PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Ten quick tips for biomarker discovery and validation analyses using machine learning

Ramon Diaz-Uriarte[1], Elisa Gómez de Lope[2], Rosalba Giugno[3], Holger Fröhlich[4,5], Petr V. Nazarov[6], Isabel A. Nepomuceno-Chamorro[7], Armin Rauschenberger[2], Enrico Glaab[2]*

1 Department of Biochemistry, School of Medicine, Universidad Autónoma de Madrid, Instituto de Investigaciones Biomédicas 'Alberto Sols' (UAM-CSIC), Madrid, Spain, **2** Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg, **3** Department of Computer Science, University of Verona, Verona, Italy, **4** Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany, **5** Bonn-Aachen International Centre for IT (b-it), Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany, **6** Department of Cancer Research, Luxembourg Institute of Health, Strassen, Luxembourg, **7** Dpto. de Lenguajes y Sistemas Informáticos, University of Seville, Seville, Spain

* enrico.glaab@uni.lu

Citation: Diaz-Uriarte R, Gómez de Lope E, Giugno R, Fröhlich H, Nazarov PV, Nepomuceno-Chamorro IA, et al. (2022) Ten quick tips for biomarker discovery and validation analyses using machine learning. PLoS Comput Biol 18(8): e1010357. https://doi.org/10.1371/journal.pcbi.1010357

Editor: Francis Ouellette, McGill University, CANADA

Published: August 11, 2022

Funding: EG acknowledges funding support by the Luxembourg National Research Fund (FNR) as part of the National Centre for Excellence in Research on Parkinson's disease (I1R-BIC-PFN-15NCER), and from the European Union's Horizon 2020 research and innovation programme as part of the projects DIGIPD (grant no. ERAPERMED 2020-314) and PERMIT (grant no. 874 825). RG was supported from the European Union's Horizon 2020 research and innovation programme under grant agreement 814978 and JPcofuND2 Personalised Medicine for Neurodegenerative Diseases project JPND2019-466-037. The funders had no role in study design, data collection and

This is a *PLOS Computational Biology* Methods paper.

Introduction

High-throughput experimental methods for biosample profiling and growing collections of clinical and health record data provide ample opportunities for biomarker discovery and medical decision support. However, many of the new data types, including single-cell omics and high-resolution cellular imaging data, also pose particular challenges for data analysis. A high dimensionality of the data in relation to small numbers of available samples (often referred to as the $p >> n$ problem), influences of additive and multiplicative noise, large numbers of uninformative or redundant data features, outliers, confounding factors and imbalanced sample group numbers are all common characteristics of current biomedical data collections. While first successes have been achieved in developing clinical decision support tools using multifactorial omics data, e.g., resulting in FDA-approved omics-based biomarker signatures for common cancer indications [1], there is still an unmet need and great potential for earlier, more accurate and robust diagnostic and prognostic tools for many complex diseases.

Here, we provide a set of broadly applicable tips to address some of the most common pitfalls and limitations for biomarker signature development, including supervised and unsupervised machine learning, feature selection and hypothesis testing approaches. In contrast to previous guidelines discussing detailed aspects of quality control, statistics or study reporting, we give a broader overview of the typical challenges and sort the quick tips to address them chronologically by the study phase (starting with study design, then covering consecutive phases of biomarker signature discovery and validation, see also the overview in Fig 1). While these tips are not comprehensive, they are chosen to cover what we consider as the most frequent, significant, and practically relevant issues and risks in biomarker development. By pointing the reader to further relevant literature on the covered aspects of biomarker discovery and validation, we hope to provide an initial guideline and entry point into the more detailed technical and application-specific aspects of this field.

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Tip 1: Choose a suitable study design

A first step in the preparation of biomarker signature discovery studies is to define the scientific objective and scope clearly and in detail. Common pitfalls to avoid include imprecise goals such as vague primary and secondary biomedical outcomes to investigate or a loosely defined study scope in terms of subject inclusion and exclusion criteria. This can lead to an inappropriate feasibility and risk assessment, to misunderstandings between the collaborators, and ultimately to a delayed or unsuccessful implementation. The collaborators should therefore agree on, and precisely define, the key study design aspects well in advance, and jointly assess the feasibility and suitability of the planned design in relation to the study goals. Apart from the definition of the specific scope, objectives, and milestones, this also includes the choice of relevant experimental conditions to study (diseases/subtypes/treatments) or prior data to include (e.g., existing clinical and health record data), the selection of a suitable tissue pool/cell type(s) and measurement platform, the biological sampling design (i.e., how the samples will be collected, if not already available), the blocking design [2], and the measurement design (i.e., the arrangement of samples in the measurement instrument and across different measurement batches [3]). Moreover, to ensure that the study is adequately powered and that biospecimen resources are used efficiently, dedicated sample size determination methods [4] and sample selection and matching methods (e.g., for confounder matching between cases and controls) [5] should be applied.

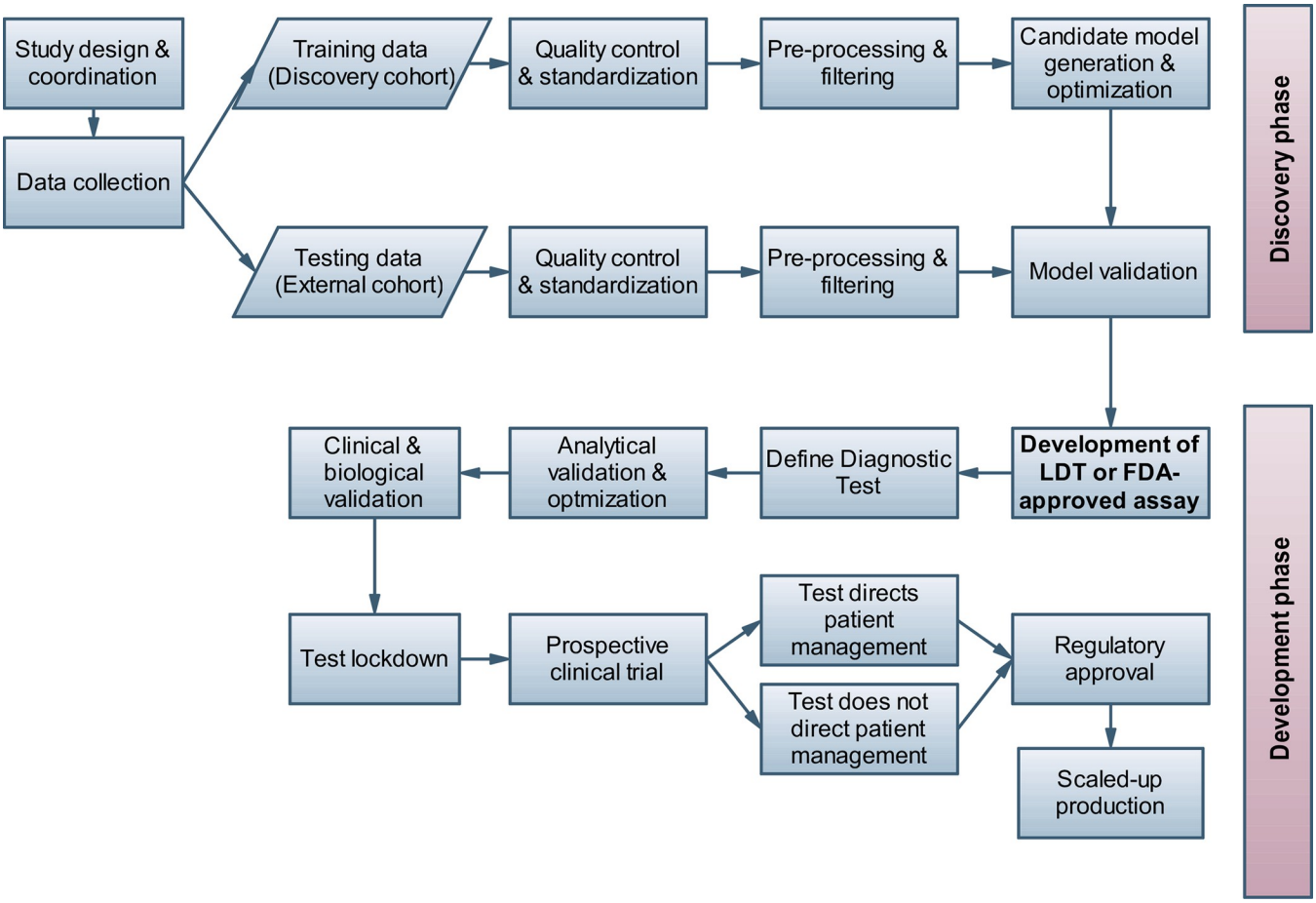

Fig 1. Schematic overview of key steps in a common biomarker test development workflow for patient stratification or disease outcome prediction.

https://doi.org/10.1371/journal.pcbi.1010357.g001

Studies that aim to assess the effects of interventions should include potential confounders as covariates. However, covariates that are common effects of treatment and outcome should not be included in the analysis because they would lead to selection and collider bias [6,7]; likewise, it is not recommended to indiscriminately include pretreatment covariates as they can induce bias amplification [6–8]. In contrast, studies that are purely predictive, without an interest in causation, do not have to be concerned about confounders, and the criteria of covariate inclusion purely depend on increasing predictive performance (see also Tips 4 to 8). Additionally, a specific and common concern with covariates in these types of studies is understanding the relative contribution of different types of variables, in particular clinical versus omics variables, which we address in Tip 3.

As part of the study design, early planning is required to ensure that legal and ethical requirements of data collection will be met throughout the study. For maintaining data security and privacy, data management and access strategies should be defined during this initial planning phase, e.g., by following specific frameworks and guidelines for this purpose [9,10]. Finally, a comprehensive and clear documentation of the study design is essential for effective project monitoring. For this purpose, we recommend following standard reporting guidelines, including visual illustrations of the study design or patient flow through the study, such as CONSORT [11] or STARD [12,13].

Tip 2: Ensure data quality, curation, and standardization

Many biomedical datasets derived from non-targeted molecular profiling or high-throughput imaging approaches are affected by multiple sources of noise and bias, and clinical datasets are often not harmonized across different patient cohorts. In general, one can distinguish between technical noise and biological variance. Current data analytical methods have only a limited ability to discriminate between them. Therefore, quality control and filtering analyses, data curation, annotation, and standardization are important initial steps in biomedical data processing pipelines. Relevant quality controls typically include statistical outlier checks and computing data type-specific quality metrics, as implemented in established software packages, e.g., the *fastQC/FQC* package for next-generation sequencing (NGS) data [14], *arrayQualityMetrics* for microarray data [15], *pseudoQC*, *MeTaQuaC*, and *Normalyzer* for proteomics and metabolomics data [16–18]. Further dedicated quality assurance methods have been developed for cellular and neuroimaging data [19,20], clinical data [21,22], and digital biomarkers [23]. All quality checks should be applied both before and after preprocessing of the raw data to ensure that all quality issues have been resolved and no artificial patterns were introduced by inadequate preprocessing methods.

Apart from the initial processing and filtering, the curation of clinical data also involves dedicated checks and data transformations, e.g., ensuring that the values fall within acceptable ranges (e.g., checking maximum and minimum age and body mass index values), resolving inconsistencies (e.g., different units or value encodings), and transforming the data to standard formats (e.g., OMOP [24], CDISC [25], ICD10/11 [26], SNOMED CT [27]). Beyond these curation steps, a minimum set of required complementary annotations should be made available for subsequent data analyses and dissemination. Standard formats for providing annotations for the most common experimental and clinical data types have already been established, e.g., the MIAME [28] and MINSEQE [28,29] guidelines for microarray and NGS experiments and corresponding standards for metabolomics and proteomics data (e.g., MIAPE [30] and MSI [31]). These standards should be adopted already in the early data processing stages.

Finally, as part of the data curation and standardization, it is recommendable to compare and evaluate multiple options to define primary and secondary study endpoints and other key input and outcome variables (e.g., comparing different definitions of tumor grades or disease

stages or different disease ontologies [32]). Considering multiple definitions of the same disease outcomes can help to address lack of clarity or loss of information associated with the use of only a single outcome definition.

Tip 3: Integrate different data types effectively and assess the value of clinical versus omics data

Studies that have access to multiple datasets or use variables of qualitatively different kinds (e.g., clinical and omics) need to integrate these data effectively. In the machine learning literature, traditionally 3 different strategies for multimodal data integration have been suggested, namely early, intermediate, and late integration [33,34]. Early integration methods focus on extraction of common features from several data modalities. A typical example is canonical correlation analysis (CCA) and sparse variants of CCA [35,36]. In a second step, conventional machine learning methods can then be applied based on the extracted common feature space.

Late integration algorithms first learn separate models for each data modality and then combine predictions made by these models, for example, with the help of a meta-model trained on the outputs of data source specific sub-models. The latter strategy is called stacked generalization, stacking, or super learning [37–39].

Intermediate integration algorithms are the youngest branch of data fusion approaches. The idea is to join data sources while building the predictive model. A classic example of this strategy is support vector machine (SVM) learning with linear combinations of multiple kernel functions [34]. More recently, multimodal neural network architectures have been devised for this purpose [40].

A related problem to data integration is the selection of the most useful data type(s), when multiple available datasets contain redundant information, but have different informative value. A common example for this in biomedicine is assessing the clinical utility of omics data, or any other type of high-dimensional experimental measurement data, when we already have data from traditional clinical markers. The key question here is whether predictors built from omics data provide an added value for decision-making. Addressing this question requires comparative evaluations in addition to an integrative analysis and using the traditional clinical data as the baseline [41–44].

For more detailed guidelines and relevant method comparisons, we refer the reader to a broader overview of machine learning methods for omics data integration [45], representative case studies on combining omics and clinical data [46], and generic multi-omics integration approaches [47,48].

Tip 4: Choose adequate preprocessing and filtering approaches

Raw biomedical data is often influenced by a variety of preanalytical factors, resulting in systematic biases and a shifting and scaling of the measured signals. Many artifacts and normalization issues are data type specific and need to be addressed using dedicated preprocessing and filtering methods. Tailored software solutions have been made available to preprocess clinical data [21], NGS data [49], microarray data [50], different types of metabolomics and proteomics data [18], and cellular and brain imaging data [51–54]. Although no generic rules and methods exist for all data types, the following considerations apply to most datasets. For attributes with a large proportion of missing values (e.g., more than 30% of values missing), researchers may want to consider a complete removal. For features with smaller numbers of missing values, imputation methods or machine learning algorithms that tolerate a limited occurrence of missing values may be applied, depending on the type of missingness [55]. To filter out uninformative attributes, the removal of features with zero or small variance is also recommended, and further alternative filtering methods using the sum of absolute covariances

[55,56] or tests of the unimodality or multimodality of the data distribution have been proposed [57]. After filtering, additional standardization, transformation, or scaling steps may also be warranted. For example, standardization can help to make clinical features on different scales more comparable, and, for linear models, assumptions about the linearity, distribution, and constant variance of the response are often better met after using transformations such as Box-Cox [58,59]. Moreover, functional omics data often displays a dependence of the feature signal variance on the average signal intensity, which can be addressed by a variance stabilizing transformation [60–62]. Finally, the successful application of data filtering and preprocessing should be checked and evaluated, e.g., by repeating initial quality control analyses (see Tip 2) and assessing global shape and distribution characteristics of the processed data using low-dimensional visualizations (e.g., principal coordinate analysis [63], non-metric multidimensional scaling [64], t-SNE [65], and UMAP [66]) and dedicated software tools for omics visualization [67].

Tip 5: Compare and select relevant modeling methods

After data preprocessing, appropriate statistical and machine learning methods need to be chosen for the analysis. Model selection strongly depends on the analysis goals, e.g., whether a probabilistic model of the data or a prediction of a categorical outcome is needed, and whether the study focus is on model interpretability or model performance. To preselect suitable algorithms for comparative evaluation, the number of input and output features, the number of available samples, and the type of features (categorical, numerical, ordinal) need to be considered [57,68]. The selection of the modeling procedure can also be informed by low-dimensional data visualizations and distribution plots [69–71]. However, low-dimensional intuitions of patterns in high-dimensional data can also be misleading, if the sample distances in the original feature space are not well preserved and partly reflect idiosyncrasies of the visualization method [72]. To facilitate model selection for the non-expert, automated machine learning (AutoML) approaches have been proposed, which use combinatorial search algorithms and heuristics to replace manual tasks in model selection [73]. But not all models are suitable for all types of data. For example, training a deep neural network with high-dimensional data of a few hundred patient samples is likely to result in a highly overfitted model. Hence, it is necessary to carefully choose the right types of models a priori and not purely rely on brute force compute power. To facilitate the choice for the reader, an overview of commonly used unsupervised and supervised machine learning algorithms, including popular implementations in the programming languages R and Python, references to methodology descriptions, and best practice example applications is provided in Tables A and B in S1 Text, respectively.

Once suitable modeling procedures have been chosen, comparing multiple representative approaches is recommended. This can be achieved by applying cross-validation or bootstrapping methods, followed by comparing different performance metrics using statistical tests [74,75] (see also Tip 6). However, overfitting should be avoided, e.g., by using nested cross-validation, and the significance scores for performance statistics should be adjusted for multiple hypothesis testing [75]. Apart from *p*-value significance scores, confidence intervals and similar measures of uncertainty should be assessed [76–79], taking into account the limitations of individual uncertainty measures [80]. Finally, in addition to assessing individual machine learning algorithms, the integration of modeling approaches using ensemble learning (for both supervised and unsupervised problems) or consensus clustering (for unsupervised problems) may be explored to combine the benefits of different modeling methods [81,82].

While extensive model evaluations and comparisons are generally beneficial, the success and feasibility of the model selection scheme will also depend on realistic time planning and consideration of the run-time requirements for the preselected algorithms [83]. At the end of a

comparative model evaluation, several algorithms may display a very similar prediction performance. Hence, secondary selection criteria, such as interpretability or stability of feature selection should be considered. In summary, researchers should carefully plan all model selection steps and choose suitable and objective evaluation criteria before running computationally expensive analyses.

Tip 6: Optimize model parameters and feature selection without overfitting

Biomedical datasets often have many more features than samples (the "$p >> n$" problem). This increases the risk for creating overfitted models, because data points are sparsely distributed in a very high dimensional space, resulting in statistically unstable models. Two popular approaches to prevent overfitting are ridge and lasso regularization [84,85], which shrink the squared, or respectively, absolute model coefficients towards zero. Alternatively, combining ridge and lasso regularization, the elastic net [85,86] can handle correlated variables more effectively than the lasso [85,87]. By optimizing the regularization parameter, which determines the extent to which estimated model coefficients are shrunk towards zero, we can prevent overfitting (too little shrinkage) and underfitting (too much shrinkage). The most common way of optimizing this and other hyperparameters is to perform a grid search with cross-validation, but there are more efficient alternatives [88,89], as well as Bayesian procedures, in which the prior performs the role of the penalty [90–92].

A common mistake in model optimization is to not only perform unsupervised but also supervised feature selection outside cross-validation. For example, removing features because of their low variance or their high correlation with other input features is a suitable global filtering method, but removing features from both training and test set data because of their low correlation with the target variable is an error [84]. Supervised attribute selection must take place inside cross-validation to avoid information leakage and overoptimistic estimates of predictive performance resulting from selection bias [93,94]. This also applies if the aim is to compare different approaches (e.g., data pre-preprocessing, feature transformation) before selecting the most predictive one. Moreover, if cross-validation is applied for both hyperparameter optimization and performance estimation (see Tip 7), a nested cross-validation scheme is required, i.e., while an outer cross-validation loop is used for performance estimation, an inner cross-validation loop is used for hyperparameter optimization. An alternative to selecting single hyperparameters by cross-validation is to combine multiple hyperparameters by stacked generalization [37,95,96]. Furthermore, predictive models avoiding explicit hyperparameter optimization may be chosen, e.g., random forests [97–99].

Finally, for many biomedical applications, natural structures among features or complementary information on the features can be exploited as an additional information source for model building. For example, among causally related features, we might want to prioritize the selection of upstream over downstream features in a known causal graph [100] to account for pairs or groups of functionally related features [101,102] or to transfer information from previous studies (i.e., prior weights or prior effects) into the learning procedure. These approaches to integrate prior knowledge into the learning phase have the potential to render models more predictive and more interpretable.

Tip 7: Assess model performance in an unbiased and robust fashion

Once the data have been prepared and modeling approaches selected, a metric has to be chosen to assess model performance. The performance metric selection is problem specific, and it

is often recommended to consider multiple metrics to distinguish between different error types (e.g., type 1 versus type 2 error) and consider different penalties for outliers (e.g., quadratic versus non-quadratic loss functions). This is particularly important for imbalanced study groups [103], often observed in biomedical projects (e.g., identifying approximately 0.3% breast cancer patients in a population-wide mammography screening). Researchers may consider using balanced accuracy measures or ensure balancing during model training by applying over/under-sampling or data augmentation methods (test set samples should however always remain independent from the training set and synthetic redundancy introduced by oversampling should be avoided) [104–106]. Moreover, a prior sample size calculation and clearly defined study goals can help to ensure that enough samples for each study group are available for both modeling and performance assessment. In general, researchers should ensure that machine learning models are well calibrated, i.e., the distribution of predicted probabilities is close to the true probabilities of class membership. The most common calibration techniques and calibration measures for this purpose have been reviewed previously [107].

Common performance measure choices include the balanced accuracy, the F1 score, Matthew's correlation coefficient, sensitivity/specificity for supervised binary classification, the mean squared error or absolute error and (adjusted) R^2 for regression tasks [59,84,92], and internal validity indices, such as the average Silhouette width or Calinski–Harabasz index for unsupervised clustering [108,109]. However, the choice of the performance metric does not only depend on the outcome variable type but also the specific analysis goals and applications (see [110] for an empirical study of different performance metrics). Moreover, for classifiers that provide predicted probabilities for group membership rather than pure categorical outcome predictions, dedicated performance measures are available to avoid the subjective choice of threshold values for outcome categorization (a problem that affects accuracy, sensitivity, and specificity measures [111,112]). These include Brier's score, the concordance index, the area under the receiver operating characteristic curve (AUC), the precision-recall curve (PR AUC), and the kappa curve (AUK), which can also be applied to survival data [111–116]. Depending on the clinical scenario, the uniform weighting of type 1 and type 2 errors in classical performance measures may sometimes provide counterintuitive classifier rankings, and the use of decision-analytic tools, which take into account the costs of different error types, should be considered [112,117].

When estimating a model's generalization performance from observational data, the variability in biomedical datasets is often high, due to both technical and biological sources of variation. To address this challenge, bootstrapping methods, such as .632+ bootstrap, can be used to obtain more robust performance estimates [118]. Another well-accepted approach is repeated or iterated *k*-fold cross-validation, which often gives less biased estimates of the true generalization performance [119]. When selecting the parameter *k*, the user should be aware of the balance between bias (low *k*) and variability (high *k*, e.g., for leave-one-out cross-validation) [118,120]. Bolstered error estimation is a further robust alternative approach dedicated specifically to datasets with small sample size [121,122]. Finally, it is important to remember that high estimated performance on a single test dataset does not equate to generalizability on other datasets and to clinical or biomedical relevance [123] (see also Tip 8). More detailed practical guidance on the use of relevant algorithms and software tools for model performance assessment, including best practice examples, is provided in [85,92,124–126].

Tip 8: Improve and validate the generalization capability of the model

Depending on the goals of a biomarker study (e.g., whether the study involves a clinical validation or only preclinical biomarker research) and the study type (e.g., whether the study is

prospective or retrospective), different options are available to improve and evaluate an initial biomarker signature obtained from a discovery cohort. Clinical biomarker studies require that the final model is locked and recorded before testing on an independent validation cohort. The subjects in the validation cohort have to be representative of the intended patient population and fulfill the same inclusion and exclusion criteria as the discovery cohort [127,128]. Depending on whether the discovery and validation cohorts cover distinct geographic regions, environments, and ethnic backgrounds, the generalization capability of the final model may be restricted significantly by the population coverage and diversity of the included cohorts.

Studies focusing on early preclinical stages of biomarker discovery have more flexibility in collecting additional data to optimize and confirm the generalization capability of an initial machine learning model. Apart from straightforward optimization strategies, such as increasing the size of the discovery cohort and thereby the size of the training dataset for modeling, a wide range of external data sources can be exploited to further improve a model. For example, integrative meta-analyses of in-house data and relevant public or collaborator-derived clinical and omics data can be applied to improve the feature selection for a model [129], or prior knowledge from cellular pathway databases and the biomedical literature can be used to filter predictive molecular biomarkers depending on their involvement in disease-associated pathways [130] or to derive more robust pathway- or network-based predictive features [131]. Furthermore, cellular or animal models for the disease condition of interest can provide additional data for biomarker validation, which is often freely available in public data repositories. Functional validation studies involving the modulation of candidate biomarker molecules or pathways via knockdown and overexpression experiments in a disease model may provide information on causal associations with measurable disease phenotypes [132]. While all these information sources provide effective means for the initial confirmation and filtering of candidate markers, after having optimized a biomarker signature and locked down the final machine learning model, the final clinical evaluation will always require an adequately powered external validation on a distinct, representative patient cohort.

Tip 9: Ascertain that the model meets the required level of interpretability and explainability

Depending on the goals of a biomedical prediction or stratification project, the success of applied machine learning methods might not only depend on the predictive performance of generated models but also their interpretability, biological plausibility, and insightfulness. When interpretability and explainability are relevant objectives and criteria for the study success, researchers should consider so-called "white-box" learning algorithms, i.e., modeling approaches that link input features to the outcome variable of interest in a more transparent and easier to understand fashion than the more complex, but often also more accurate, "black-box" modeling methods.

For settings requiring a high level of model interpretability, a wide variety of machine learning approaches is available to find a suitable compromise between model generalization capability and explainability. Common examples for learning approaches favoring interpretability are linear modeling methods [92] and rule-based machine learning methods, such as classification and regression trees [133,134], combinatorial rule learning approaches [135,136], and probabilistic and fuzzy rule learning methods [137,138]. While linear modeling approaches enable a relevance scoring and ranking of features by their absolute weights in a model, rule-based learning approaches can provide additional information on feature associations by computing statistics on their co-occurrence in decision rule sets [139]. Apart from these generic learning methods, more recently, domain-specific interpretable prediction and clustering

approaches, which exploit prior biological knowledge from cellular pathways and molecular networks [140–142], have gained interest. In addition, there is a quickly growing literature on Explainable AI (XAI) techniques to interpret also very complex black-box models, such as neural networks. Examples include Shapley Additive Explanations [143], LIME [144], Explainable Boosting Machines [145], and symbolic meta-modeling [146]. A systematic review of those and further methods can be found in [147,148].

In summary, white-box modeling methods are not required for all applications, but being able to understand a stratification or prediction model derived from biomedical data and assess its biological plausibility is often beneficial, and particularly important in clinical decision support applications. In these settings, the transparency, credibility, and trustworthiness of machine learning models is equally important as the evidence for predictive power [149].

Tip 10: Translate biomarker discoveries to in vitro diagnostics or diagnostic medical devices

Most biomarker signature discoveries are obtained using non-targeted, high-throughput measurement approaches, which cover large numbers of candidate biomarkers, but lack sensitivity and are not certified for diagnostic applications. If the long-term study goal is to develop biomarker findings into a clinically validated diagnostic test, then it is typically not only necessary to validate the biomarker signature on an external cohort but also to translate the original high-throughput measurement approach to a more targeted and sensitive measurement technology, which fulfills the requirements for clinical biomarker applications in terms of technical reliability and robustness.

Typical examples for this transition from non-targeted methodologies (e.g., omics profiling of patient biospecimens) to a targeted approach are the replacement of high-throughput transcriptomics profiling by targeted qRT-PCR or digital PCR measurements, or the replacement of mass spectrometry (MS)-based proteomics by targeted immunoassays, after developing and producing specific antibodies targeting the omics-derived peptide or protein fragment biomarkers. While the original discovery analyses are conducted on measurements for several thousands of biomarkers (e.g., 50k genetic transcripts), the targeted analyses focus only on small numbers of candidate biomarkers (e.g., 10 to 20 transcripts), selected using machine learning and cross-validation analyses of the original discovery data. The transition from non-targeted to targeted approaches normally does not only just require a new validation of the targeted version of the biomarker signatures but also adjustments of model parameters. If sufficient training data is available for the targeted method according to a sample size calculation, this model adjustment can be obtained by simply refitting the model on the new data. However, to guide the model building process and exploit the prior data from non-targeted analyses, it may additionally be worthwhile to consider applying transfer learning approaches. Transfer learning techniques use information from pre-trained machine learning models (e.g., information on the feature relevance or feature effects with respect to a clinical outcome of interest) to apply it to a new but similar data analysis task, in order to exploit the prior information to build more robust and accurate models (see [150] for a review of methodologies).

After a biomarker model has been refitted successfully to targeted measurement data, there are 2 main possible pathways for translating the model into a clinical biomarker test: The development of an in vitro diagnostic (IVD) or a diagnostic medical device. IVDs are tests applied to human body fluid or tissue samples to assess an individual's health status. In contrast to other medical devices, they do not involve any direct action on the patient. By contrast, diagnostic medical devices can come in direct contact with the patient and include active devices with different levels of associated risk (in different countries, medical devices are

categorized into different regulatory classes, depending on the risk and the required regulatory control). In the EU, all medical devices must be CE marked before they can reach the market ("CE" stands for "conformité européenne" and indicates that a product has been assessed by the manufacturer and deemed to meet EU safety, health, and environmental protection requirements). Further details on regulatory pathways for machine learning-based IVD and diagnostic medical device development and a comparison of associated policies in Europe and the United States can be found in a dedicated article [151]. Finally, researchers should take into consideration relevant FDA guidelines, in particular the "Good Machine Learning Practice for Medical Device Development: Guiding Principles" [152], which highlights the different types of multidisciplinary expertise required throughout the total product life cycle of a medical device.

Conclusions

Biomarker signature discovery and development involve complex interdisciplinary collaborations and several interdependent tasks and decisions, ranging from the initial choice of study design parameters to the approaches for data collection and preprocessing, and the strategies for model building and validation. Many of the challenges in these projects are study and problem specific and cannot be fully addressed by general guidelines and recommendations. However, a variety of common pitfalls, issues, and limitations are shared across the majority of biomarker discovery and validation studies, and dedicated strategies and methods to circumvent or alleviate these common problems are already available.

In this article, we have chronologically summarized some of the most frequent challenges that occur during the typical phases of biomarker projects and suggested methods and software tools that may help to avoid unsuitable study designs, prevent analysis and validation errors, and increase chances for success. Since the practical implementation for many of the covered topics would require more detailed explanations, we have directed the reader to relevant literature with more in-depth information for each tip. For an overview of related existing guidelines and data and methods standardization efforts, we also recommend to study the "Criteria for the use of omics-based predictors in clinical trials" by the US National Cancer Institute [153] with a focus on omics-derived biomarkers and the standard framework "Assessing Credibility of Computational Modeling through Verification and Validation: Application to Medical Devices" with a broad applicability beyond the specific framework focus on medical devices [154]. Furthermore, as a guidance on how to document and present biomarker results derived from machine learning approaches, we refer the reader to the TRIPOD Statement on "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" [155] and the more generic "Standards for Reporting of Diagnostic Accuracy (STARD)" [12,13]. In practice, project managers should also ensure that the required multidisciplinary expertise for all project phases is well represented in the project consortium, and that measures for effective cross-disciplinary communication throughout the project are set in place.

As further steps in the future, community-driven standardization efforts, involving researchers, practitioners, and regulators in the field, are still needed to develop more comprehensive and detailed documentation and validation standards, minimum requirements, and study type-specific guidelines to further improve the quality of biomarker stratification and prediction projects.

Supporting information

S1 Text. Supporting Text S1 for the manuscript "Ten Quick Tips for Biomarker Discovery and Validation Analyses Using Machine Learning". Table A in S1 Text. Unsupervised

learning algorithms. Overview of widely used unsupervised machine learning algorithms, including implementations in the programming languages R and Python, references to methodology descriptions, and best practice example applications. **Table B in S1 Text. Supervised learning algorithms.** Overview of widely used supervised machine learning algorithms, including implementations in the programming languages R and Python, references to methodology descriptions, and best practice example applications.
(PDF)

Acknowledgments

We thank Prof. Anne-Laure Boulesteix and Dr. Francisco Azuaje for helpful comments and suggestions during our expert consultation workshop on machine learning in personalized medicine.

References

1. Moshkovskii S, Pyatnitsky M, Lokhov P, Baranova A. OMICS for Tumor Biomarker Research. Biomarkers. Cancer. 2014:1–22. https://doi.org/10.1007/978-94-007-7744-6_14–1
2. Casler MD. Blocking Principles for Biological Experiments. Applied Statistics in Agricultural, Biological, and Environmental Sciences. 2018. p. 53–72. https://doi.org/10.2134/appliedstatistics.2015.0074.c3
3. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics. 2016; 17:29–39. https://doi.org/10.1093/biostatistics/kxv027 PMID: 26272994
4. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, et al. Harmonization of quality metrics and power calculation in multi-omic studies. Nat Commun. 2020; 11:3092. https://doi.org/10.1038/s41467-020-16937-8 PMID: 32555183
5. de Graaf MA, Jager KJ, Zoccali C, Dekker FW. Matching, an appealing method to avoid confounding? Nephron Clin Pract. 2011; 118:c315–c318. https://doi.org/10.1159/000323136 PMID: 21293153
6. Hernan MA, Robins JM. Causal Inference. CRC Press; 2020.
7. Pearl J. Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge University Press; 2009.
8. Ding P, VanderWeele TJ, Robins JM. Instrumental variables as bias amplifiers with general outcome and confounding. Biometrika. 2017; 104:291–302. https://doi.org/10.1093/biomet/asx009 PMID: 29033459
9. Aramesh K. An Ethical Framework for Global Governance for Health Research. Springer. Nature. 2019.
10. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data. 2018; 5. https://doi.org/10.1186/s40537-017-0110-7
11. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008; 337:a2390. https://doi.org/10.1136/bmj.a2390 PMID: 19001484
12. Korevaar DA, Cohen JF, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. Updating standards for reporting diagnostic accuracy: the development of STARD 2015. Res Integr Peer Rev. 2016; 1:7. https://doi.org/10.1186/s41073-016-0014-7 PMID: 29451535
13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015; 351:h5527. https://doi.org/10.1136/bmj.h5527 PMID: 26511519
14. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017:3137–3139. https://doi.org/10.1093/bioinformatics/btx373 PMID: 28605449
15. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics. 2009; 25:415–416. https://doi.org/10.1093/bioinformatics/btn647 PMID: 19106121
16. Wang S, Yang H. pseudoQC: A Regression-Based Simulation Software for Correction and Normalization of Complex Metabolomics and Proteomics Datasets. Proteomics. 2019; 19:e1900264. https://doi.org/10.1002/pmic.201900264 PMID: 31474000

17. Kuhring M, Eisenberger A, Schmidt V, Kränkel N, Leistner DM, Kirwan J, et al. Concepts and Software Package for Efficient Quality Control in Targeted Metabolomics Studies: MeTaQuaC. Anal Chem. 2020; 92:10241–10245. https://doi.org/10.1021/acs.analchem.0c00136 PMID: 32603093

18. Chawade A, Alexandersson E, Levander F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. J Proteome Res. 2014; 13:3114–3120. https://doi.org/10.1021/pr401264n PMID: 24766612

19. Huguet J, Falcon C, Fusté D, Girona S, Vicente D, Molinuevo JL, et al. Management and Quality Control of Large Neuroimaging Datasets: Developments From the Barcelonaβeta Brain Research Center. Front Neurosci. 2021; 15:633438. https://doi.org/10.3389/fnins.2021.633438 PMID: 33935631

20. Qiu M, Zhou B, Lo F, Cook S, Chyba J, Quackenbush D, et al. A cell-level quality control workflow for high-throughput image analysis. BMC Bioinformatics. 2020; 21:280. https://doi.org/10.1186/s12859-020-03603-5 PMID: 32615917

21. Gu W, Yildirimman R, Van der Stuyft E, Verbeeck D, Herzinger S, Satagopam V, et al. Data and knowledge management in translational research: implementation of the eTRIKS platform for the IMI OncoTrack consortium. BMC Bioinformatics. 2019; 20:164. https://doi.org/10.1186/s12859-019-2748-y PMID: 30935364

22. Prokscha S. Practical Guide to Clinical Data Management. 3rd ed. CRC Press; 2011.

23. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med. 2019; 2. https://doi.org/10.1038/s41746-019-0090-4 PMID: 30868107

24. Reinecke I, Zoch M, Wilhelm M, Sedlmayr M, Bathelt F. Transfer of Clinical Drug Data to a Research Infrastructure on OMOP—A FAIR Concept. Stud Health Technol Inform. 2021; 287:63–67. https://doi.org/10.3233/SHTI210815 PMID: 34795082

25. Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med. 2009; 48:408–413. https://doi.org/10.3414/ME9236 PMID: 19621114

26. Buescher PA. The International Classification of Diseases (ICD). 2003.

27. Rossander A, Lindsköld L, Ranerup A, Karlsson D. A State-of-the Art Review of SNOMED CT Terminology Binding and Recommendations for Practice and Research. Methods Inf Med. 2021. https://doi.org/10.1055/s-0041-1735167 PMID: 34583415

28. Brazma A. Minimum Information About a Microarray Experiment (MIAME)—successes, failures, challenges. ScientificWorldJournal. 2009; 9:420–423. https://doi.org/10.1100/tsw.2009.57 PMID: 19484163

29. Taylor CF, Field D, Sansone S-A, Aerts J, Apweiler R, Ashburner M, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol. 2008; 26:889–896. https://doi.org/10.1038/nbt.1411 PMID: 18688244

30. Taylor CF. Minimum Reporting Requirements for Proteomics: A MIAPE Primer. Proteomics. 2006:39–44. https://doi.org/10.1002/pmic.200600549 PMID: 17031795

31. Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee DY, Lu Y, et al. Quality control for plant metabolomics: reporting MSI-compliant studies. Plant J. 2008; 53:691–704. https://doi.org/10.1111/j.1365-313X.2007.03387.x PMID: 18269577

32. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012; 40:D940–D946. https://doi.org/10.1093/nar/gkr972 PMID: 22080554

33. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform. 2018; 19:325–340. https://doi.org/10.1093/bib/bbw113 PMID: 28011753

34. Support vector machine applications in computational biology. Kernel Methods in Computational Biology. The MIT Press; 2004.

35. Yoon G, Carroll RJ, Gaynanova I. Sparse semiparametric canonical correlation analysis for data of mixed types. Biometrika. 2020; 107:609–625. https://doi.org/10.1093/biomet/asaa007 PMID: 34621080

36. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. Neural Comput. 2004; 16:2639–2664. https://doi.org/10.1162/0899766042321814 PMID: 15516276

37. Wolpert DH. Stacked generalization. Neural Netw. 1992:241–259. https://doi.org/10.1016/s0893-6080(05)80023-1

38. Džeroski S, Ženko B. Is Combining Classifiers with Stacking Better than Selecting the Best One? Mach Learn. 2004:255–273. https://doi.org/10.1023/b:mach.0000015881.36452.6e

39. Valdes G, Interian Y, Gennatas E, Van der Laan M. The Conditional Super Learner. IEEE Trans Pattern Anal Mach Intell. 2021. https://doi.org/10.1109/TPAMI.2021.3131976 PMID: 34851823

40. Gao J, Li P, Chen Z, Zhang J. A Survey on Deep Learning for Multimodal Data Fusion. Neural Comput. 2020; 32:829–864. https://doi.org/10.1162/neco_a_01273 PMID: 32186998

41. Volkmann A, De Bin R, Sauerbrei W, Boulesteix A-L. A plea for taking all available clinical information into account when assessing the predictive value of omics data. BMC Med Res Methodol. 2019; 19:162. https://doi.org/10.1186/s12874-019-0802-0 PMID: 31340753

42. De Bin R, Boulesteix A-L, Benner A, Becker N, Sauerbrei W. Combining clinical and molecular data in regression prediction models: insights from a simulation study. Brief Bioinform. 2020; 21:1904–1919. https://doi.org/10.1093/bib/bbz136 PMID: 31750518

43. Rodríguez-Girondo M, Salo P, Burzykowski T, Perola M, Houwing-Duistermaat J, Mertens B. Sequential double cross-validation for assessment of added predictive ability in high-dimensional omic applications. Ann Appl Stat. 2018; 12:1655–1678.

44. Truntzer C, Mostacci E, Jeannin A, Petit J-M, Ducoroy P, Cardot H. Comparison of classification methods that combine clinical data and high-dimensional mass spectrometry data. BMC Bioinformatics. 2014; 15:385. https://doi.org/10.1186/s12859-014-0385-z PMID: 25432156

45. Zhou W. Machine Learning Methods for Omics Data. Dermatol Int. 2011.

46. De Bin R, Sauerbrei W, Boulesteix A-L. Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. Stat Med. 2014; 33:5310–5329. https://doi.org/10.1002/sim.6246 PMID: 25042390

47. Hardiman G. Systems Analytics and Integration of Big Omics Data. MDPI. 2020. https://doi.org/10.3390/genes11030245 PMID: 32111000

48. Ahmad A, Fröhlich H. Integrating heterogeneous omics data via statistical inference and learning techniques. Genom Comput Biol. 2016; 2:32.

49. Franke KR, Crowgey EL. Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms. Genomics Inform. 2020; 18:e10. https://doi.org/10.5808/GI.2020.18.1.e10 PMID: 32224843

50. Federico A, Saarimäki LA, Serra A, Del Giudice G, Kinaret PAS, Scala G, et al. Microarray Data Preprocessing: From Experimental Design to Differential Analysis. Methods Mol Biol. 2022; 2401:79–100. https://doi.org/10.1007/978-1-0716-1839-4_7 PMID: 34902124

51. Liberda D, Pięta E, Pogoda K, Piergies N, Roman M, Koziol P, et al. The Impact of Preprocessing Methods for a Successful Prostate Cell Lines Discrimination Using Partial Least Squares Regression and Discriminant Analysis Based on Fourier Transform Infrared Imaging. Cell. 2021; 10. https://doi.org/10.3390/cells10040953 PMID: 33924045

52. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp. 2002; 17:143–155. https://doi.org/10.1002/hbm.10062 PMID: 12391568

53. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 1996; 29:162–173. https://doi.org/10.1006/cbmr.1996.0014 PMID: 8812068

54. Muschelli J, Sweeney E, Crainiceanu CM. freesurfer: Connecting the Freesurfer software with R. F1000Res. 2018; 599. https://doi.org/10.12688/f1000research.14361.1 PMID: 30057753

55. He Y, Zhang G, Hsu C-H. Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies. CRC Press; 2021.

56. Tritchler D, Parkhomenko E, Beyene J. Filtering genes for cluster and network analysis. BMC Bioinformatics. 2009; 10:193. https://doi.org/10.1186/1471-2105-10-193 PMID: 19549335

57. De Bin R, Risso D. A novel approach to the clustering of microarray data via nonparametric density estimation. BMC Bioinformatics. 2011; 12:49. https://doi.org/10.1186/1471-2105-12-49 PMID: 21303507

58. Osborne J. Improving your data transformations: Applying the Box-Cox transformation. University of Massachusetts Amherst. 2010. https://doi.org/10.7275/QBPC-GK17

59. Weisberg S. Applied Linear Regression, 4th ed. John Wiley & Sons; 2014.

60. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019; 20:296. https://doi.org/10.1186/s13059-019-1874-1 PMID: 31870423

61. Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics. 2003; 19:966–972. https://doi.org/10.1093/bioinformatics/btg107 PMID: 12761059

62. Purohit PV, Rocke DM, Viant MR, Woodruff DL. Discrimination models using variance-stabilizing transformation of metabolomic NMR data. OMICS. 2004; 8:118–130. https://doi.org/10.1089/1536231041388348 PMID: 15268771

63. Principal coordinate analysis and non-metric multidimensional scaling. Statistics for Biology and Health. New York, NY: Springer New York; 2007. p. 259–264.

64. Rabinowitz GB. An introduction to nonmetric multidimensional scaling. Am J Pol Sci. 1975; 19:343.

65. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008; 9.

66. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018. https://doi.org/10.1038/nbt.4314 PMID: 30531897

67. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. Nat Methods. 2010; 7:S56–S68. https://doi.org/10.1038/nmeth.1436 PMID: 20195258

68. Bonaccorso G. Machine Learning Algorithms. Packt Publishing Ltd. 2017.

69. Huang X, Wu L, Ye Y. A review on dimensionality reduction techniques. Int J Pattern Recognit Artif Intell. 2019; 33:1950017.

70. Kraemer G, Reichstein M, Mahecha M. DimRed and coRanking—unifying dimensionality reduction in R. R J. 2018; 10:342.

71. Irizarry RA. Introduction to Data Science: Data Analysis and Prediction Algorithms with R. CRC Press; 2019.

72. Urpa LM, Anders S. Focused multidimensional scaling: interactive visualization for exploration of high-dimensional data. BMC Bioinformatics. 2019; 20:221. https://doi.org/10.1186/s12859-019-2780-y PMID: 31046657

73. Hanussek M, Blohm M, Kintz M. Can AutoML outperform humans? An evaluation on popular OpenML datasets using AutoML Benchmark. 2020 2nd International Conference on Artificial Intelligence, Robotics and Control. 2020. https://doi.org/10.1145/3448326.3448353

74. García S, Fernández A, Luengo J, Herrera F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Inf Sci. 2010; 180:2044–2064.

75. van de Wiel MA, Berkhof J, van Wieringen WN. Testing the prediction error difference between 2 predictors. Biostatistics. 2009; 10:550–560. https://doi.org/10.1093/biostatistics/kxp011 PMID: 19380517

76. Beaulieu-Prévost D. Confidence Intervals: From tests of statistical significance to confidence intervals, range hypotheses and substantial effects. Tutor Quant Methods Psychol. 2006:11–19. https://doi.org/10.20982/tqmp.02.1.p011

77. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05." Am Stat. 2019; 73: 1–19.

78. Goodman SN. Aligning statistical and scientific reasoning. Science. 2016; 352:1180–1181. https://doi.org/10.1126/science.aaf5406 PMID: 27257246

79. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016; 31:337–350. https://doi.org/10.1007/s10654-016-0149-3 PMID: 27209009

80. Huber W. A clash of cultures in discussions of the P value. Nat Methods. 2016:607–607. https://doi.org/10.1038/nmeth.3934 PMID: 27467722

81. Kunapuli G. Ensemble Methods for Machine Learning. Manning Publications; 2022.

82. Goder A, Filkov V. Consensus clustering algorithms: Comparison and refinement. Proceedings of the Tenth Workshop on Algorithm Engineering and Experiments (ALENEX). Philadelphia, PA: Society for Industrial and Applied Mathematics. 2008;2008:109–117.

83. Shalev-Shwartz S, Ben-David S. The Runtime of Learning. Understanding Machine Learning. p. 73–86. https://doi.org/10.1017/cbo9781107298019.009

84. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer Science & Business Media; 2017.

85. Efron B, Hastie T. Computer age statistical inference: Algorithms, evidence, and data science. Cambridge University Press; 2016.

86. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005; 67:301–320.

87. Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J. Evaluation of the lasso and the elastic net in genome-wide association studies. Front Genet. 2013; 4:270. https://doi.org/10.3389/fgene.2013.00270 PMID: 24363662

88. Agrawal T. Hyperparameter Optimization in Machine Learning. 2021. https://doi.org/10.1007/978-1-4842-6579-6

89. Frohlich H, Zell A. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. Proceedings 2005 IEEE International Joint Conference on Neural Networks. 2005. IEEE; 2006. https://doi.org/10.1109/ijcnn.2005.1556085

90. Cawley GC, Talbot NLC. Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters. J Mach Learn Res. 2007; 8:841–861.

91. van Erp S, Oberski DL, Mulder J. Shrinkage priors for Bayesian penalized regression. J Math Psychol. 2019; 89:31–50.

92. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media; 2013.

93. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A. 2002; 99:6562–6566. https://doi.org/10.1073/pnas.102102699 PMID: 11983868

94. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007; 99:147–157. https://doi.org/10.1093/jnci/djk018 PMID: 17227998

95. Breiman L. Stacked regressions. Mach Learn. 1996; 24:49–64.

96. Rauschenberger A, Glaab E, van de Wiel MA. Predictive and interpretable models via the stacked elastic net. Bioinformatics. 2021; 37:2012–2016. https://doi.org/10.1093/bioinformatics/btaa535 PMID: 32437519

97. Genuer R, Poggi J-M. Random Forests with R. Springer. Nature. 2020.

98. Classification: Practice—Random Forest. 2018. https://doi.org/10.4135/9781526469144

99. Diaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7:3. https://doi.org/10.1186/1471-2105-7-3 PMID: 16398926

100. Aben N, Vis DJ, Michaut M, Wessels LFA. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. Bioinformatics. 2016; 32:i413–i420. https://doi.org/10.1093/bioinformatics/btw449 PMID: 27587657

101. Rauschenberger A, Ciocănea-Teodorescu I, Jonker MA, Menezes RX, van de Wiel MA. Sparse classification with paired covariates. Adv Data Anal Classif. 2020; 14:571–588.

102. van de Wiel MA, Lien TG, Verlaat W, van Wieringen WN, Wilting SM. Better prediction by use of co-data: adaptive group-regularized ridge regression. Stat Med. 2016; 35:368–381. https://doi.org/10.1002/sim.6732 PMID: 26365903

103. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern C Appl Rev. 2012; 42:463–484.

104. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from Imbalanced Data Sets. Springer; 2018.

105. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. J Artif Intell Res. 2018; 61:863–905.

106. Brownlee J. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery; 2020.

107. Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. Calibration of machine learning models. Handbook of Research on Machine Learning Applications and Trends. IGI Global. 2010:128–146.

108. Meroufel H. Earth Observation Department, Centre of Space Techniques, Algeria. Comparative Study between Validity Indices to Obtain the Optimal Cluster. Int J Comput Electr Eng. 2017:343–350. https://doi.org/10.17706/ijcee.2017.9.1.343–350

109. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics. 2005; 21:3201–3212. https://doi.org/10.1093/bioinformatics/bti517 PMID: 15914541

110. Bruhns S. An Empirical Study of Performance Metrics for Classifier Evaluation in Machine Learning. 2008.

111. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Springer; 2015.

112. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010; 21:128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2 PMID: 20010215

113. Kaymak U, Ben-David A, Potharst R. The AUK: A simple alternative to the AUC. Eng Appl Artif Intell. 2012:1082–1089. https://doi.org/10.1016/j.engappai.2012.02.012

114. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol. 2017; 17:53. https://doi.org/10.1186/s12874-017-0332-6 PMID: 28388943

115. Bilal E, Dutkowski J, Guinney J, Jang IS, Logsdon BA, Pandey G, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. PLoS Comput Biol. 2013; 9: e1003047. https://doi.org/10.1371/journal.pcbi.1003047 PMID: 23671412

116. Herrmann M, Probst P, Hornung R, Jurinovic V, Boulesteix A-L. Large-scale benchmark study of survival prediction methods using multi-omics data. Brief Bioinform. 2021;22. https://doi.org/10.1093/bib/bbaa167 PMID: 32823283

117. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagn Progn Res. 2017; 1:19. https://doi.org/10.1186/s41512-017-0020-3 PMID: 31093548

118. Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. J Am Stat Assoc. 1997; 92:548.

119. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal. 2009; 53:3735–3745.

120. Gronau QF, Wagenmakers E-J. Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. Comput Brain Behav. 2019; 2:1–11. https://doi.org/10.1007/s42113-018-0011-7 PMID: 30906917

121. Braga-Neto U, Dougherty E. Bolstered error estimation. Pattern Recogn. 2004; 37:1267–1281.

122. Sima C, Braga-Neto UM, Dougherty ER. High-dimensional bolstered error estimation. Bioinformatics. 2011; 27:3056–3064. https://doi.org/10.1093/bioinformatics/btr518 PMID: 21914630

123. Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer. 2021; 21:199–211. https://doi.org/10.1038/s41568-020-00327-9 PMID: 33514930

124. Kuhn M, Johnson K. Applied Predictive Modeling. 2013. https://doi.org/10.1007/978-1-4614-6849-3

125. Hackeling G. Mastering Machine Learning with Scikit-Learn. 2nd ed. 2017.

126. Lantz B. Machine Learning with R: Expert techniques for predictive modeling. 3rd ed. Packt Publishing Ltd; 2019.

127. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy, Institute of Medicine. Evolution of Translational Omics: Lessons Learned and the Path Forward. In: Micheel CM, Nass SJ, Omenn GS, editors. Washington (DC): National Academies Press (US); 2014.

128. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta. 2014; 427:49–57. https://doi.org/10.1016/j.cca.2013.09.018 PMID: 24076255

129. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics. 2014; 15:91. https://doi.org/10.1186/1471-2105-15-91 PMID: 24678608

130. Cardoso AL, Fernandes A, Aguilar-Pimentel JA, de Angelis MH, Guedes JR, Brito MA, et al. Towards frailty biomarkers: Candidates from genes and pathways regulated in aging and age-related diseases. Ageing Res Rev. 2018; 47:214–277. https://doi.org/10.1016/j.arr.2018.07.004 PMID: 30071357

131. Glaab E. Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification. Brief Bioinform. 2016; 17:440–452. https://doi.org/10.1093/bib/bbv044 PMID: 26141830

132. Ilyin SE, Belkowski SM, Plata-Salamán CR. Biomarker discovery and validation: technologies and integrative approaches. Trends Biotechnol. 2004; 22:411–416. https://doi.org/10.1016/j.tibtech.2004.06.005 PMID: 15283986

133. Loh W-Y. Fifty Years of Classification and Regression Trees. Int Stat Rev. 2014:329–348. https://doi.org/10.1111/insr.12016

134. Berk RA. Classification and Regression Trees (CART). Statistical Learning from a Regression. Perspective. 2016:129–186. https://doi.org/10.1007/978-3-319-44048-4_3

135. Frank E, Witten IH. Generating Accurate Rule Sets Without Global Optimization. 2008.

136. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. PLoS ONE. 2012; 7:e39932. https://doi.org/10.1371/journal.pone.0039932 PMID: 22808075

137. Trabelsi S, Elouedi Z. Learning decision rules from uncertain data using rough sets. Computational Intelligence in Decision and Control. 2008. https://doi.org/10.1142/9789812799470_0018

138. Gopalakrishnan V, Lustgarten JL, Visweswaran S, Cooper GF. Bayesian rule learning for biomedical data mining. Bioinformatics. 2010:668–675. https://doi.org/10.1093/bioinformatics/btq005 PMID: 20080512

139. Lazzarini N, Widera P, Williamson S, Heer R, Krasnogor N, Bacardit J. Functional networks inference from rule-based machine learning models. BioData Mining. 2016. https://doi.org/10.1186/s13040-016-0106-4 PMID: 27597880

140. Wang H, Sham P, Tong T, Pang H. Pathway-Based Single-Cell RNA-Seq Classification, Clustering, and Construction of Gene-Gene Interactions Networks Using Random Forests. IEEE J Biomed Health Inform. 2020; 24:1814–1822. https://doi.org/10.1109/JBHI.2019.2944865 PMID: 31581101

141. Mallavarapu T, Hao J, Kim Y, Oh JH, Kang M. Pathway-based deep clustering for molecular subtyping of cancer. Methods. 2020; 173:24–31. https://doi.org/10.1016/j.ymeth.2019.06.017 PMID: 31247294

142. Li X-Y, Xiang J, Wu F-X, Li M. NetAUC: A network-based multi-biomarker identification method by AUC optimization. Methods. 2021. https://doi.org/10.1016/j.ymeth.2021.08.001 PMID: 34364986

143. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems. 2017. p. 4768–4777.

144. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the predictions of any classifier. arXiv [cs.LG]. 2016. http://arxiv.org/abs/1602.04938

145. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery; 2013. p. 623–631.

146. Alaa AM, van der Schaar M. Demystifying Black-box Models with Symbolic Metamodels. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2019.

147. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2020; 23. https://doi.org/10.3390/e23010018 PMID: 33375658

148. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020; 58:82–115.

149. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020; 20:310.

150. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016; 3. https://doi.org/10.1186/s40537-016-0043-6

151. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health. 2021; 3:e195–e203. https://doi.org/10.1016/S2589-7500(20)30292-2 PMID: 33478929

152. U.S. Food and Drug Administration. Good machine learning practice for medical device development. In: U.S. Food and Drug Administration [Internet]. 2021 Oct 27 [cited 2022 Apr 5]. Available from: https://www.fda.gov/media/153486/download.

153. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, et al. Criteria for the use of omics-based predictors in clinical trials. Nature. 2013; 502:317–320. https://doi.org/10.1038/nature12564 PMID: 24132288

154. Assessing Credibility of Computational Modeling Through Verification and Validation: Application to Medical Devices. Am Soc Mech Eng. 2018.

155. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015; 13:1. https://doi.org/10.1186/s12916-014-0241-z PMID: 25563062