupervised or supervised, supervised and reinforced learning.

In unsupervised learning, an algorithm uncovers patterns and relationships within the data, frequently aiming to group or organize either patients or metabolites based on inherent similarities or differences. Unsupervised learning does not rely on known output values or explicit feedback during the training process. The two main categories of unsupervised learning are (i) clustering, which groups data points into clusters or segments based on their inherent similarities, (ii) dimensionality reduction, which reduces the complexity of the data by representing it in a lower-dimensional space while retaining essential information [21].

Commonly occurring in clinical decision support, supervised methods take input features describing the patient (which, in the case of metabolomics, contain measured metabolite levels and clinical descriptors) and map it to a specific output. Depending on the type of the output, we recognize (i) classification where output values are categorical (e.g., type of disorder), (ii) regression which fits the data to predict numeric output values (e.g., probability of having a specific disorder) [22].

Although the majority of the tasks fulfilled by CDSSs can be attributed to either classification or regression, some specialized sub-tasks call for more dedicated tools. Survival analysis, for instance, employs statistical methods to predict the time until an event of interest (like death, failure, relapse, or any other specific outcome) occurs [23]. Although simple regression models could address this issue, they fail to fully capture the nuances hidden in variability related to time and individual attributes of patients and require more strict assumptions about data. Moreover, survival analysis accommodates censored data where the event of interest might not have occurred for some individuals within the considered time period.

Metabolomics, given its immediate reaction to external agents (e.g., drug administration), is well-suited for monitoring a patient's status. However, due to the high dimensionality of the metabolomics data, integrating it with classical algorithms for survival analysis is very challenging. Despite this, certain algorithms, like Random Survival Forest have demonstrated the ability to produce reliable outcomes based on metabolomics data [24].

The important caveat is that machine learning is deeply ingrained in modern metabolomics. For example, machine learning-based methods are commonly used in metabolite identification during spectrum processing [25] (Fig. 2, Spectrum processing). Nevertheless, as this step does not directly pertain to developing the CDSS, it is not discussed in this review.

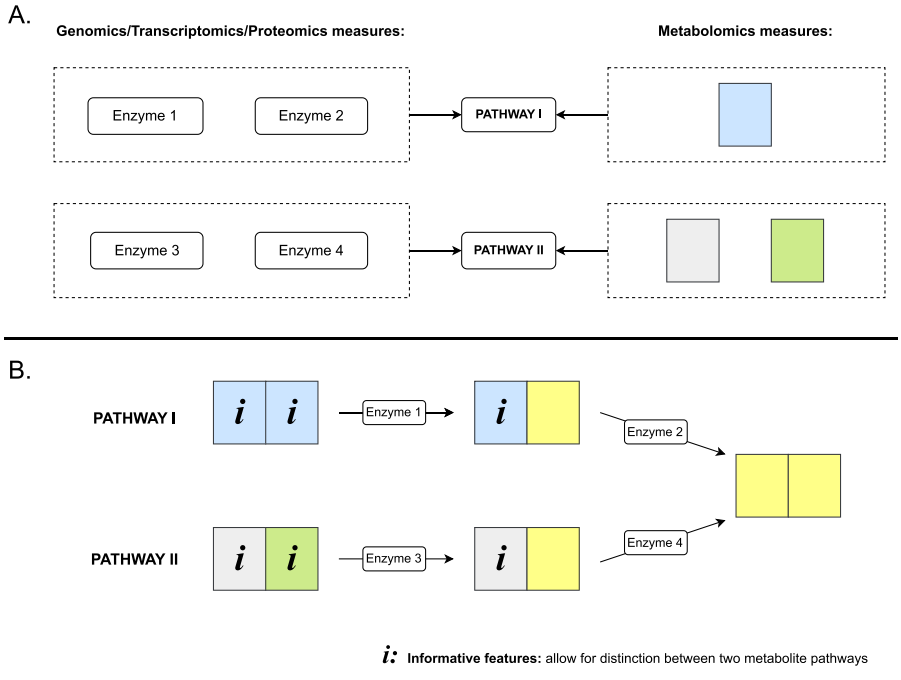

Fig. 2. The design scheme of clinical predictive model design. Analytical techniques (NMR or MS) process samples taken from patients. Next, the processing of spectral data leads to the abundance of metabolites. Finally, machine learning algorithms train using metabolite abundance and clinical information.

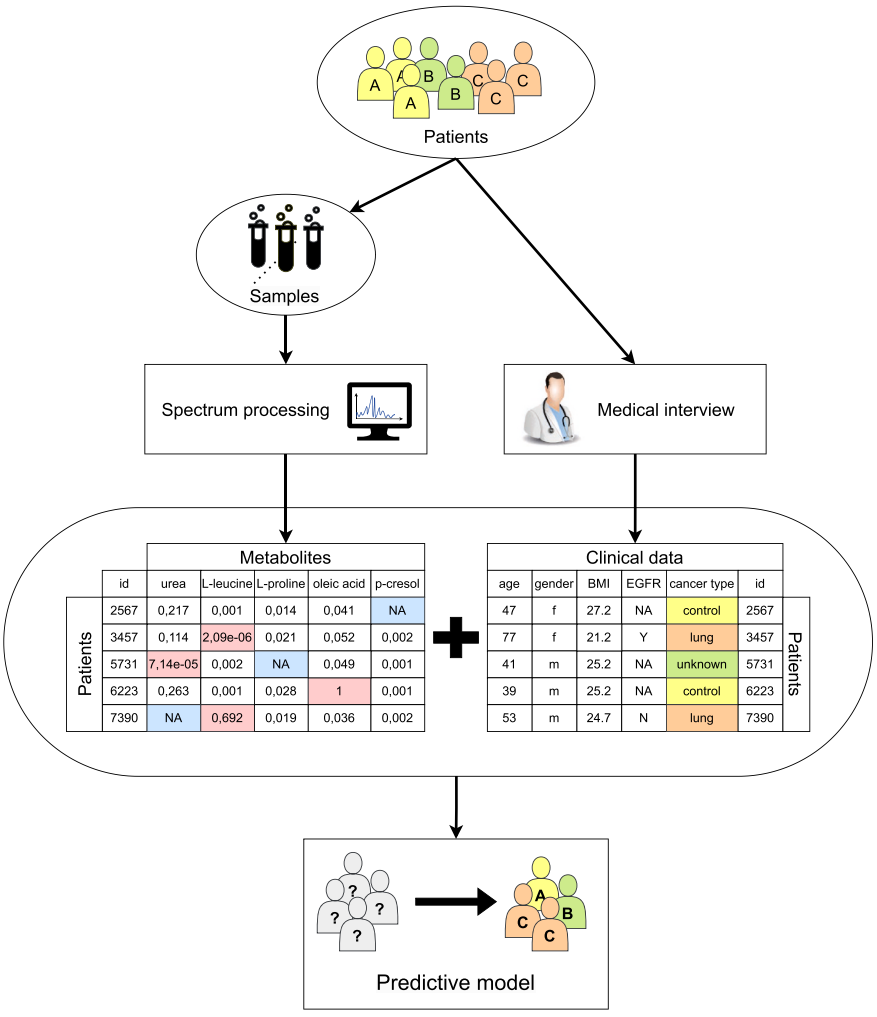

Fig. 3. The intricate elusiveness of metabolomics data. A: Genomics/transcriptomics/proteomics provide evidence on the level of enzymes in the biological pathway. Due to limitations in metabolite identification, only levels of fragments of metabolites (blue, grey and green boxes) are measured. B: Presentation of both pathways. The measurements of the yellow metabolite provide no information that discerns between the two pathways. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2.2. *Dimensionality of metabolomics data*

The first of the challenges in analyzing metabolomics data is its dimensionality. This problem pertains not only to the large number of features (Fig. 2, 'Metabolites' and 'Clinical Data') but to the lack of balance between the dimensions of the dataset. Data generated by metabolomics often suffers from the problem of 'big p, small n' (or '$p \gg n$'), defined as the smaller number of samples compared to a large number of features [26]. In such cases, classical statistical models like ordinary least squares (OLS) fail to provide a unique solution for estimating model coefficients due to the imbalance between predictors and observations. This imbalance hampers model performance, affecting both prediction quality and stability, as it was proved in Linear Mixed Models simulation studies [27]. Additionally, as the variable space increases, the distances between observations in the multidimensional space also grow, which makes the data too sparse to capture underlying trends. Furthermore, large feature space leads to the inclusion of excessive irrelevant variables, which exacerbates these issues, causing the curse of dimensionality [28] and overfitting [29], where the model captures noise in training data rather than the genuine signal.

This problem is especially profound in untargeted studies because they yield up to thousands of identified metabolites [30]. However, even targeted methods (e.g., widely targeted metabolomics) can measure hundreds of metabolites related to a significantly smaller number of samples [31]. Thus, the subsequent analysis often requires dimensionality reduction through feature selection or feature engineering (extraction). The former means choosing the subset of features that turned out to be the most informative, while the latter refers to generating a smaller set of new features by transforming or combining the original ones.

2.2.1. *Feature selection*

There are three main classes of feature selection methods: filter, wrapper and embedded [32]. Most filter (screening) methods consider each feature separately, looking for its relevance to the label of the data (in the case of classification) or its correlation to the continuous output. Although these methods are not computationally intensive and scale linearly with the number of features, they may ignore information hidden under the interactions. At the same time, considering the features and their interactions jointly yields larger feature space and makes searching for informative features even more complex. Examples of such methods are methods based on the information gain as implemented in the lilikoi R package [33] or MetaboAnalyst [34]. They offer a wide array of filtering methods (e.g. ANOVA, Chi-squared test, Fisher's exact test, Pearson correlation, Mann-Whitney *U* test or *t*-test) that are widely used in the field [35].

Wrapper methods allow for selecting the subset of features optimal for model setting. Essentially, they assess the quality of prediction (or

classification) of a plethora of models built on multiple different feature subsets to include the effect of the relationships between different metabolites. However, they are more prone to overfitting and computationally intensive than filters. An example of such method is Boruta, which is based on the Random Forest algorithm. Briefly, it iteratively eliminates variables less pertinent than random features [36]. Another example of a wrapper dedicated to metabolomics data is a pipeline combining neural networks and genetic algorithms [37].

In contrast to wrapper methods, embedded methods utilize a single model for feature selection and prediction (or classification). Consequently, the feature selection is deeply ingrained in the very design of model development. As a result, these methods inherently capture the effects of relationships between features. Examples of embedded methods for metabolomics data are regularized regression, tree-based methods, support vector machines [38] as well as random forests [39] or Lasso-based regularization methods like HSIC Lasso [40].

Instead of relying on a single filter or wrapper, it is possible to propose a solution based on the ensemble of feature filtering methods. Even in cases involving small cohorts of critically ill patients, the ensemble method successfully identifies a set of crucial metabolites, showcasing its resilience and efficacy [41].

2.2.2. Feature engineering

Feature engineering involves generating new features by transforming (manipulating) or combining the original ones to preserve essential information [42]. In metabolomics, as well as in other omics sciences, a highly popular example of feature engineering is normalization. Normalization is crucial in metabolomics to address unwanted sample-to-sample variations coming from different sample concentrations or instrumental drift [43]. Common normalization methods include total sum normalization, internal or external standard metabolites normalization and probabilistic quotient normalization [44,45].

Another aspect of feature engineering involves data transformations (scalings) aimed at aligning with the assumptions of various statistical tests. In metabolomics, numerous transformation methods serve three primary purposes: addressing heteroscedasticity, normalizing skewed distributions, and converting multiplicative metabolite relationships into additive ones [46].

Feature engineering also encompasses feature extraction techniques. PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are the most popular feature extraction methods. Their commonness in metabolomics led to algorithms tailored specifically to this data type [47,48]. However, the downside of these methods, examples of unsupervised learning, is that the newly created features do not offer any interpretability, as they do not have any real-world meanings (please refer to section 2.3 regarding the importance of the model interpretability).

Other feature engineering methods leverage *a priori* biological knowledge to create more representative features from specific metabolites. For example, the weighted gene co-expression network analysis (WGCNA) algorithm, adapted to metabolomics data [49], groups metabolites using both their co-abundance and data from the KEGG Pathway Database. Contrary to the features produced by methods like PCA or LDA, these clusters maintain interpretability as they embody distinct metabolomic pathways. This approach was included, among others, in a framework for integrative analysis of '-omics' data combining metabolome and microbiome information [50].

However, some methods to generate such sparse and knowledge-driven representations do not require co-abundance data and rely only on the information from metabolomic pathways. These features also retain some biological interpretability while resisting the noise associated with the abundance [51]. While widely applied in other 'omics' fields [52], this method may pose challenges due to the noise resulting from misidentifying metabolites or their low information content (see section 3.1).

2.3. Assessment of the model quality

In traditional machine learning, metrics like accuracy or AUC (area under the receiver operating characteristic curve) are standard measures of prediction quality. However, these measures are less practical or may even be harmful in the clinical environment. The assessment of machine learning-based models for clinical data hinges on evaluating their clinical utility, a critical aspect in healthcare applications.

In the context of classification-based CDSSs, clinical utility refers to the system's ability to provide meaningful and actionable insights to healthcare professionals, ultimately leading to improved patient outcomes. It encompasses the model's accuracy in diagnosing or predicting medical conditions and its potential to influence treatment decisions, reduce healthcare costs, or enhance the overall patient care experience. Clinical utility assessments often employ the model's sensitivity and specificity, which are essential for understanding its diagnostic capabilities [53].

False positives and false negatives play a pivotal role in evaluating these models. False positives occur when a model incorrectly identifies a patient as having a particular condition when they do not. In a clinical setting, false positives can lead to unnecessary medical tests, treatments, or patient stress, potentially increasing healthcare costs and causing anxiety. Conversely, false negatives occur when a model fails to identify a patient with a condition, potentially delaying diagnosis and treatment. Misclassification in both scenarios can lead to tragic consequences. At the same time, striking a balance between minimizing false positives and false negatives constitutes a challenging and responsible aspect of model assessment [54]. Common metrics in this domain include positive predictive value (PPV), reflecting the likelihood of accurately detecting a condition, and negative predictive value (NPV), denoting the probability of correctly rejecting a condition. Moreover, the assessment can be done using the clinical utility measures such as intention to treat estimation, which is an emulated trial [55], net benefit (NB) utilizes the exchange rate of benefits and harms gained by applying the model [56], or the number needed to benefit (NNB), which represents the number of people needed to screen to find those that need treatment [57].These metrics provide additional weights to wrong predictions (false positives and false negatives).

The development of each clinical model should follow the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines. It consists of a 27-item checklist, that helps to improve the reporting quality of studies that develop, validate, or update prediction models [58,59]. Moreover, fulfilling all the requirements by the model should allow them to be compared with each other.

Furthermore, assessing ML-based models in clinical data requires ongoing monitoring and adjustment. As real-world patient populations and clinical practices evolve, these models need to adapt and adjust. Establishing feedback loops with healthcare professionals and experts becomes vital for refining model performance, addressing concerns related to false positives and false negatives, and ensuring the technology consistently delivers meaningful clinical utility [60]. To at least partially automatize these feedback loops, CDSSs should be deployed using the Continuous Integration (CI)/Continuous Delivery (CD) framework, where the impact of ongoing updates on the quality of predictions is consistently monitored and evaluated [61].

Unfortunately, one of the biggest drawbacks of the machine-learning models is the lack of interpretability, making it challenging to explain their predictions or which features are important. It is especially pronounced in a high-stake environment as clinics, where a single decision often decide about human well-being or even life. Here, the entire field of Explainable Artificial Intelligence (XAI) tries to provide human-understandable explanations of decisions made by black-box models. The usage of XAI in with CDSS would help to assess if the predictions are made using right premises, ensuring that the model works as intended [62].

3. Challenges and chances in the development of CDSSs

3.1. Pathway analysis as the independent conformation of CDSSs

Pathway analysis (PA) in metabolomics is pivotal role in unraveling the complex web of interconnections among metabolites within biological systems. The concentration levels of metabolites within cells are intricately linked to the underlying metabolic network structure, giving rise to various correlations among these molecular compounds [63]. However, it is important to note that these correlations often defy the simple assumption that strongly correlated metabolites are direct neighbors within the metabolic network. Metabolites that are physically or temporally distant may exhibit significant correlations, while metabolites in close proximity may not display such strong relationships [64].

PA methods can be divided into three classes. The first one, over-representation analysis (ORA) pinpoints pathways or groups of metabolites that exhibit greater similarities with a specific set of molecules of interest than would be anticipated by random chance [65]. To verify that, ORA tests (usually using Fisher's exact test) if a concentration of metabolites in a given sample is significantly different than the frequency in the reference set of metabolites [66]. Since this framework involves multiple testing, its results require post-processing, either in the form of p-value adjustment or computation of q-values. Thus, ORA is susceptible to variability of metabolomics, where some techniques identify broader sets of metabolites, because a different amount of detected compounds could severely alter the final result. Moreover, changing a reference set of metabolites for a given sample can greatly change identified pathways. Lastly, the post-processing of ORA results often involves arbitrary decisions, such as setting the thresholds for significant associations [67].

In the case of functional class scoring (FCS), instead of considering independently every metabolite, the analysis is performed on statistics representing pathways. Thus, FCS methods, like GSEA [68], could detect many small alterations between numerous metabolites in a single pathway, instead of focusing only on large changes like ORA. Obviously, this transformation hampers inference based on single metabolites, which could be beneficial for tasks like biomarker identification. Moreover, FCS methods need to correct the noise that might be introduced by considering at the same time numerous signals from many metabolities [69].

The last class, topological pathway analysis (TPA), includes topological measures of the metabolites in the statistics, considering their non-equivalence within the path [70]. Here it can be seen as a variant of FCS, where instead of treating all metabolites equally, we weight them according to their importance in the pathway. Since pathway can be represented as networks this importance is usually considered to be a some measure of centrality [71]. The precise pathway representation is critical, because TPA methods, like MetPA [71], are more vulnerable to incomplete pathway data than FCS approaches.

One of the benefits of the pathway analysis is its ability to validate the metabolomics-based CDSSs independently. Here, knowledge from PA could confirm that metabolites considered by the model are indeed associated with a specific phenomenon. For example, metabolites identified by binary logistic regression as biomarkers for therapeutic interventions in Parkinson's disease were independently confirmed using PA to accelerate the self-assembly of alpha-synuclein, a protein associated with the onset and progress of this disorder [72].

3.2. Reliability of machine learning methods

While the potential of machine learning to provide highly accurate automated diagnoses is promising, the current lack of robust scientific evidence regarding its clinical benefits is a significant concern. Despite the hype surrounding it, a systematic review of deep learning's performance in disease diagnosis using medical images compared to health professionals reveals multiple issues [73]. Notably, only a small fraction of studies have evaluated algorithm performance in external cohorts, which means that they do not assess the capacity of these algorithms to analyze previously unseen data [74]. Moreover, they suffer from methodological flaws and unrealistic evaluation conditions, for example, very small and homogenous test data [75]. Unfortunately, many proposals to automate clinical diagnoses with machine learning remain unvalidated, raising concerns as these models are frequently adopted without rigorous assessment [76].

Furthermore, machine learning's strong emphasis on predictive performance often overshadows the potential advantages of utilizing simple clinical variables. Research indicates that clinical information can significantly enhance predictions based on high-dimensional data. Additionally, there seems to be a shift towards classifying patients into binary categories rather than predicting a continuum of risk, which may not align with the intended goal of this technology [77]. However, the discretization of data streamlines computing critical measures of the model quality as the clinical utility (Section 2.4).

3.3. Causal inference

In precision medicine, machine learning promises to tailor medical treatments to individuals. However, achieving this objective presents substantial challenges for two fundamental reasons. First, machine learning struggles with causal inference, as it cannot establish cause and effect without making specific assumptions. Causal inference necessitates a well-defined causal question, the ability to identify the causal effect and observe all relevant treatment options among all interest groups. These assumptions frequently require human input or true artificial intelligence, making it challenging for machine learning alone to identify the optimal treatment for an individual. For example, the human understanding of the causality of a given phenomenon is necessary to properly select potential covariates while excluding spurious associations [7]. Growing interest in developing methods that can infer causal relationships into decision-making processes, remains an intense point of ongoing research. Presently, several approaches have emerged, including Mendelian randomization, Bayesian networks and probabilistic graphical models, which encode causal relationships within graphical structures or generalized random forests [78].

The second challenge is the inherent complexity and probabilistic nature. This complexity makes it impossible to predict individual outcomes with certainty. Statistics, including machine learning, excel at understanding and comparing probabilities between groups but cannot predict individual events with absolute certainty. The probabilistic nature of health processes implies that individual outcomes will always contain an element of chance, irrespective of the dataset's size or the model's complexity [79]. Addressing this problem requires prioritizing performance measures that involve risk assessment (e.g., clinical utility, described in section 2.3).

3.4. Accuracy of ML-based diagnosis in the context of noise in metabolomics data

In the realm of untargeted metabolomics, the precise management of noise within datasets is of utmost importance, especially when assessing the accuracy of machine learning-based diagnosis models. Liquid chromatography coupled with mass spectrometry (LC-MS) is the preferred analytical technique, renowned for its exceptional sensitivity and comprehensive coverage of small molecular species [80]. Nonetheless, datasets generated by this method often contain numerous detected features that can present significant challenges. Global profiling of metabolites using a high-resolution mass spectrometer may result in a set of even ten thousand features [81]. These features frequently lack informative value and can originate from background signals or procedural artifacts introduced during sample processing [82]. Moreover, different adducts and in-source fragments are also detected and need to be

identified and grouped into a single representative feature *per* analyte. This computationally intensive process is crucial from the point of view of further data analysis steps [83]. The presence of artificial features in the data set will not only affect the statistical outcome but may also provide misleading information if such a feature is improperly identified as a real biological molecule. Over the years, signal processing has been largely automated, and multiple software tools for LC-MS metabolomics data processing have been developed, *e.g.*, XCMS, mzMine, OpenMS or MS-DIAL [81]. However, features generated by different tools may not overlap even by more than 50 % [84]. Recently, new tools for data processing have appeared, such as SLAW [85] or Asari [81], which has been included in the newest 6.0 version of MetaboAnalyst [86].

Furthermore, the inherent complexities and imperfections in feature detection and integration, which are crucial for machine learning algorithms, exacerbate the problem by introducing spurious or inadequately integrated features. The consequences of insufficient noise filtration should not be underestimated, as the simulation study indicated that it leads to improper estimation of longitudinal effects [87]. This is crucial, particularly in the context of machine learning-based diagnosis, as it could undermine the accuracy and reliability of diagnostic models [88].

In addition, missing values are another indirect noise source in metabolomics data. Missing values come from various sources, such as technical errors during sample preparation, instrument malfunction, or the inherent limitations of detection or quantification thresholds impacting the reliability and comprehensiveness of analysis. Depending on their origin, missing values are categorized into three schemes: Missing Completely at Random (MCAR), where missing values occur randomly without any relation to observed or unobserved data; Missing at Random (MAR), where the probability of a value being missing depends only on observed data; and Missing Not at Random (MNAR), where the probability of a value being missing depends on unobserved data or the value itself [89]. Since missing values mostly don't allow for subsequent multivariate statistical inference [90], researchers have adopted various strategies for handling missing values, including imputation techniques, aimed at preserving the integrity of the overall dataset.

While simple methods like mean imputation offer ease of implementation, they may introduce bias, particularly when data is not MCAR. On the other hand, more sophisticated approaches such as regression or K-nearest neighbors imputation capture complex relationships but can be sensitive to outliers. Multiple imputation stands out as a robust solution by generating multiple datasets and considering conditional distributions, albeit requiring careful consideration of assumptions about the missing data mechanism and data distribution. Deep learning-based imputation methods hold promise, although they often demand large datasets and computational resources.

Thus, the choice of imputation method is a highly complex issue that depends on multiple factors. At the same time, one shouldn't assume that the sole purpose of imputation is just to fill in missing values; rather, we should view it as an integral part of statistical analysis and select a method that harmonizes well with other steps of analysis and leads to the most reliable statistical results possible. The optimal method should be tailored to the data structure, missing data mechanism, and imputation objectives. For instance, conditional mean imputation might be suitable for quantitative data with relatively stable distributions, while imputation via conditional distribution sampling is preferable for data with complex dependencies among variables and high variability in general. Currently, 41 methods for imputing missing values in metabolomics data are available through the web server and R package called imputomics [5].

4. Conclusion

Combining clinical data gathered by medical interviews with metabolite levels has already led to the development of CDSSs. One such system used serum metabolites with clinical indices, which could effectively discriminate between TBI patients who reported good and poor outcomes [91]. Another is the SCORE2 algorithm, which estimates the 10-year risk of fatal and non-fatal cardiovascular disease (CVD) in individuals aged 40–69 without prior CVD or diabetes in Europe, enhancing the identification of individuals at higher risk [92]. Moreover, widely used clinical evaluation systems, sequential organ failure assessment (SOFA) score and the acute physiology and chronic health evaluation II (APACHE II) score, combined with metabolomics data, have also proven to be more effective in the identification of critically ill patients who are at an increased risk of mortality due to sepsis infection [93]. It shows that combining metabolomics data with clinical data in prediction models can be an excellent strategy to improve the ability of anatomical-based scoring systems to identify patients at risk.

It is crucial to understand that CDSSs based on machine learning come with risks. Without evaluating their clinical utility, these machine-learning models bring little benefit in practical applications. Moreover, the lack of data analysis techniques tailored to the unique structure of metabolomics data prevents the development of well-performing models. Therefore, we hope that our brief presentation of the current state-of-the-art (consult Supplementary Table 2 for summary of mentioned tools or tools containing described methods) will stimulate the usage of metabolomics-based CDSSs as this branch of '-omics' still has yet to unravel its full potential.

CRediT authorship contribution statement

Michał Burdukiewicz: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Jarosław Chilimoniuk:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Krystyna Grzesiak:** Methodology, Visualization, Writing – original draft, Writing – review & editing. **Adam Krętowski:** Supervision, Writing – review & editing. **Michał Ciborowski:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The research was supported by Polish Ministry of Science and Higher Education within the project "Excellence Initiative–Research University" and Medical University of Białystok grant (B.SUB.24.548 to M.C.). We would like to express our deep appreciation to the Metabolomics and Proteomics Laboratory (Clinical Research Centre, Medical University of Białystok) for valuable discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trac.2024.117819.

References

[1] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome Biol. 18 (2017) 83, https://doi.org/10.1186/s13059-017-1215-1.
[2] X. Dai, L. Shen, Advances and trends in omics technology development, Front. Med. 9 (2022).

[3] P. Miggiels, B. Wouters, G.J.P. van Westen, A.-C. Dubbelman, T. Hankemeier, Novel technologies for metabolomics: more for less, TrAC Trends Anal. Chem. 120 (2019) 115323, https://doi.org/10.1016/j.trac.2018.11.021.
[4] G.J. Patti, O. Yanes, G. Siuzdak, Metabolomics: the apogee of the omics trilogy, Nat. Rev. Mol. Cell Biol. 13 (2012) 263–269, https://doi.org/10.1038/nrm3314.
[5] J. Chilimoniuk, K. Grzesiak, J. Kała, D. Nowakowski, A. Krętowski, R. Kolenda, M. Ciborowski, M. Burdukiewicz, Imputomics: web server and R package for missing values imputation in metabolomics data, Bioinformatics 40 (3) (2024), https://doi.org/10.1093/bioinformatics/btae098 btae098.
[6] J.L. Markley, R. Brüschweiler, A.S. Edison, H.R. Eghbalnia, R. Powers, D. Raftery, D.S. Wishart, The future of NMR-based metabolomics, Curr. Opin. Biotechnol. 43 (2017) 34–40, https://doi.org/10.1016/j.copbio.2016.08.001.
[7] K.F. Arnold, V. Davies, M. de Kamps, P.W.G. Tennant, J. Mbotwa, M.S. Gilthorpe, Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning, Int. J. Epidemiol. 49 (2020) 2074–2082, https://doi.org/10.1093/ije/dyaa049.
[8] S. Alseekh, A. Aharoni, Y. Brotman, K. Contrepois, J. D'Auria, J. Ewald, J.C. Ewald, P.D. Fraser, P. Giavalisco, R.D. Hall, M. Heinemann, H. Link, J. Luo, S. Neumann, J. Nielsen, L. Perez de Souza, K. Saito, U. Sauer, F.C. Schroeder, S. Schuster, G. Siuzdak, A. Skirycz, L.W. Sumner, M.P. Snyder, H. Tang, T. Tohge, Y. Wang, W. Wen, S. Wu, G. Xu, N. Zamboni, A.R. Fernie, Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices, Nat. Methods 18 (2021) 747–756, https://doi.org/10.1038/s41592-021-01197-1.
[9] X. Zhang, X. Zhu, C. Wang, H. Zhang, Z. Cai, Non-targeted and targeted metabolomics approaches to diagnosing lung cancer and predicting patient prognosis, Oncotarget 7 (2016) 63437–63448, https://doi.org/10.18632/oncotarget.11521.
[10] K. Beebe, A.D. Kennedy, Sharpening precision medicine by a Thorough Interrogation of metabolic Individuality, Comput. Struct. Biotechnol. J. 14 (2016) 97–105, https://doi.org/10.1016/j.csbj.2016.01.001.
[11] N. Vinayavekhin, A. Saghatelian, Untargeted metabolomics, Curr. Protoc. Mol. Biol. 90 (30) (2010), https://doi.org/10.1002/0471142727.mb3001s90, 1.1-30.1.24.
[12] W. Lu, B.D. Bennett, J.D. Rabinowitz, Analytical strategies for LC–MS-based targeted metabolomics, J. Chromatogr. B, Hyphenated Techniques for Global Metabolite Profiling 871 (2008) 236–242, https://doi.org/10.1016/j.jchromb.2008.04.031.
[13] U. Vrhovsek, D. Masuero, M. Gasperotti, P. Franceschi, L. Caputi, R. Viola, F. Mattivi, A Versatile targeted metabolomics method for the rapid quantification of multiple classes of Phenolics in Fruits and Beverages, J. Agric. Food Chem. 60 (2012) 8831–8840, https://doi.org/10.1021/jf2051569.
[14] Z. Jia, Q. Qiu, R. He, T. Zhou, L. Chen, Identification of metabolite interference is necessary for accurate LC-MS targeted metabolomics analysis, Anal. Chem. 95 (2023) 7985–7992, https://doi.org/10.1021/acs.analchem.3c00804.
[15] R.D. Beger, W. Dunn, M.A. Schmidt, S.S. Gross, J.A. Kirwan, M. Cascante, L. Brennan, D.S. Wishart, M. Oresic, T. Hankemeier, D.I. Broadhurst, A.N. Lane, K. Suhre, G. Kastenmüller, S.J. Sumner, I. Thiele, O. Fiehn, R. Kaddurah-Daouk, For "precision medicine and Pharmacometabolomics task Group"-metabolomics Society Initiative, 2016. Metabolomics enables precision medicine: "A white paper, Community perspective.", Metabolomics 12 (2016) 149, https://doi.org/10.1007/s11306-016-1094-6.
[16] S. Medlock, J.C. Wyatt, V.L. Patel, E.H. Shortliffe, A. Abu-Hanna, Modeling information flows in clinical decision support: key insights for enhancing system effectiveness, J. Am. Med. Inform. Assoc. 23 (2016) 1001–1006, https://doi.org/10.1093/jamia/ocv177.
[17] N.-Q.-K. Le, T.-T.-D. Nguyen, Y.-Y. Ou, Identifying the molecular functions of electron transport proteins using radial basis function networks and biochemical properties, J. Mol. Graph. Model. 73 (2017) 166–178, https://doi.org/10.1016/j.jmgm.2017.01.003.
[18] N.-Q.-K. Le, Y.-Y. Ou, Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins, BMC Bioinf. 17 (2016) 501, https://doi.org/10.1186/s12859-016-1369-y.
[19] P.S. Reel, S. Reel, E. Pearson, E. Trucco, E. Jefferson, Using machine learning approaches for multi-omics data analysis: a review, Biotechnol. Adv. 49 (2021) 107739, https://doi.org/10.1016/j.biotechadv.2021.107739.
[20] H.V. Thi, T.-N. Hoang, N.Q.K. Le, D.-T. Chu, Application of data science and bioinformatics in RNA therapeutics, Prog. Mol. Biol. Transl. Sci. 203 (2024) 83–97, https://doi.org/10.1016/bs.pmbts.2023.12.019.
[21] P. Meinicke, T. Lingner, A. Kaever, K. Feussner, C. Göbel, I. Feussner, P. Karlovsky, B. Morgenstern, Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps, Algorithms Mol. Biol. 3 (2008) 9, https://doi.org/10.1186/1748-7188-3-9.
[22] T. Ghosh, W. Zhang, D. Ghosh, K. Kechris, Predictive modeling for metabolomics data, in: S. Li (Ed.), Computational Methods and Data Analysis for Metabolomics, Methods in Molecular Biology, Springer US, New York, NY, 2020, pp. 313–336, https://doi.org/10.1007/978-1-0716-0239-3_16.
[23] R. Singh, K. Mukhopadhyay, Survival analysis in clinical trials: Basics and must know areas, Perspect. Clin. Res. 2 (2011) 145, https://doi.org/10.4103/2229-3485.86872.
[24] S. Dietrich, A. Floegel, M. Troll, T. Kühn, W. Rathmann, A. Peters, D. Sookthai, M. von Bergen, R. Kaaks, J. Adamski, C. Prehn, H. Boeing, M.B. Schulze, T. Illig, T. Pischon, S. Knüppel, R. Wang-Sattler, D. Drogan, Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis, Int. J. Epidemiol. 45 (2016) 1406–1420, https://doi.org/10.1093/ije/dyw145.
[25] K. Dührkop, M. Fleischauer, M. Ludwig, A.A. Aksenov, A.V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, S. Böcker, Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information, Nat. Methods 16 (2019) 299–302, https://doi.org/10.1038/s41592-019-0344-8.
[26] J. Chen, Z. Chen, EXTENDED BIC FOR SMALL-n-LARGE-P SPARSE GLM, Stat. Sin. 22 (2012) 555–574.
[27] R. Hyde, L. O'Grady, M. Green, Stability selection for mixed effect models with large numbers of predictor variables: a simulation study, Prev. Vet. Med. 206 (2022) 105714, https://doi.org/10.1016/j.prevetmed.2022.105714.
[28] E. Keogh, A. Mueen, Curse of dimensionality, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning and Data Mining, Springer, US, Boston, MA, 2017, pp. 314–315, https://doi.org/10.1007/978-1-4899-7687-1_192.
[29] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12, https://doi.org/10.1021/ci0342472.
[30] A.C. Schrimpe-Rutledge, S.G. Codreanu, S.D. Sherrod, J.A. McLean, Untargeted metabolomics strategies—challenges and emerging Directions, J. Am. Soc. Mass Spectrom. 27 (2016) 1897–1905, https://doi.org/10.1007/s13361-016-1469-y.
[31] J. Liu, Z.-B. Li, Q.-Q. Lu, Y. Yu, S.-Q. Zhang, P.-F. Ke, F. Zhang, J.-C. Li, Metabolite profile of COVID-19 revealed by UPLC-MS/MS-based widely targeted metabolomics, Front. Immunol. 13 (2022).
[32] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517, https://doi.org/10.1093/bioinformatics/btm344.
[33] F.M. AlAkwaa, B. Yunits, S. Huang, H. Alhajaji, L.X. Garmire, Lilikoi: an R package for personalized pathway-based classification modeling using metabolomics data, GigaScience 7 (2018) giy136, https://doi.org/10.1093/gigascience/giy136.
[34] Z. Pang, J. Chong, G. Zhou, D.A. de Lima Morais, L. Chang, M. Barrette, C. Gauthier, P.-É. Jacques, S. Li, J. Xia, MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights, Nucleic Acids Res. 49 (2021) W388–W396, https://doi.org/10.1093/nar/gkab382.
[35] J. Xia, D.I. Broadhurst, M. Wilson, D.S. Wishart, Translational biomarker discovery in clinical metabolomics: an introductory tutorial, Metabolomics 9 (2013) 280–299, https://doi.org/10.1007/s11306-012-0482-9.
[36] J. Labory, E. Njomgue-Fotso, S. Bottini, Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data, Comput. Struct. Biotechnol. J. 23 (2024) 1274–1287, https://doi.org/10.1016/j.csbj.2024.03.016.
[37] A. Lisitsyna, F. Moritz, Y. Liu, L. Al Sadat, H. Hauner, M. Claussnitzer, P. Schmitt-Kopplin, S. Forcisi, Feature selection pipelines with classification for non-targeted metabolomics combining the neural network and genetic algorithm, Anal. Chem. 94 (2022) 5474–5482, https://doi.org/10.1021/acs.analchem.1c03237.
[38] J. Hageman, Chapter 10 - relevant metabolites' selection strategies, in: J. Troisi (Ed.), Metabolomics Perspectives, Academic Press, 2022, pp. 381–398, https://doi.org/10.1016/B978-0-323-85062-9.00010-6.
[39] A. Godlewski, M. Czajkowski, P. Mojsak, T. Pienkowski, W. Gosk, T. Lyson, Z. Mariak, J. Reszec, M. Kondraciuk, K. Kaminski, M. Kretowski, M. Moniuszko, A. Kretowski, M. Ciborowski, A comparison of different machine-learning techniques for the selection of a panel of metabolites allowing early detection of brain tumors, Sci. Rep. 13 (2023) 11044, https://doi.org/10.1038/s41598-023-38243-1.
[40] Y. Takahashi, M. Ueki, M. Yamada, G. Tamiya, I.N. Motoike, D. Saigusa, M. Sakurai, F. Nagami, S. Ogishima, S. Koshiba, K. Kinoshita, M. Yamamoto, H. Tomita, Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection, Transl. Psychiatry 10 (2020) 1–12, https://doi.org/10.1038/s41398-020-0831-9.
[41] L.B. Kosyakovsky, E. Somerset, A.J. Rogers, M. Sklar, J.R. Mayers, A. Toma, Y. Szekely, S. Soussi, B. Wang, C.-P.S. Fan, R.M. Baron, P.R. Lawler, Machine learning approaches to the human metabolome in sepsis identify metabolic links with survival, Intensive Care Med. Exp 10 (2022) 24, https://doi.org/10.1186/s40635-022-00445-8.
[42] I. Guyon, A. Elisseeff, An introduction to feature extraction, in: I. Guyon, M. Nikravesh, S. Gunn, L.A. Zadeh (Eds.), Feature Extraction: Foundations and Applications, Studies in Fuzziness and Soft Computing, Springer, Berlin, Heidelberg, 2006, pp. 1–25, https://doi.org/10.1007/978-3-540-35488-8_1.
[43] A.M. De Livera, D.A. Dias, D. De Souza, T. Rupasinghe, J. Pyke, D. Tull, U. Roessner, M. McConville, T.P. Speed, Normalizing and integrating metabolomics data, Anal. Chem. 84 (2012) 10768–10776, https://doi.org/10.1021/ac302748b.
[44] M. Gotsmy, J. Brunmair, C. Büschl, C. Gerner, J. Zanghellini, Probabilistic quotient's work and pharmacokinetics' contribution: countering size effect in metabolic time series measurements, BMC Bioinf. 23 (2022) 379, https://doi.org/10.1186/s12859-022-04918-1.
[45] B.B. Misra, Data normalization strategies in metabolomics: current challenges, approaches, and tools, Eur. J. Mass Spectrom. 26 (2020) 165–174, https://doi.org/10.1177/1469066720918446.
[46] J. Sun, Y. Xia, Pretreating and normalizing metabolomics data for statistical analysis, Genes Dis 11 (2024) 100979, https://doi.org/10.1016/j.gendis.2023.04.018.
[47] G. Nyamundanda, L. Brennan, I.C. Gormley, Probabilistic principal component analysis for metabolomic data, BMC Bioinf. 11 (2010) 571, https://doi.org/10.1186/1471-2105-11-571.
[48] D. Yuan, Y. Liang, L. Yi, Q. Xu, O.M. Kvalheim, Uncorrelated linear discriminant analysis (ULDA): a powerful tool for exploration of metabolomics data, Chemom. Intell. Lab. Syst. 93 (2008) 70–79, https://doi.org/10.1016/j.chemolab.2008.04.005.

[49] G. Pei, L. Chen, W. Zhang, WGCNA application to proteomic and metabolomic data analysis, Methods Enzymol. 585 (2017) 135–158, https://doi.org/10.1016/bs.mie.2016.09.016.
[50] H.K. Pedersen, S.K. Forslund, V. Gudmundsdottir, A.Ø. Petersen, F. Hildebrand, T. Hyötyläinen, T. Nielsen, T. Hansen, P. Bork, S.D. Ehrlich, S. Brunak, M. Oresic, O. Pedersen, H.B. Nielsen, A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links, Nat. Protoc. 13 (2018) 2781–2800, https://doi.org/10.1038/s41596-018-0064-z.
[51] K.T. Do, D.J.N.-P. Rasp, G. Kastenmüller, K. Suhre, J. Krumsiek, MoDentify: phenotype-driven module identification in metabolomics networks at different resolutions, Bioinformatics 35 (2019) 532–534, https://doi.org/10.1093/bioinformatics/bty650.
[52] P. Langfelder, S. Horvath, Eigengene networks for studying the relationships between co-expression modules, BMC Syst. Biol. 1 (2007) 54, https://doi.org/10.1186/1752-0509-1-54.
[53] L.J. Lesko, I. Zineh, S.-M. Huang, What is clinical utility and Why should We care? Clin. Pharmacol. Ther. 88 (2010) 729–733, https://doi.org/10.1038/clpt.2010.229.
[54] E.W. Steyerberg, Y. Vergouwe, Towards better clinical prediction models: seven steps for development and an ABCD for validation, Eur. Heart J. 35 (2014) 1925–1931, https://doi.org/10.1093/eurheartj/ehu207.
[55] M.C. Sachs, A. Sjölander, E.E. Gabriel, Aim for clinical utility, not just predictive accuracy, Epidemiol. Camb. Mass 31 (2020) 359–364, https://doi.org/10.1097/EDE.0000000000001173.
[56] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, The BMJ 352 (2016) i6, https://doi.org/10.1136/bmj.i6.
[57] V.X. Liu, D.W. Bates, J. Wiens, N.H. Shah, The number needed to benefit: estimating the value of predictive analytics in healthcare, J. Am. Med. Inform. Assoc. JAMIA 26 (2019) 1655–1659, https://doi.org/10.1093/jamia/ocz088.
[58] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B.V. Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. van Smeden, A.-L. Boulesteix, J. C. Camaradou, L.A. Celi, S. Denaxas, A.K. Denniston, B. Glocker, R.M. Golub, H. Harvey, G. Heinze, M.M. Hoffman, A.P. Kengne, E. Lam, N. Lee, E.W. Loder, L. Maier-Hein, B.A. Mateen, M.D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, P. Logullo, TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378, https://doi.org/10.1136/bmj-2023-078378.
[59] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, BMJ 350 (2015) g7594, https://doi.org/10.1136/bmj.g7594.
[60] J. Feng, R.V. Phillips, I. Malenica, A. Bishara, A.E. Hubbard, L.A. Celi, R. Pirracchio, Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare, Npj Digit. Med. 5 (2022) 1–9, https://doi.org/10.1038/s41746-022-00611-y.
[61] M. Steidl, M. Felderer, R. Ramler, The pipeline for the continuous development of artificial intelligence models—current state of research and practice, J. Syst. Softw. 199 (2023) 111615, https://doi.org/10.1016/j.jss.2023.111615.
[62] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, The Precise4Q consortium, 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (2020) 310, https://doi.org/10.1186/s12911-020-01332-6.
[63] R. Steuer, Review: on the analysis and interpretation of correlations in metabolomic data, Brief. Bioinform. 7 (2006) 151–158, https://doi.org/10.1093/bib/bbl009.
[64] D. Camacho, A. de la Fuente, P. Mendes, The origin of correlations in metabolomics data, Metabolomics 1 (2005) 53–63, https://doi.org/10.1007/s11306-005-1107-3.
[65] A. Rosato, L. Tenori, M. Cascante, P.R. De Atauri Carulla, V.A.P. Martins dos Santos, E. Saccenti, From correlation to causation: analysis of metabolomics data using systems biology approaches, Metabolomics 14 (2018) 37, https://doi.org/10.1007/s11306-018-1335-y.
[66] Y. Lu, Z. Pang, J. Xia, Comprehensive investigation of pathway enrichment methods for functional interpretation of LC–MS global metabolomics data, Brief. Bioinform. 24 (2022) bbac553, https://doi.org/10.1093/bib/bbac553.
[67] C. Wieder, C. Frainay, N. Poupin, P. Rodríguez-Mier, F. Vinson, J. Cooke, R.P. Lai, J.G. Bundy, F. Jourdan, T. Ebbels, Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis, PLoS Comput. Biol. 17 (2021) e1009105, https://doi.org/10.1371/journal.pcbi.1009105.
[68] G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M.N. Artyomov, A. Sergushichev, Fast Gene Set Enrichment Analysis, 2021, https://doi.org/10.1101/060012.
[69] M.A. García-Campos, J. Espinal-Enríquez, E. Hernández-Lemus, Pathway analysis: state of the art, Front. Physiol. 6 (2015) 383, https://doi.org/10.3389/fphys.2015.00383.
[70] P. Khatri, M. Sirota, A.J. Butte, Ten Years of pathway analysis: current approaches and Outstanding challenges, PLoS Comput. Biol. 8 (2012) e1002375, https://doi.org/10.1371/journal.pcbi.1002375.
[71] J. Xia, D.S. Wishart, MetPA: a web-based metabolomics tool for pathway analysis and visualization, Bioinformatics 26 (2010) 2342–2344, https://doi.org/10.1093/bioinformatics/btq418.
[72] Y. Shao, T. Li, Z. Liu, X. Wang, X. Xu, S. Li, G. Xu, W. Le, Comprehensive metabolic profiling of Parkinson's disease by liquid chromatography-mass spectrometry, Mol. Neurodegener. 16 (2021) 4, https://doi.org/10.1186/s13024-021-00425-8.
[73] X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J.R. Ledsam, M.K. Schmid, K. Balaskas, E. J. Topol, L.M. Bachmann, P.A. Keane, A.K. Denniston, A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digit. Health 1 (2019) e271–e297, https://doi.org/10.1016/S2589-7500(19)30123-2.
[74] G.S. Collins, J.A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.-M. Yu, K.G. Moons, D.G. Altman, External validation of multivariable prediction models: a systematic review of methodological conduct and reporting, BMC Med. Res. Methodol. 14 (2014) 40, https://doi.org/10.1186/1471-2288-14-40.
[75] S. Vollmer, B.A. Mateen, G. Bohner, F.J. Király, R. Ghani, P. Jonsson, S. Cumbers, A. Jonas, K.S.L. McAllister, P. Myles, D. Grainger, M. Birse, R. Branson, K.G. M. Moons, G.S. Collins, J.P.A. Ioannidis, C. Holmes, H. Hemingway, Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, BMJ 368 (2020) l6927, https://doi.org/10.1136/bmj.l6927.
[76] M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J.P.A. Ioannidis, G.S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, BMJ 368 (2020) m689, https://doi.org/10.1136/bmj.m689.
[77] A. Volkmann, R. De Bin, W. Sauerbrei, A.-L. Boulesteix, A plea for taking all available clinical information into account when assessing the predictive value of omics data, BMC Med. Res. Methodol. 19 (2019) 162, https://doi.org/10.1186/s12874-019-0802-0.
[78] E. Porcu, J. Sjaarda, K. Lepik, C. Carmeli, L. Darrous, J. Sulc, N. Mounier, Z. Kutalik, Causal inference methods to integrate omics and complex Traits, Cold Spring Harb. Perspect. Med. 11 (2021) a040493, https://doi.org/10.1101/cshperspect.a040493.
[79] J. Wilkinson, K.F. Arnold, E.J. Murray, M. van Smeden, K. Carr, R. Sippy, M. de Kamps, A. Beam, S. Konigorski, C. Lippert, M.S. Gilthorpe, P.W.G. Tennant, Time to reality check the promises of machine learning-powered precision medicine, Lancet Digit. Health 2 (2020) e677–e680, https://doi.org/10.1016/S2589-7500(20)30200-4.
[80] E.J. Want, P. Masson, F. Michopoulos, I.D. Wilson, G. Theodoridis, R.S. Plumb, J. Shockcor, N. Loftus, E. Holmes, J.K. Nicholson, Global metabolic profiling of animal and human tissues via UPLC-MS, Nat. Protoc. 8 (2013) 17–32, https://doi.org/10.1038/nprot.2012.135.
[81] S. Li, A. Siddiqa, M. Thapa, Y. Chi, S. Zheng, Trackable and scalable LC-MS metabolomics data processing using asari, Nat. Commun. 14 (2023) 4113, https://doi.org/10.1038/s41467-023-39889-1.
[82] N.G. Mahieu, G.J. Patti, Systems-level annotation of a metabolomics data set reduces 25 000 features to Fewer than 1000 unique metabolites, Anal. Chem. 89 (2017) 10397–10406, https://doi.org/10.1021/acs.analchem.7b02380.
[83] L. Perez de Souza, A.R. Fernie, Computational methods for processing and interpreting mass spectrometry-based metabolomics, Essays Biochem. (2023), https://doi.org/10.1042/EBC20230019. EBC20230019.
[84] O.D. Myers, S.J. Sumner, S. Li, S. Barnes, X. Du, Detailed investigation and comparison of the XCMS and MZmine 2 Chromatogram construction and chromatographic Peak detection methods for Preprocessing mass spectrometry metabolomics data, Anal. Chem. 89 (2017) 8689–8695, https://doi.org/10.1021/acs.analchem.7b01069.
[85] A. Delabriere, P. Warmer, V. Brennsteiner, N. Zamboni, SLAW: a scalable and self-Optimizing processing Workflow for untargeted LC-MS, Anal. Chem. 93 (2021) 15024–15032, https://doi.org/10.1021/acs.analchem.1c02687.
[86] Z. Pang, Y. Lu, G. Zhou, F. Hui, L. Xu, C. Viau, A.F. Spigelman, P.E. MacDonald, D. S. Wishart, S. Li, J. Xia, MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation, Nucleic Acids Res. gkae253 (2024), https://doi.org/10.1093/nar/gkae253.
[87] B. Erdős, J.A. Westerhuis, M.E. Adriaens, S.D. O'Donovan, R. Xie, C.M. Singh-Povel, A.K. Smilde, I.C.W. Arts, Analysis of high-dimensional metabolomics data with complex temporal dynamics using RM-ASCA+, PLoS Comput. Biol. 19 (2023) e1011221 https://doi.org/10.1371/journal.pcbi.1011221.
[88] S. Li, Y. Park, S. Duraisingham, F.H. Strobel, N. Khan, Q.A. Soltow, D.P. Jones, B. Pulendran, Predicting network Activity from high throughput metabolomics, PLoS Comput. Biol. 9 (2013) e1003123, https://doi.org/10.1371/journal.pcbi.1003123.
[89] D.B. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592, https://doi.org/10.1093/biomet/63.3.581.
[90] T.J. Davis, T.R. Firzli, E.A. Higgins Keppler, M. Richardson, H.D. Bean, Addressing missing data in GC × GC metabolomics: Identifying Missingness type and evaluating the impact of imputation methods on Experimental Replication, Anal. Chem. 94 (2022) 10912–10920, https://doi.org/10.1021/acs.analchem.1c04093.
[91] M. Orešič, J.P. Posti, M.H. Kamstrup-Nielsen, R.S.K. Takala, H.F. Lingsma, I. Mattila, S. Jäntti, A.J. Katila, K.L.H. Carpenter, H. Ala-Seppälä, A. Kyllönen, H.-R. Maanpää, J. Tallus, J.P. Coles, I. Heino, J. Frantzén, P.J. Hutchinson, D. K. Menon, O. Tenovuo, T. Hyötyläinen, Human serum metabolites associate with Severity and patient outcomes in traumatic brain injury, EBioMedicine 12 (2016) 118–126, https://doi.org/10.1016/j.ebiom.2016.07.015.
[92] SCORE2 working group and ESC Cardiovascular risk collaboration, SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe, Eur. Heart J. 42 (2021) 2439–2454, https://doi.org/10.1093/eurheartj/ehab309.
[93] R.J. Langley, E.L. Tsalik, J.C. van Velkinburgh, S.W. Glickman, B.J. Rice, C. Wang, B. Chen, L. Carin, A. Suarez, R.P. Mohney, D.H. Freeman, M. Wang, J. You, J. Wulff, J.W. Thompson, M.A. Moseley, S. Reisinger, B.T. Edmonds, B. Grinnell, D. R. Nelson, D.L. Dinwiddie, N.A. Miller, C.J. Saunders, S.S. Soden, A.J. Rogers, L. Gazourian, L.E. Fredenburgh, A.F. Massaro, R.M. Baron, A.M.K. Choi, G. R. Corey, G.S. Ginsburg, C.B. Cairns, R.M. Otero, V.G. Fowler, E.P. Rivers, C.

W. Woods, S.F. Kingsmore, An integrated clinico-metabolomic model improves prediction of death in sepsis, Sci. Transl. Med. 5 (2013) 195ra95, https://doi.org/10.1126/scitranslmed.3005893.