PERMIT – PERsonalised MedicIne Trials

Guidelines on validation of AI stratification

Work Package 4 - Deliverable 4.2

Deliverable no	4.2
Deliverable Title	Guidelines on validation of AI stratification
Contractual delivery month	July 2021
Responsible Partner	ELIXIR-LU/UNILU
Author(s)	Enrico Glaab, Armin Rauschenberger (Luxembourg Centre for Systems Biomedicine, University of Luxembourg)
Other Contributors	Rita Banzi, Chiara Gerardi, Vanna Pistotti

Executive summary

Background
Validating artificial intelligence (AI) models for patient sub-group stratification robustly and reliably, and ensuring their biological interpretability, involves a wide range of organizational, experimental, statistical and knowledge integration challenges. The choice of suitable methodologies depends strongly on the study type and study goals, and on the specific characteristics of the studied disease indication (e.g. the clinical and molecular heterogeneity of the disease, the numbers and relative sizes of known or putative disease sub-groups, the accessibility of relevant tissue and body fluid biosamples from patients and control subjects, among others). A structured overview of the common challenges and limitations to address in the cross-validation, external validation and interpretation of biomedical AI models, and a detailed discussion of expert-based recommendations to circumvent and address these challenges, are still lacking.

PERMIT has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N. 874825

Methods

Guided by a recent scoping review of the biomedical literature on validation methods for AI-based stratification projects, expert consultation workshops including biostatisticians, machine learning scientists and bioinformaticians working at the interface of AI and biomedicine were held to identify key practical challenges in the validation and biological interpretation of AI-based stratification models, and propose structured guidelines and recommendations to address these challenges.

Results

A chronologically structured collection of the main practical challenges occurring in the validation and design of explainable AI-based stratification models was devised during the workshops, and practical guidelines and common errors to avoid were identified for each challenge. This structured list of recommendations may help to facilitate the generation of biologically interpretable models, circumvent and alleviate common shortcomings and limitations in AI-based biomarker validation, and provide a basis to design study-type specific minimum documentation and reporting guidelines. The generic recommendations are complemented by relevant references to sub-topic specific guidelines and technical methodology publications, to provide a starting point and overview for the reader on how to obtain further information on domain-specific and technical aspects in the implementation of computational validation pipelines.

DISCLAIMER

This document contains information which is the proprietary to the PERMIT Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to any third party, in whole or in parts, except with prior written consent of the PERMIT Coordinator.

The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any particular purpose. The document reflects only the authors' view. The PERMIT Consortium is not responsible for any use that may be made of the information it contains. The user uses the information at its solo risk and liability.

Document log

Issue	Date (yyyy-mm-dd)	Comment	Author/partner
1	2022 – 02 -11		UNILU – Enrico Glaab
2	2022-11-23	Table of contents has been revised	UNILU- Enrico Glaab

Table of Contents

PERMIT has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N. 874825

Background

In biomarker development projects, the design of suitable artificial intelligence (AI) models is often considered as the main challenge, whereas the subsequent model validation is more commonly regarded as a standard task that does not involve complex preparations and method choices. However, validating AI-based stratification models involves several interdependent tasks and non-obvious decisions: (1) Ensuring that available cohorts and data are representative and used efficiently, (2) guaranteeing sufficient robustness and reliability of the performance estimates for each assessed modeling approach, (3) extracting the required information on strengths and weaknesses of different modeling methods for benchmarking (e.g. in terms of different error types and robustness against outliers), and (4) interpreting multivariate biomarker signatures and assessing their biological plausibility and clinical utility. Addressing all these tasks effectively is a multi-disciplinary challenge that involves organizational, experimental, statistical and knowledge integration skills, and the buildup of a corresponding team that is able to cross communication barriers between different relevant disciplines. Furthermore, the choice of adequate experimental and statistical methodologies for a reliable and cost-effective validation of AI-based patient stratification models depends strongly on the study type, study objectives and characteristics of the investigated disease (e.g. the inter-individual and sub-group heterogeneity of the disease in terms of clinical, molecular and imaging characteristics, the numbers and relative sizes of known or putative disease sub-types, the accessibility of relevant tissue and body fluid biosamples from patients and control subjects, among others).

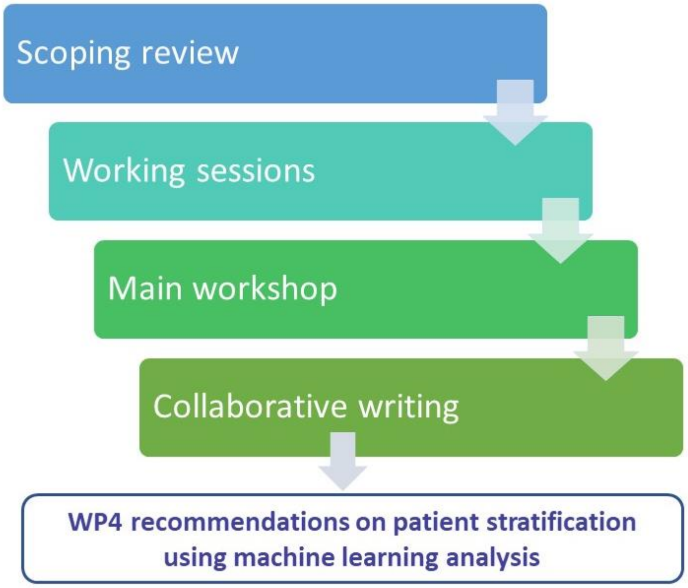
In this report, we have therefore integrated the information obtained from a prior scoping survey of the relevant literature and dedicated expert consultation workshops to provide a structured overview of some of the most frequent challenges in the validation and biological interpretation of AI-based patient stratification models. Apart from addressing common limitations, pitfalls and performance bottlenecks in AI-based stratification for each identified challenge, we provide the consensus recommendations from the expert consultation workshops to circumvent, alleviate or solve these challenges. Specifically, we present an overview of different quantitative validation metrics, cross-validation and external validation schemes for common diagnostic or prognostic model assessment tasks, and suggest criteria on how to select, compare or combine different approaches (including pointers to relevant software packages, workflows and existing guidelines). In addition to aspects relevant for the implementation of a validation study, we discuss considerations for the early planning phase, such as prior statistical power estimation and the representativeness of the validation cohort and data. We also cover the topic of model interpretability and explainability, which is an essential part of the model assessment for most biomedical applications, significantly influences the model selection, and is closely linked to the evaluation of the predictive model performance. Examples of interpretable “white-box” modeling approaches are given, which avoid the complexity of commonly used “black-box” models and help to ensure the transparency, biological plausibility and trustworthiness of prediction models.
Finally, we provide an outlook on initiatives for workflow automation and standardization to facilitate model evaluation in the future, and structured machine learning approaches that aim to provide more biologically interpretable and robust biomarker signature models.

For each discussed challenge and recommendation for biomarker signature validation and interpretation, the reader is also referred sub-topic specific literature on the covered aspects, in order to provide an initial guidance to investigate project- and application-specific aspects in more detail. While the present report focuses on methods for validating AI models for patient stratification, a further dedicated report is provided on methods for building corresponding AI models (see deliverable 4.1).

Approaches (Methods)

We followed four steps in developing the final recommendations: 1) Scoping review of the literature, 2) Working sessions with experts in the field, 3) Workshop with a large group of stakeholders (main workshop), 4) Collaborative writing of final set of recommendations (**Figure 1**).

Figure 1. Development process


Scoping review
Working sessions
Main workshop
Collaborative writing
WP4 recommendations on patient stratification using machine learning analysis

2.1 Scoping review

We first conducted a scoping review, which was previously reported in the deliverable D2.3, to collect available evidence on the use of machine learning for patient stratification in the context of Personalised Medicine (PM). The key findings of the scoping review, covering the biomarker discovery studies using machine learning analysis of omics data which have led to clinically validated diagnostic and prognostic tools, have been published in the journal BMJ Open [1].

2.2 Working sessions

We conducted two working sessions with experts in the PM field, which aimed to discuss the main features of 1) machine learning methods for stratification; 2) validation methods for stratification models. Moreover, a preliminary joint meeting organised by WP4 and WP3 focused on sample size calculation.

2.2.1 Participants

An initial list of stakeholders of the PM ecosystem was created by the PERMIT Project Manager (PG) and then shared with all WP coordinators for further suggestions on expert names related to their WP domain. From the shared list, WP4 coordinator selected some experts in PM and domain experts in biostatistics, machine learning and bioinformatics, who had prior experiences in artificial intelligence-based stratification. These experts were involved both in the working sessions on machine learning methods for stratification (see deliverable 4.1), and the workshop on validation methods (see section 2.2.2 below).

2.2.2 Key points/questions for the discussion in working session 2: validation methods for stratification models

The preparatory material for the second working session was included in the 6-page document provided to the meeting participants prior to working session I (see deliverable 4.1, Appendix 1), covering background information and relevant literature references on both the generation and validation of machine learning methods for patient stratification. The main questions addressed in this workshop were analogous to those in working session 1, but referring to machine learning validation schemes and metrics instead of modelling methods, i.e.:

- Where do the participants see the main gaps and limitations in current validation schemes and performance metrics for machine learning based patient stratification models?
- Do the findings from the literature scoping review on validation workflows presented at the beginning of the meeting match with the participants' experiences, and should the inclusion and structuring of covered topics on gaps, pitfalls and limitations in validation methods and scoring metrics be revised or extended?
- Which generic methodologies and measures would the stakeholders recommend to avoid common pitfalls in biomedical machine learning validation studies, and which best practices, existing guidelines, standardization efforts, or generic advice should be considered to improve quality in future studies?

2.3 Main workshop

The workshop aimed to agree on and discuss the final selection, structuring and phrasing of the proposed key challenges and recommendations on machine learning and validation methods for patient stratification. In particular, the main agreed discussion topics were the specific chronological structuring of recommendations by study phase (1. Planning phase: Study design; 2. Discovery phase: Data collection & pre-processing; 3. Discovery phase: Model building & optimization; 4. Validation phase: Evaluation & interpretation), the inclusion/exclusion and filling of gaps for the working session derived generic best practice recommendations and pitfalls/issues to avoid for each of the study phases, and the refinement of the recommendation formulations in terms of highlighting key aspects and finalizing the included references to existing technical method explanations and guideline documents.

2.3.1 Participants

The following list provides information on the names, affiliations and the domain expertise for the invited workshop participants:

Name	Affiliation	Country	Domain expertise
Anne-Laure Boulesteix	Ludwig-Maximilians-University Munich	Germany	biostatistics and machine learning
Francisco Azuaje	Genomics England	United Kingdom	biomedical data science, artificial intelligence, AI for healthcare
Holger Fröhlich	Fraunhofer Institute for Algorithms and Scientific Computing SCAI	Germany	artificial intelligence, biomedical data science, bioinformatics
Isabel A. Nepomuceno Chamorro	Universidad Pablo de Olavide	Spain	data analytics science, bioinformatics, data mining
Petr Nazarov	Luxembourg Institute of Health	Luxembourg	biostatistics, machine learning, bioinformatics
Paolo Frasconi	University of Florence	Italy	machine learning, computer science
Ramon Diaz-Uriarte	Universidad Autónoma de Madrid	Spain	bioinformatics, computational biology, statistical computing
Rosalba Giugno	University of Verona	Italy	bioinformatics, systems biology, data mining

Armin Rauschenberger	University of Luxembourg	Luxembourg	biostatistics, machine learning (representative of PERMIT WP4)
Enrico Glaab	University of Luxembourg	Luxembourg	bioinformatics, machine learning, biostatistics (representative of PERMIT WP4)

Table 1: List of invited workshop participants with affiliations and domain expertise information.

2.3.2 Development of the questions to address in the workshop

As the invited participants had already taken part in the working sessions for WP4 and received the relevant pre-reading material for these sessions (see Annex I), no further preparatory material was circulated. Based on the discussions carried out during the working sessions, the coordinator of WP4 drafted the questions to be addressed in the main workshop, which are reported in Table 2 below.

1) Planning phase & study design:
   - Which main mitigation strategies should be proposed to address insufficient statistical power in the study design?
   - How can informative and uninformative censoring in stratification studies (e.g. due to dropouts) be addressed most effectively?
   - Which methodological recommendations can be given to handle imbalances in the study groups (e.g. small relative numbers of control subjects recruited)?

2) Data collection & pre-processing phase:
   - Which generic measures can be recommended to promote adequate quality control in AI stratification studies?
   - Which most common errors and pitfalls to avoid in data pre-processing and filtering should be highlighted?
   - How should researchers assess the appropriateness of normalization methods for high-dimensional experimental data used in AI stratification projects?

3) Model building & discovery phase:
   - Which best practices for model selection can be recommended to identify the most robust and predictive modelling approach for a particular stratification task among a selection of candidate methods?
   - Which approaches should be proposed to prevent model overfitting and underfitting?
   - How should researchers optimize an AI model and select suitable performance metric(s)?

4) Model validation & interpretation phase:
   - How can researchers ensure the robustness of a validation study for AI stratification?
   - Which are the main common methodological errors to avoid in AI model validation projects?

- Which generic practical recommendations can be given to ensure that AI stratification models are sufficiently interpretable and explainable?

Table 2. Questions designed to guide the discussions on methodology recommendations for AI-based stratification in the main workshop

2.4 Collaborative writing and technical validation

The WP4 coordinators and a subset of experts who participated in the workshop drafted the list of the recommendations, working collaboratively on a joint Google Document. The aim was to reach a consensus on the selection, priority and wording of a text covering the ten most relevant generic challenges for AI-based patient stratification projects, and practical recommendations to address each of these challenges (using the “Ten simple rules” format of the PLoS Computational Biology journal). The rules were structured chronologically to reflect the different project phases in stratification studies, starting with the planning and study design phase and ending with the validation phase and generic study-phase independent recommendations.

2.5 Implementation

A workshop, which is planned for March 2022, will be organized to discuss with all stakeholders how the recommendations developed will be implemented. Representatives from all key stakeholders, within the consortium and beyond, will be engaged in the preparation of this interactive workshop, and will be asked to describe how the recommendations can be integrated and implemented in their particular field as well as what measures should be taken to increase dissemination and uptake.

Results

The following table (Table 3) covers the identified common challenges and risks in the validation of AI-based patient stratification models, and the possible mitigation strategies proposed during the expert consultation workshops to address these challenges. The covered challenges reflect the validation phase, and general risk affecting all study phases (the initial planning and model building phases are discussed in a dedicated report as part of deliverable 4.1). As indicated in brackets in the table, the expert discussions and collected results covered both challenges and recommendations specific to supervised AI learning methods, and unsupervised for stratification. The likelihood and impact of the identified risks/challenges were rated jointly by consensus among the experts on a scale on a three-point scale (low, medium or high; see columns 2 and 3 in Table 3). For each challenge/risk, recommended mitigation strategies were also identified by the experts, or collected from the prior review of the machine learning literature (deliverable 2.3) and then validated by the experts during the consultation workshops. In addition to the collection or challenges and risks specific to the validation stage of AI stratification projects, a further list of the key study-phase independent challenges/risks was also created, using the same format (see Table 4).

Challenge/Risk	Likelihood (low, medium or high)	Impact (low, medium or high)	Mitigation strategies
Validation phase: Model performance assessment, evaluation of external generalization capacity, and interpretation			
Model performance assessment is not robust enough (high standard deviation in cross-validated performance estimates, specific to supervised methods)	medium to high	medium to high	• Use robust bootstrapping or cross-validation techniques (e.g. bolstered cross-validation) and multiple performance metrics to assess predictive performance • Ensure a sufficient number of samples is available for performance assessment through a prior power estimation
The predictive model does not generalize across different cohorts and populations (specific to supervised methods)	high	high	• Plan an external validation on a distinct cohort (covering distinct continents / ethnic backgrounds) • Consider both the discovery study and the external validation study in the prior power estimation to enable both robust model building and robust validation of the generalization capability • Plan for a targeted external model validation using more sensitive measurement techniques (e.g. qRT-PCR, digital PCR for transcriptomics studies) • Consider a meta-analysis of relevant public or collaborator-derived omics datasets from other cohorts for feature selection prior to model building
Insufficient model interpretability / explainability	medium to high	low to high	• When interpretability and explainability are relevant objectives and metrics of success of the study, choose "white-box" learning algorithms (understanding that the best predictive performance might be achieved by algorithms and methods with low interpretability).

Overoptimistic assessment of omics-data utility	medium to high	high	• Consider structured machine learning approaches guided by prior biological knowledge from cellular pathways and molecular networks to build more biologically interpretable models • Include traditional clinical data in predictive model building and assess the clinical utility (added value for decision making) of omics data relative to the clinical data.

Table 3: Identified challenges and risks during the planning and discovery phases of AI-based patient stratification projects (column 1), including a qualification of risk likelihood (column 2) and impact (column 3), and proposed mitigation strategies (column 4).

Challenge/Risk	**Likelihood** (low, medium or high)	**Impact** (low, medium or high)	**Mitigation strategies**
Generic risks affecting multiple study phases			
Ineffective communication between collaborators from different disciplines	medium to high	high	• Regular meetings with documented results and action points • Appointment of shared research group members with interdisciplinary expertise as mediators between groups from different disciplines • Introductory presentations on the methodologies and challenges in project-relevant disciplines to create awareness of limitations, threats and opportunities for the collaboration
Personal data security breach	low	high	• Data protection and security training is given to all project-assigned personnel • Appointment of a data security officer • Data handling procedures are planned and implemented by strictly following all relevant data protection regulations (e.g. European Union's General Data Protection Regulation, GDPR)

Hiring of personnel for the study is delayed	medium	medium to high	• Plan sufficient time for hiring in between the study design phase and the study implementation phase • Ensure that a back-up plan is in place in case relevant personnel cannot be hired in time (e.g. reallocation of tasks and timelines for existing personnel)
Dropout of personnel during the study (e.g. due to illness, job change)			• Ensure that for each key task in the project reserve personnel is available in the case of a dropout • Ensure sufficient buffer time in the project for delays due to dropouts

Table 4: Identified generic and study phase independent challenges and risks in AI-based patient stratification projects (column 1), including a qualification of risk likelihood (column 2) and impact (column 3), and proposed mitigation strategies (column 4).

Discussion and Conclusions

The discussions with machine learning and data science experts on the basis of a prior literature review revealed several common challenges in the validation and interpretation of AI-based patient stratification approaches, and provided best practice suggestions and general recommendations on how to circumvent or alleviate frequent issues and limitations. While widely accepted cross-validation and external validation schemes are already available, and first standards for AI model validation, verification and documentation have been proposed [2–4], these standards do not cover many of the possible pitfalls in early data processing and filtering stages, lack a discussion of novel bootstrapping and bolstered error estimation approaches, and no widely accepted guidelines, recommendations or standards are available for ensuring model interpretability and explainability.

Similar to AI-based model building (see deliverable 4.1), the choice of suitable model validation and interpretation methods strongly depends on the specific study design, the disease indication and goals of the study, which may require adaptions of generic validation methodologies or even a study-specific design of the model evaluation and interpretation approaches. However, both the scoping review of the relevant literature and the experiences shared by the invited experts during the consultation workshops confirmed that generic validation methodologies, guidelines and practical recommendations are applicable to the great majority of AI-based stratification studies, and common risks, threats and related mitigation measures have general relevance beyond individual diseases or specific study designs.

The following recommendations, grouped by research phase, reflect the key conclusions on generic validation and interpretation methodologies from the expert consultation workshops:

1.) Model performance assessment

The evaluation of AI stratification models involves multiple criteria, including the predictive performance, robustness, replicability on external datasets / cohorts, biological plausibility and interpretability, and utility in clinical decision making in terms of benefit/risk relation. However, the first and most important validation criterion for the assessment of new, tentative AI models is the predictive performance, which ultimately determines whether the model has an additive informative value for practical diagnostic, prognostic or monitoring applications.

As a first step, a quantitative metric has to be chosen to assess model performance and select the most useful combinations of modeling methods and input data types. The choice of performance metrics is a problem-specific task, and it is often recommendable to consider multiple metrics to distinguish between different error types (e.g. type 1 vs. type 2 error) and consider different penalties for outliers (e.g. quadratic vs. non-quadratic loss functions). This is particularly important for imbalanced study groups [5], often observed in biomedical studies (e.g. identifying ~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- or under-sampling or data augmentation methods (test set samples should however always be independent from the training set and synthetic redundancy introduced by oversampling should be avoided) [6–8]. Moreover, a prior sample size estimation and clearly defined study goals can help to ensure that a sufficient number of samples for each study group is available for both modeling and performance assessment. Common performance measure choices include the balanced accuracy and sensitivity/specificity for supervised binary classification, the mean squared error or absolute error for regression tasks, and the internal validity indices Silhouette width or Calinski-Harabasz index for unsupervised clustering [9,10] (note that for clustering tasks, only if a ground truth is known, external validity can be assessed, e.g. using the adjusted Rand index). However, the choice of the performance metric does not only depend on the type of the outcome variable but also the specific analysis goals and applications (see [11] 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 categorization (a problem that affects accuracy, sensitivity and specificity measures). These include deviance/likelihood, 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 be applied to multiple outcome types, including survival data [12,13]. 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 consider the costs of different error types, should be considered.

When estimating a model’s generalization performance from observational data, the variability in biomedical datasets is often very high, due to both technical and biological sources of variation. Bootstrapping (sampling with replacement) methods, such as .632+ bootstrap, can be used as a means to obtain robust performance estimates [14]. Another well-accepted approach is repeated or iterated *k*-fold cross-validation (sampling without replacement), which often gives less biased estimates of the true generalization performance [15]. When selecting the parameter *k*, the user should be aware of the balance between bias (low *k*) and variability (high *k*, e.g. leave-one-out cross-validation) [16]. Bolstered error estimation is a further alternative approach dedicated specifically to datasets with small sample size [17,18]. 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 [19]. More detailed discussion and practical guidance on the use of relevant algorithms and software tools for model performance assessment is provided in [20–22].

2.) Data integration and selection of the most informative data types

In the machine learning literature traditionally three different strategies for multi-modal data integration have been suggested, namely early, intermediate and late integration [23,24]. Early integration methods focus on extraction of common features from several data modalities. A classic example is Canonical Correlation Analysis (CCA) and sparse variants of CCA [88,89]. 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 [25,26].

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 [24]. More recently, multi-modal neural network architectures have been devised [27].

A related, but different problem to data integration is the selection of the most useful data type(s), when multiple available datasets contain redundant information, but have different overall 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 the experimental data provide added value for decision making, which requires comparative evaluations in addition to integrative analysis, and using the traditional clinical data as the baseline [28].

For more detailed guidelines and method comparisons, we refer the reader to a broader overview of machine learning methods for omics data integration [29], representative case studies on combining omics and clinical data [30], and generic multi-omics integration approaches [31].

3.) Model optimization and validation

Depending on the goals of an AI stratification study (e.g., whether the study aims at a clinical biomarker validation or only at preclinical biomarker evaluation) and the study type (e.g., whether the study is prospective or retrospective) different options are available to improve and validate an initial biomarker model obtained from a discovery cohort. Clinical biomarker studies require that the final model derived from a discovery cohort is locked down and recorded before testing the model on an independent validation cohort, and the subjects in the validation cohort have to fulfill the same inclusion and exclusion criteria as for the discovery cohort [32,33]. 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 more on preclinical research and the earlier stages of biomarker discovery have more flexibility in collecting additional information to further optimize and confirm the generalization capability of an initial biomarker model. Apart from straightforward optimization strategies, such as increasing the size of the discovery cohort(s) and thereby the size of the training dataset for modeling, or using more sensitive and targeted measurement technologies for data collection (e.g. replacing high-throughput transcriptomics profiling approaches by targeted qRT-PCR or digital PCR measurements, focusing on pre-selected genes), a wide range of other information sources and experiments can be used 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 used to improve the feature selection for a model [34], or prior knowledge from cellular pathway databases and the biomedical literature can be integrated into the analysis to filter predictive features depending on their involvement in disease-associated pathways [35] or to derive more robust pathway- or network-based predictive features [36]. Furthermore, cellular or animal models for the disease condition of interest can provide additional data for the validation of molecular biomarker features, and functional validation studies involving the modulation of candidate biomarker molecules or pathways via knockdown and overexpression experiments may enable the direct testing of a functional association with measurable disease phenotypes [37]. While all these information sources provide effective means for the initial confirmation and filtering of candidate markers, after the researchers have optimized a biomarker signature and locked down the final machine learning model, the final clinical evaluation will always require a robust external validation on a distinct patient cohort.

4.) Model interpretability and explainability assessment

Depending on the specific goals of a prediction or stratification analysis of biomedical data, the success of a machine learning approach might not only be judged by the quantitative evaluation of the predictive performance of the resulting model, but also by its interpretability and biological insightfulness and plausibility. When interpretability and explainability are relevant objectives and criteria for the success of the study, researchers should choose so-called "white-box" learning algorithms, i.e. simpler modeling approaches which link the input features to the outcome variable of interest in a more transparent and easier to understand fashion than more complex, but often also more accurate “black-box” modeling methods. Although simple white-box models may often achieve a good predictive performance in many applications, especially when the underlying relationship between input and output variables is well captured by simple linear relationships or small sets of intuitively interpretable if-then-else decision rules, investigators need to understand that for many complex datasets the best predictive performance might only be achieved by algorithms with low interpretability, due to an inherent complexity of feature relationships in the data.

However, 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. Prime examples for interpretable learning approaches are linear modeling methods [38], and rule-based machine learning methods, such as classification and regression trees [39,40],

PERMIT has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N. 874825

combinatorial rule learning approaches [41,42], and probabilistic and fuzzy rule learning methods [43,44]. While linear modeling approaches enable the scoring and ranking of features by their absolute weight 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 [45]. Apart from these generic learning methods with a focus on interpretability, more recently, domain-specific prediction and clustering approaches, which exploit prior biological knowledge from cellular pathway and molecular network structures [46–48], have gained interest and traction. While these white-box modeling methods are not required for all applications, being able to understand a stratification or prediction model derived from biomedical data and assess its biological plausibility is 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 the statistical power and clinical benefit of diagnostic or prognostic tools [49].

5.) Ensuring effective cross-disciplinary communication throughout the project

Biomarker development studies involve cross-disciplinary collaboration between experimental researchers, computer scientists and clinicians throughout the duration of the project. Therefore, a key determinant of the study success, which concerns all study phases, is the effectiveness of the communication between all participants. This is particularly important already during the initial study design, where the study plan and goals, and the participants' responsibilities need to be clearly defined, including the planning of contingencies and mitigation measures, to prevent risks and misunderstandings in the implementation phase.

Regular meetings within individual working groups (e.g. on a weekly basis) and between different collaborating groups (e.g. on a monthly basis) with a clear agenda, documented results and action items are essential to align research activities, monitor progress, and jointly plan and decide the next steps. Introductory presentations and interdisciplinary workshops in the early project phases can help to bridge the gaps in domain knowledge between representatives of different experimental, computational and clinical specializations. To complement meetings in person, web-based systems for video conferencing and project management should be set up, and a wide variety of relevant open source software for this purpose is already available [50–54]. In general, computational project management software can facilitate efficient milestone tracking, document and data sharing, and the planning, communication and decision making during the project.

Finally, apart from the communication within the project team, effective communication should also involve external stakeholders, e.g. by organizing outreach activities and information events to engage with patients and their representative organizations, and creating public awareness of ongoing research efforts through online and traditional media coverage.

Summary

The choice of methodologies for the validation and interpretation of AI-based patient stratification models depends both on the study goals and the study type. However, while study-specific model evaluation and interpretation approaches may be required in some cases, generic cross-validation and external training/test set validation schemes are applicable to the great majority of studies. The selection of performance metrics must fit with type of the primary and secondary endpoint variables (differentiating between categorical, numerical or survival outcomes), but it is generally advisable to consider multiple outcome metrics in combination, due to the different limitations associated with individual metrics for the same type of outcome. For categorical outcomes in particular, it is often useful to consider probabilistic predictions and dedicated evaluation metrics for probabilistic outcomes in addition to categorical predictions.

To prevent information leaks and an overoptimistic estimation of a model's generalization error, it is important to carefully separate the test data used for model evaluation from the data used for model improvement, e.g. by applying a nested cross-validation. For the same reason, global supervised feature selection prior to supervised machine learning has to be avoided, and feature selection should either be integrated into the cross-validation, and/or unsupervised feature selection methods should be applied.

Finally, a key aspect for biomedical model evaluation is to ensure the biological interpretability and explainability of the model. A wide variety of dedicated model building algorithms, with a focus on maximizing model interpretability while retaining a high predictive performance, is available for this purpose, and includes rule-based machine learning methods as well as structured learning approaches, which exploit prior information on feature interrelations (e.g. pathway or network associations between biomolecular features in omics datasets). Overall, the main generic recommendations for the validation of AI-based stratification models are to ensure the availability of adequate cohorts and data already in the study design phase using power calculations and clearly defined exclusion and inclusion criteria, to choose modeling approaches with a sufficient degree of model interpretability, and to use widely accepted nested cross-validation and external validation schemes.

Next Steps

The scoping review of the biomedical literature and the discussions during the expert consultation workshops on the validation of AI stratification models revealed several existing best practice recommendations, common errors to avoid, first standardization efforts for the validation and reporting of AI models [2–4], and initial generic suggestions on how to ensure model interpretability and explainability [49,55]. However, to integrate and further develop these recommendations into comprehensive and widely accepted standard guidelines, and to promote their implementation in practice, further community-driven standardization efforts, dissemination activities, software support tools and frameworks, and regulatory and incentive policies are required.

A first objective will be to establish an organizational framework for knowledge exchange and consensus discussions between representatives of different stakeholders in the field, i.e. researchers in AI and biomedicine, regulators, funding bodies, science publishers and science policy makers. A corresponding initiative would help to ensure that a jointly discussed collection of standards and recommendations for biomedical AI models is sufficiently comprehensive, has broad support in the research community, and is regularly updated by stakeholder consensus meetings.

For clinical AI-based stratification models, the development of new regulatory standards in the EU should take into consideration the existing relevant guidelines by the U.S. Food and Drug

Administration, in particular the “Good Machine Learning Practice for Medical Device Development: Guiding Principles” [56], which highlights the importance of leveraging multidisciplinary expertise throughout the total product life cycle for an AI-based diagnostic or medical device. Apart from establishing detailed regulatory standards for clinical validation, introducing standards for the documentation and reporting of preclinical AI-based research on patient stratification could be supported by extensions of journal submission guidelines, as well as adjusted submission and reporting policies for public research grants.

Beyond regulatory standards, a wide range of community-driven, complementary initiatives could help to improve quality of AI model validation projects. Standard software frameworks and easy-to-use workflow automation tools, e.g. using container-based virtualization [57], could greatly facilitate the implementation of scientifically valid, rigorous, and fully reproducible cross-validation and external testing workflows for biomedical AI models. Such workflow automation would also help to promote the sharing and re-use of successful workflows for new applications and automated tracing of data provenance to simplify monitoring and re-evaluation of all data processing stages. Best practice guidelines and study-type specific recommendations on how to avoid common pitfalls and shortcomings in the planning and implementation of AI studies, which go beyond minimum regulatory standards, will need to be further developed and broadly disseminated in journals, scientific magazines, blogs and social media platforms. Finally, similar to the guidelines for AI model building (see deliverable 4.1), the jointly established community guidelines for model validation should also be integrated into current academic research education programmes, to prepare the next generation of scientists, clinicians and regulators for future technological developments and challenges in validating AI-based biomedical tools.

References

1. Glaab E, Rauschenberger A, Banzi R, Gerardi C, Garcia P, Demotes J. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. *BMJ Open*. 2021;11(12):e053674. doi:10.1136/bmjopen-2021-053674
2. ASME. Assessing Credibility of Computational Modeling and Simulation Results through Verification and Validation : Application to Medical Devices. *Asme V&V 40-2018*. 2018:40. https://www.asme.org/products/codes-standards/vv-40-2018-assessing-credibility-computational.
3. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698
4. McShane LM, Cavenagh MM, Lively TG, et al. Criteria for the use of omics-based predictors in clinical trials. *Nature*. 2013. doi:10.1038/nature12564
5. 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(4):463-484. doi:10.1109/tsmcc.2011.2161285
6. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. *Learning from Imbalanced*

Data Sets. Springer; 2018. https://play.google.com/store/books/details?id=8Fp0DwAAQBAJ LB - BQbS.

7. Fernandez A, Garcia S, Herrera F, Chawla N V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J Artif Intell Res*. 2018;61:863-905. doi:10.1613/jair.1.11192
8. Brownlee J. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery; 2020. https://books.google.com/books/about/Imbalanced_Classification_with_Python.html?hl=&id=jaXJDwAAQBAJ LB - EkBb.
9. Meroufel H, Department EO, Techniques C of S, Algeria. Comparative Study between Validity Indices to Obtain the Optimal Cluster. *Int J Comput Electr Eng*. 2017;9(1):343-350. doi:10.17706/ijcee.2017.9.1.343-350
10. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics*. 2005;21(15):3201-3212. doi:10.1093/bioinformatics/bti517
11. Bruhns S. *An Empirical Study of Performance Metrics for Classifier Evaluation in Machine Learning*.; 2008. https://books.google.com/books/about/An_Empirical_Study_of_Performance_Metric.html?hl=&id=KMZIQwAACAAJ LB - EI3i.
12. Kaymak U, Ben-David A, Potharst R. The AUK: A simple alternative to the AUC. *Eng Appl Artif Intell*. 2012;25(5):1082-1089. doi:10.1016/j.engappai.2012.02.012
13. 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(1):53. doi:10.1186/s12874-017-0332-6
14. Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. *J Am Stat Assoc*. 1997;92(438):548. doi:10.2307/2965703
15. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal*. 2009;53(11):3735-3745. doi:10.1016/j.csda.2009.04.009
16. Gronau QF, Wagenmakers E-J. Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. *Comput Brain Behav*. 2019;2(1):1-11. doi:10.1007/s42113-018-0011-7
17. Braga-Neto U, Dougherty E. Bolstered error estimation. *Pattern Recognit*. 2004;37(6):1267-1281. doi:10.1016/j.patcog.2003.08.017
18. Sima C, Braga-Neto UM, Dougherty ER. High-dimensional bolstered error estimation. *Bioinformatics*. 2011. doi:10.1093/bioinformatics/btr518
19. 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(3):199-211. doi:10.1038/s41568-020-00327-9
20. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media; 2013. https://play.google.com/store/books/details?id=yPfZBwAAQBAJ LB - iXxV.
21. Kuhn M, Johnson K. Applied Predictive Modeling. 2013. doi:10.1007/978-1-4614-6849-3
22. Hackeling G. *Mastering Machine Learning with Scikit-Learn, Second Edition*.; 2017. https://books.google.com/books/about/Mastering_Machine_Learning_with_Scikit_L.html?hl=&id=S2sYtAEACAAJ LB - k32i.

23. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform*. 2018;19(2):325-340. doi:10.1093/bib/bbw113

24. Support vector machine applications in computational biology. In: *Kernel Methods in Computational Biology*. The MIT Press; 2004. doi:10.7551/mitpress/4057.003.0005

25. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5(2):241-259. doi:10.1016/s0893-6080(05)80023-1

26. Džeroski S, Ženko B. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Mach Learn*. 2004;54(3):255-273. doi:10.1023/b:mach.0000015881.36452.6e

27. Gao J, Li P, Chen Z, Zhang J. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Comput*. 2020;32(5):829-864. doi:10.1162/neco_a_01273

28. 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(1):162. doi:10.1186/s12874-019-0802-0

29. Zhou W. *Machine Learning Methods for Omics Data Integration*.; 2011. https://books.google.com/books/about/Machine_Learning_Methods_for_Omics_Data.html?hl=&id=4lh7AQAACAAJ LB - a3q9.

30. 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(30):5310-5329. doi:10.1002/sim.6246

31. Hardiman G. *Systems Analytics and Integration of Big Omics Data*. MDPI; 2020. https://books.google.com/books/about/Systems_Analytics_and_Integration_of_Big.html?hl=&id=9ufcDwAAQBAJ LB - nW7W.

32. Trials C on the R of O-BT for PPO in C, Services B on HC, Policy B on HS, Medicine I of. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Washington (DC): National Academies Press (US); 2014. doi:10.17226/13297

33. Horvath AR, Lord SJ, StJohn A, et al. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta*. 2014;427:49-57. doi:10.1016/j.cca.2013.09.018

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

35. Cardoso AL, Fernandes A, Aguilar-Pimentel JA, et al. Towards frailty biomarkers: Candidates from genes and pathways regulated in aging and age-related diseases. *Ageing Res Rev*. 2018;47:214-277. doi:10.1016/j.arr.2018.07.004

36. Glaab E. Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification. *Brief Bioinform*. 2015. doi:10.1093/bib/bbv044

37. Ilyin SE, Belkowski SM, Plata-Salamán CR. Biomarker discovery and validation: technologies and integrative approaches. *Trends Biotechnol*. 2004;22(8):411-416. doi:10.1016/j.tibtech.2004.06.005

38. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. Springer Science & Business Media; 2013. https://play.google.com/store/books/details?id=qcI_AAAAQBAJ LB - lSZk.

39. Berk RA. Classification and Regression Trees (CART). *Stat Learn from a Regres Perspect*. 2016:129-186. doi:10.1007/978-3-319-44048-4_3

40. Loh W-Y. Fifty Years of Classification and Regression Trees. *Int Stat Rev*. 2014;82(3):329-348.

doi:10.1111/insr.12016

41. Frank E, Witten IH. *Generating Accurate Rule Sets Without Global Optimization*.; 2008. https://books.google.com/books/about/Generating_Accurate_Rule_Sets_Without_Gl.html?hl=&id=T85tzgEACAAJ LB - 3yxf.
42. 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(7):e39932.
43. Trabelsi S, Elouedi Z. LEARNING DECISION RULES FROM UNCERTAIN DATA USING ROUGH SETS. *Comput Intell Decis Control*. 2008. doi:10.1142/9789812799470_0018
44. Gopalakrishnan V, Lustgarten JL, Visweswaran S, Cooper GF. Bayesian rule learning for biomedical data mining. *Bioinformatics*. 2010;26(5):668-675. doi:10.1093/bioinformatics/btq005
45. Lazzarini N, Williamson S, Heer R, Krasnogor N, Bacardit J. Functional networks inference from rule-based machine learning models. *BioData Min*. 2016;9:1-23. doi:10.1186/s13040-016-0106-4
46. 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 Heal Inf*. 2020;24(6):1814-1822. doi:10.1109/JBHI.2019.2944865
47. Mallavarapu T, Hao J, Kim Y, Oh JH, Kang M. Pathway-based deep clustering for molecular subtyping of cancer. *Methods*. 2020;173:24-31. doi:10.1016/j.ymeth.2019.06.017
48. Li X-Y, Xiang J, Wu F-X, Li M. NetAUC: A network-based multi-biomarker identification method by AUC optimization. *Methods*. 2021. doi:10.1016/j.ymeth.2021.08.001
49. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1). doi:10.1186/s12911-020-01332-6
50. Andre E, Le Breton N, Lemesle A, Roux L, Gouaillard A. Comparative Study of WebRTC Open Source SFUs for Video Conferencing. *2018 Princ Syst Appl IP Telecommun*. 2018. doi:10.1109/iptcomm.2018.8567642
51. Open Source Tools for Managing Projects. *Introd to Softw Proj Manag*. 2016:281-288. doi:10.1201/b16534-14
52. Pereira AM, Gonçalves RQ, Von Wangenheim CG, Buglione L. Comparison of open source tools for project management. *Int J Softw Eng Knowl Eng*. 2013;23(02):189-209. doi:10.1142/s0218194013500046
53. Mishra A, Mishra D. Software project management tools. *Softw Eng Notes*. 2013;38(3):1-4. doi:10.1145/2464526.2464537
54. Abramova V, Pires F, Bernardino J. Open source vs proprietary project management tools. In: *New Advances in Information Systems and Technologies*. Advances in intelligent systems and computing. Cham: Springer International Publishing; 2016:331-340. doi:10.1007/978-3-319-31232-3_31
55. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021. doi:10.1016/j.jbi.2020.103655
56. U.S. Food and Drug Administration. Good Machine Learning Practice for Medical Device Development: Guiding Principles. fda.gov. https://www.fda.gov/medical-devices/software-

medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles. Published 2021.

57. Wüst S, Schwerdel D, Müller P. Container-based virtualization technologies. *PIK - Prax der Informationsverarbeitung und Kommun*. 2017. doi:10.1515/pik-2017-0001