Computational and Structural Biotechnology Journal 28 (2025) 281–293

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

An artificial intelligence-based platform for personalized predictions of Metacognitive Training effectiveness

Caroline König [a,*], Pedro Copado [a], Alfredo Vellido [a], Àngela Nebot [a], Cecilio Angulo [b], Maria Lamarca [c,d,e], Vanessa Acuña [f], Fabrice Berna [g], Steffen Moritz [h], Łukasz Gawęda [i], Susana Ochoa [c,d]

[a] Soft Computing Research Group (SOCO), Intelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Centre, Universitat Politècnica de Catalunya (UPC Barcelona Tech), Jordi Girona 1-3, Barcelona, 08034, Spain
[b] Knowledge Engineerig Research Group (GREC), Intelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Centre, Universitat Politècnica de Catalunya (UPC Barcelona Tech), Jordi Girona 1-3, Barcelona, 08034, Spain
[c] MERITT Group, Institut de Recerca Sant Joan de Déu, Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, 08830, Barcelona, Spain
[d] Consorcio de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain
[e] Clinical and Health Psychology Department, School of Psychology, Universitat Autònoma de Barcelona, Bellaterra, 08193, Barcelona, Spain
[f] Departamento de Psiquiatría, Escuela de Medicina, Facultad de Medicina, Universidad de Valparaíso, Valparaíso, Chile
[g] University of Strasbourg, University Hospital of Strasbourg, Inserm, Strasbourg, France
[h] Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, 20246, Germany
[i] Experimental Psychopathology Lab, Institute of Psychology, Polish Academy of Sciences, Warsow, Poland

ARTICLE INFO

Keywords:
Personalized medicine
Explainable artificial intelligence
Feature selection
Fairness
Metacognitive Training
Mental health

ABSTRACT

This study introduces a machine learning (ML)-based platform aimed at predicting the effectiveness of Metacognitive Training (MCT). The platform is meant to function as an experimental prototype in the scope of a clinical research project for a decision support system to assist clinicians in tailoring treatment plans for patients with psychosis. It integrates eight ML models to evaluate MCT effectiveness under a wide range of mental health questionnaires to assess a broad spectrum of psychological symptoms. By incorporating diverse measures, the platform aims to capture a comprehensive understanding of patient profiles, enabling more precise and tailored predictions for treatment personalization. Furthermore, the transparency requirements for artificial intelligence (AI) systems, as outlined in the AI Act regulation of the European Union, are addressed through the implementation of explainable AI models, using post-hoc explanations based on SHAP analysis for each predictive model. Ethical concerns related to ensuring gender-neutral behavior in the system are tackled by conducting a disparate impact analysis, which evaluates biases present in the models enhancing the system's accountability and alignment with ethical and regulatory standards.

1. Introduction

Personalized medicine applications can be understood as data-centric endeavors for which advanced analytical methods such as machine learning (ML) can be used to increase precision in medical care [1,2]. These methods enable the design of customized treatments and improve patient outcomes by aligning medical decisions with individual needs and characteristics. In psychology, the domain of this study, personalized healthcare has already been applied in various mental illness treatments to decide between multiple treatment alternatives [3,4]. Treatments differ by their clinical efficacy and cost-effectiveness, aspects that significantly influence healthcare decision-making for regulators, clinicians, and patients [5]. The precise determination of which individuals are most likely to benefit from a particular treatment and the forecast of cost distribution at the patient level is not straightforward [6], but it is essential to enable the delivery of personalized treatment recommendations.

Mental illnesses are a major public health challenge worldwide, contributing both to the health and economic burdens [7]. In Europe, schizophrenia affects approximately 2% of the population [8], con-

[*] Corresponding author.
E-mail address: caroline.leonore.konig@upc.edu (C. König).

https://doi.org/10.1016/j.csbj.2025.07.051
Received 5 May 2025; Received in revised form 24 July 2025; Accepted 28 July 2025

Available online 5 August 2025

tributing to significant disability and societal costs, particularly due to work absences and early retirement [9]. Although antipsychotic medications are widely prescribed, they often do not improve functional outcomes. Psychological treatments, on the other hand, have shown effectiveness in improving symptoms, even in the absence of medication [10]. One such intervention, Metacognitive Training (MCT) [11,12], was introduced from a psychological perspective to address schizophrenia. MCT aims to reduce positive symptoms by targeting cognitive biases [13]. This training is composed of 10 modules, each addressing key aspects such as attribution style, jumping-to-conclusions bias, confirmation bias, social cognition, false memories, and affective symptoms, along with additional modules focused on self-esteem and stigma [14]. As summarized in the recent meta-review by [15] there is evidence on the effectiveness of MCT in reducing delusions, hallucinations, cognitive biases [16] and patient satisfaction [17]. Additionally, MCT has been found to alleviate negative symptoms to some extent while enhancing self-esteem and overall functioning. Some studies suggest that variables such as gender [18], anxiety, self-esteem, quality of life, and severity level may act as moderators influencing treatment response [19–21].

This study reports the design and development of a ML-based prototype platform aimed at predicting the effectiveness of MCT in the context of the European ERAPERMED 2022-292 research project “Towards a Personalized Medicine Approach to Psychological Treatment of Psychosis”, henceforth referred to as PERMEPSY (www.permepsy.org). The participants of the project are five clinical partners, namely Parc Sanitari San Joan de Déu (PSSJD) from Spain as project coordinator, University Medical Center Hamburg-Eppendorf (UKE) from Germany, Polish Academy of Sciences (PAoS) from Poland, the Universidad de Valparaiso (UV) from Chile, University Hospital of Strasbourg (Inserm) from France, and Universitat Politècnica de Catalunya (UPC) from Spain as a technical partner.

This web-based predictive platform represents an experimental prototype developed within the framework of this research project as a decision support system to assist clinicians in tailoring MCT treatment plans for individuals with psychosis. In line with research on personalized care in mental health, the model offers considerable potential for helping therapists adapt MCT modules to the distinct psychological profiles and therapeutic needs of each patient. By using patient-specific information from individual-level MCT data collected before treatment including psychological, sociodemographic, and diagnostic variables, the system moves beyond standard methods and supports a more personalized, targeted approach for individual MCT therapy. Specific components of the intervention such as module selection or symptom focus can be adjusted based on individual cognitive biases, emotional patterns, and pre-treatment symptom profiles [22]. While the model estimates the overall effectiveness of MCT using standardized questionnaire data, it does not perform automated personalization or module selection. Instead, it serves as a decision support tool, offering therapists relevant insights that may guide clinical judgment when adapting MCT to individual patient needs.

The study outlines the integration of artificial intelligence (AI) technology, namely ML-based methods, for predicting MCT effectiveness. The effectiveness of treatment is evaluated using a wide range of mental health questionnaires to assess a broad spectrum of psychological symptoms so that the predictive platform can capture a comprehensive understanding of patient profiles, allowing for more precise and tailored predictions for treatment personalization. Additionally, safety considerations regarding the use of AI technology in medical applications are taken into account. The European Union (EU) AI Act, adopted in March 2024 by its 27 member states, establishes the world’s first comprehensive legal framework dedicated to AI [23], ensuring that AI systems are human-centered, trustworthy, and ethically aligned. Its primary objective is to safeguard health, safety and fundamental rights while mitigating potential risks associated with AI-driven systems. Since this study serves as a preliminary proof of concept, it was not feasible to address all the regulatory requirements of a high-risk AI system, under which the predictive platform would fall as a medical application. However, measures ensuring transparency and gender fairness were incorporated to align with ethical and legal standards. The EU AI Act establishes transparency requirements to ensure that AI systems are explainable and accountable, particularly in high-risk applications such as healthcare [24]. Advanced ML models are often regarded as black-box systems, as the model lacks a human-understandable explanation of its reasoning due to the complexity of the underlying functional models. To address this limitation, extensive research on Explainable AI (XAI) [25,26] has been conducted in recent years, aiming to enhance models with human-understandable explanations of their reasoning. Two of the most widely used explainable ML approaches are Local Interpretable Model-agnostic Explanations (LIME) [27] and SHapley Additive exPlanations (SHAP) [28]. They provide post-hoc explanations for model predictions, enabling an understanding of how each attribute contributes to the overall outcome. In this work, SHAP is used to generate model explanations, addressing the AI system’s explainability requirements. Furthermore, the EU AI Act incorporates fairness requirements to mitigate bias in AI systems, ensuring that AI-driven decisions do not reinforce discrimination or inequality. The Handbook on European non-discrimination law [29] has defined a set of protected attributes that cannot be the basis for inferior treatment, such as sex or gender identity among others. Bias in ML models can originate from the training data or the algorithms themselves. Various techniques for the detection and mitigation of bias have been defined, ensuring fair and equitable AI outcomes [30]. However, not all differences in model predictions indicate discrimination. In certain domains, such as healthcare, variations in AI-driven decisions may be justified by biological differences or other legitimate factors rather than unfair bias [31]. Therefore, whether a model’s decision constitutes discrimination depends on the context and nature of the evaluated decision, rather than solely on the presence of different outcomes. Despite this nuanced distinction on bias, evaluating the system for gender-neutral behavior remains essential in medical applications. In this study, gender-neutral behavior within the system is analyzed by employing disparate impact analysis [32], a method used to identify and evaluate potential biases present in ML models.

This article showcases the outcomes of the design and development of a prototype for a personalized medicine platform as an AI system. It highlights the technical milestones achieved during the development, alongside addressing ethical and safety considerations and identifying limitations encountered in this preliminary phase. The remainder of the article is structured as follows. Section 2 describes the data under study, while section 3 describes the methods and section 4 presents the results of the predictive models, the software development and the study of ethical concerns of the system. The article concludes in section 5 with a discussion of the results and future lines of work.

2. Materials

The data under analysis are part of the PERMEPSY MCT database [33] that includes harmonized data from 22 international retrospective studies with information on the evolution of patients who received individual MCT. This database integrates the records of 698 patients and 563 attributes comprising sociodemographic information, such as age, gender, diagnosis, marital status, employment, and living situation, as well as the psychological evaluation of the patients before starting treatment (pre-evaluation) and after finishing treatment (post-evaluation), assessed under a range of 12 psychological indicators.

The analysis of the MCT-related retrospective data for the purposes of the PERMEPSY project has received approval from the respective ethical committees of the involved entities, namely the Research Ethics Committee of *Fundació San Joan de Déu* on 27/04/2023 by approval PIC-68-23, the *Lokale Psychologische Ethikkommission am Zentrum für Pschosoziale Medizin* of UKE on 29/03/2023 by approval LPEK-0603, the *Comitè Ético Científico del Servicio de Salud Valparaíso San Antonio* by approval

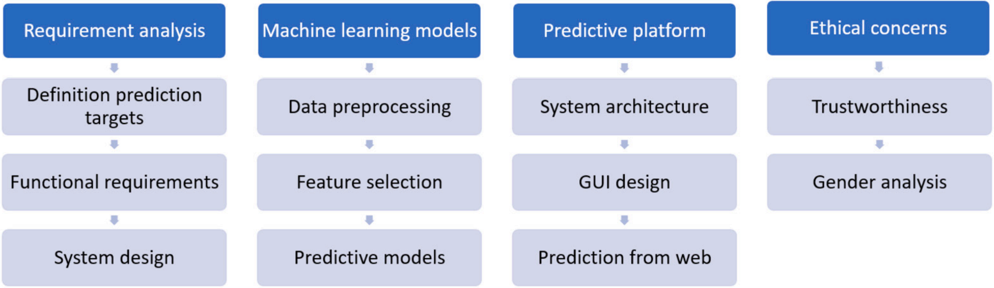


Fig. 1. Overview of development tasks.

N°54/2023 on 27/09/2023, and the Ethics Committee of the UPC on 11/12/2023 by approval 2023.13.

The database contains a substantial number of missing values, a consequence of the integration of data collected across 22 independent retrospective MCT-related studies. Efforts to address data integration as part of data harmonization have been outlined in [33]. Among the symptoms addressed by MCT, those evaluated through the PANSS indicator represented one of the most frequently available variables in the database and were previously analyzed in detail in [34]. To assess the homogeneity of features across the 22 data sources, analyses of the data distribution were carried out. The Chi-square test was applied to categorical variables, and the Kruskal–Wallis test to numerical ones, to determine whether the MCT data originated from a common distribution [35]. These statistical tests provide evidence of consistency and variability across sources. Most variables yielded p-values below 0.05, reflecting significant heterogeneity and suggesting meaningful differences in distributions across studies. These findings highlight the structural complexity of integrating data from heterogeneous studies conducted under differing protocols, time frames, and contextual conditions. Such circumstances can hinder harmonization and downstream analyses [36]. Details regarding missing values and the statistical assessments of homogeneity are presented in Appendices F and G, respectively. Despite the heterogeneity of the data, preparing it for further analysis using predictive models requires addressing missing values through multivariate imputation techniques. In this study, Multiple Imputation by Chained Equations (MICE) was applied to enhance the completeness of the dataset [37].

3. Methods

The development of the web-based predictive platform within the scope of the PERMEPSY project integrates data science and software development as described in the overview of tasks depicted in Fig. 1. The initial phase focuses on an analysis of requirements, enabling the system's definition by establishing key metrics to assess MCT effectiveness and determining the functional requirements of the predictive platform, which are essential for the subsequent software development. The second phase of the project focuses on the development of predictive models by applying ML-based analytical methods. The software development of the web-based predictive platform incorporates these predictive models, facilitating seamless accessibility and intuitive usage for end users via the internet. Finally, the project addresses ethical concerns associated with the use of AI technology in medical applications, focusing on trustworthiness, fairness, and transparency to ensure responsible and equitable use.

3.1. Analysis of requirements

The development of the prototype begins with a thorough analysis of the system requirements to lay the foundations of a well-structured design of the system. The purpose of the system is the prediction of MCT effectiveness on a personalized basis from the sociodemographic and mental health information of a patient. The system should provide a smooth and friendly experience for the end-user, ensuring ease of access and operation. The analytical process involves defining the prediction targets for the development of the predictive models, alongside analyzing the functional requirements for an efficient software design of the predictive platform.

3.1.1. Definition of prediction targets

The prediction of the effectiveness of MCT treatment focuses on the evaluation of the patient's health state after having completed the MCT treatment (post-evaluation) and taking into account the patient's health state before starting the treatment (pre-evaluation), as well as their sociodemographic information. This information is the baseline for the prediction (see Fig. 2 for an illustration of the main system process). Several well-known psychological indicators [33] are selected as key metrics for the evaluation of the patient's mental health state and constitute the prediction targets of the predictive models, as explained in the following:

1. The *PANSS* positive score (*PANSS_P*), as a key metric to assess schizophrenia-related positive symptoms [38].
2. The *Psychotic Symptom Rating Scale* (*PSY_T*), to evaluate the severity of positive symptoms, namely delusions and hallucinations [39] as the sum of the PSYRATS hallucinations (*PSY_H*) and the PSYRATS delusions (*PSY_D*) scores.
3. An alternative delusion score (*A_DEL*) derived from PANSS scores.
4. The *Rosenberg Self-Esteem Scale* (*RSES*) [40], as a metric of self-esteem.
5. The *Beck Cognitive Insight Scale* (*BCIS*) [41], as a metric of cognitive insights relying on the BCIS self-reflectiveness (*BCIS_R*) and the BCIS self-certainty (*BCIS_C*) subscores.
6. MCT completion (*MCT_C*), as a binary variable to measure whether the patient is likely to complete MCT treatment or not.

For details about the mathematical definition of each prediction target, as well as the method to calculate its ratio of change (which takes into account the difference between pre-evaluation (*PRE*) and post-evaluation (*POST*) with regard to the total range of values of the indicator), the reader is referred to Appendix A in the supplementary material. The ratio of change Δ is a derived metric that allows a straightforward interpretation of trends, expressed as a percentage. Positive values indicate the patient's improvement, while negative values reflect deterioration.

Overall, seven numeric target variables related to psychological indicators — *PANSS_P*, *PSY_D*, *PSY_H*, *A_DEL*, *RSES*, *BCIS_R*, *BCIS_C* — along with the binary variable *MCT_C*, are thus established as prediction targets. The selected prediction targets reflect symptom related and metacognitive domains aligned with the therapeutic principles of MCT. Specifically, *A_DEL* reflects the degree of conviction in delusional beliefs, while *BCIS_R* (self-reflectiveness) and *BCIS_C* (self-certainty) assess cognitive insight. Impaired cognitive insight, such as low self-reflectiveness and high self-certainty, indicates reduced metacognitive flexibility, which has been associated with poor response to psychotherapeutic interventions [41]. In the context of MCT, such insight profiles

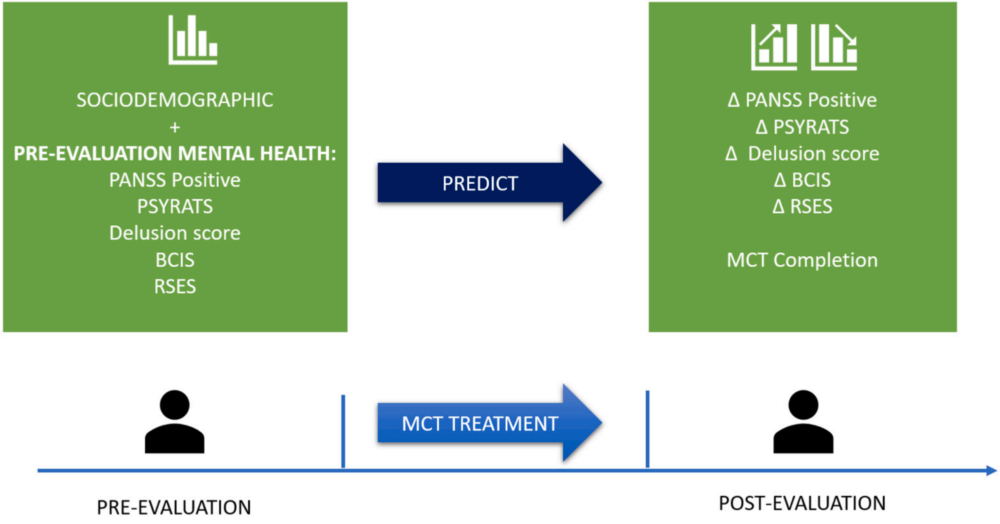


Fig. 2. Overview of the system process.

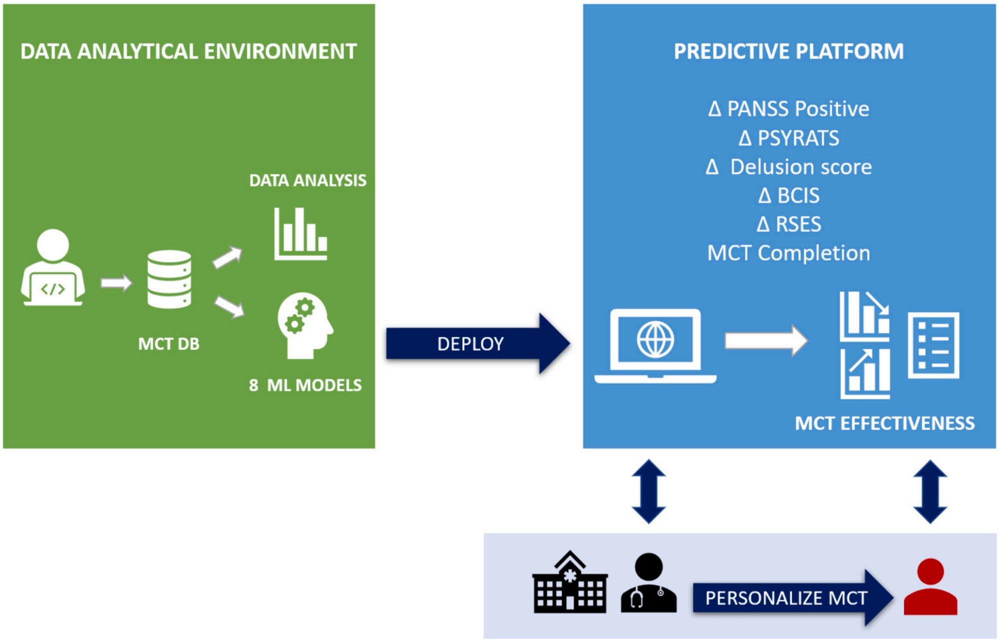


Fig. 3. Overview of systems and stakeholders.

are particularly relevant for tailoring intervention modules that target rigid thinking styles and promote belief reconsideration.

3.1.2. System design

According to the requirement analysis, the main system functionalities are defined as: I) Facilitate a system for entering patient profile information, including sociodemographic data and mental health assessments during pre-evaluation. II) Perform inference using eight ML-based models corresponding to specific prediction targets. III) Present prediction results both numerically and graphically. IV) Provide explanations for the prediction outcomes based on the baseline features.

Regarding non-functional requirements, the system must ensure ease of use and operation, supporting timely predictions from eight ML models. Additionally, it should address key principles such as data privacy, fairness, and transparency, in compliance with the EU AI Act 2024.

Fig. 3 illustrates the interaction between the various stakeholders of the system and the corresponding elements produced by each subsystem: I) A data analytical environment for the development of predictive models using Python and scikit-learn libraries [42]. This environment ensures secure and protected analysis of the MCT database, with the predictive models subsequently exported and deployed to the web-based platform for seamless integration. II) A web-based software system to ensure seamless access to the predictive platform over the Internet for end users. The Django framework [43] was chosen as the technological solution for the efficient integration of Python-based ML models within a web server framework. III) The end-users of the predictive platform are clinicians who access it remotely to gather insights into the effectiveness of MCT for individual patients. By entering baseline information, clinicians receive predictive data regarding treatment effectiveness, which serves as valuable support for tailoring and personalizing MCT treatment to the needs of each patient.

3.2. Development of ML models

In addition to the binary variable related to MCT completion, seven numeric variables were defined as key metrics for psychological symptoms, serving as prediction targets for the system. The prediction of numeric variables is approached as a regression task, while the prediction of the MCT completion variable is treated as a binary classification problem. Several ML models were evaluated, and their hyperparameters fine-tuned to optimize performance and ensure the most accurate predictions.

Regression models were evaluated calculating the prediction error measured by the root mean square error (RMSE). For binary classification, the model's performance was assessed using accuracy, specificity, sensitivity and the Matthews correlation coefficient (MCC) [44]. While the specificity and sensitivity refer to the recall of the negative and positive classes, respectively, the MCC is a balanced metric that takes into account both classes and yields values in range 1 to -1, where 1 stands for perfect classification and -1 for complete misclassification. The MCC metric has been deemed to be appropriate for classification under class imbalance [45].

Model selection follows a two-step process. Initially, a broader feature set undergoes feature selection using measurements of feature importance derived from a Random Forest (RF) model. Based on this analysis, the most relevant features for each model are identified, and a refined subset is selected as the input feature space to construct the definitive predictive models, as elaborated in subsequent sections.

3.2.1. Feature selection

The input variables for the predictive models include data related to the patient's sociodemographic profile and a variety of psychological indicators from pre-evaluation. The MCT database consists of 13 sociodemographic variables and 12 principal psychological indicators, collectively comprising 274 variables. Feature importance analysis identifies the most relevant features in the prediction of each target variable. RF [46] models are employed for this purpose, as their intrinsic capability to assess feature importance through permutation-based modeling makes them a reliable choice for this type of analysis [47].

3.2.2. Predictive models

After selecting a subset of relevant variables for the prediction models, various ML models were constructed using the reduced feature set. For numerical target variables, baseline models like Linear Regressors (LR) and Decision Tree Regressors (DT) are considered alongside more advanced ensemble models, such as RF, XGBoost (XGB), and LightGBM (LGBM) Regressors [48], an efficient variant of Gradient Boosting machines [49]. Ensemble models use a set of *weak* classifiers to construct a stronger model. RF employs a *bagging* strategy to aggregate multiple weak classifiers, while LGBM and XGB utilize a boosting technique, sequentially adapting weak classifiers to construct a more robust predictive model [50].

For the binary classification of the *MCT_C* variable, classification variants of the previous algorithms were employed, with Logistic Regression (LoR) [51] serving as the baseline model. Class imbalance, arising from the high ratio of the positive class (MCT completion) to the negative class (No MCT completion), at 667:31, is mitigated through SMOTE-based oversampling [52] applied to 80% of the training data. This approach ensures balanced representation of both classes during the model training process.

The experimental setup for model training involves a hyperparameter tuning process. This includes evaluating combinations of several relevant hyperparameters through a randomized search comprising 20 iterations and 5-fold cross-validation. The optimal hyperparameters identified for each model are detailed in Appendix E of the supplementary material.

The definitive prediction models were trained using the optimal hyperparameters identified for each model type. The data was split randomly, with 80% allocated for training and 20% for validation. This procedure is repeated 30 times to ensure reliable statistics of model performance.

3.2.3. Model explanations

SHAP analysis [53], a method derived from Shapley values in game theory [54], was used to create post-hoc explanation models. SHAP builds a surrogate model for the predictions of the original black-box ML model. The surrogate model aims to assess the sensitivity of each feature in the prediction of the model by representing it as Shapley additive values. The surrogate model breaks down the final prediction into feature-specific contributions, serving as a post-hoc explanation of the model's reasoning for a given prediction. In particular, the Tree SHAP algorithm was used to explain the output of ensemble models [55].

3.3. Fairness

ML models are required to be fair on sensitive attributes to prevent disparate impact for protected demographic groups [23]. The disparate impact doctrine is one of the most predominant legal theories used to evaluate unintended bias in systems [32]. While not established as a legal requirement, the generalization of the 80 percent rule, advocated by the US Equal Employment Opportunity Commission (EEOC) [56], is widely adopted to define the acceptable amount of bias. This rule establishes that the maximum allowable disparity in selection rates between protected and unprotected groups should be at least 80%, ensuring that AI-driven decisions do not disproportionately disadvantage certain demographic groups.

As described in [57], the Disparate Impact ("80% rule") establishes that the acceptable amount of bias ϵ for a particular metric is the impact ratio (IR) between the selection ratio (SR) of the unprivileged group b and the privileged group a. A common choice for ϵ is 0.8, which corresponds to the 80% rule [56]. The acceptable amount of bias is therefore an IR within the range $[\epsilon; \frac{1}{\epsilon}]$.

$$SR = \frac{Favorable(gender)}{Total(gender)} \tag{1}$$

$$IR = \frac{Female\_SR}{Male\_SR} \tag{2}$$

For the binary MCT completion model, the evaluation of the IR corresponds to the relation between the selection ratio of females and males (See formula (1) and (2)). Linking the decision on admitting a patient to therapy to the criteria of MCT completion, the predicted likelihood of a patient completing the treatment, makes it a potentially sensitive classification, subject to fairness considerations. The remaining models assess the patient's responsiveness to MCT effectiveness formulated as regression models. To evaluate disparate impact on the regression model, a paired t-test on the RMSE and predictions between gender groups is carried out to determine whether there are significant differences in the model's prediction [58,59]. While the t-test on the RMSE evaluates whether the model is equally accurate by gender, the t-test on the predictions evaluates if there are significant differences in effectiveness by gender, which might be explainable by documented knowledge regarding gender differences in MCT effectiveness [18].

4. Results

4.1. Feature selection

The analysis of feature relevance involves two steps. First, models are selected. Then, feature relevance is assessed for the chosen model. Two explainable models were trained: Linear Regression (LR) and RF regression for numeric target variables, and LoR and a RF classifier for the MCT_C variable. As shown in Fig. 4, the RF model outperforms other models in predicting RSES. Therefore, it was selected for feature relevance analysis. See Appendix B in the supplementary material for a detailed description of the 15 most relevant features for each target, identified through variable permutation on the RF model. Interestingly, sociodemographic variables such as age, gender, and substance consumption are often found to be highly predictive. Furthermore, the

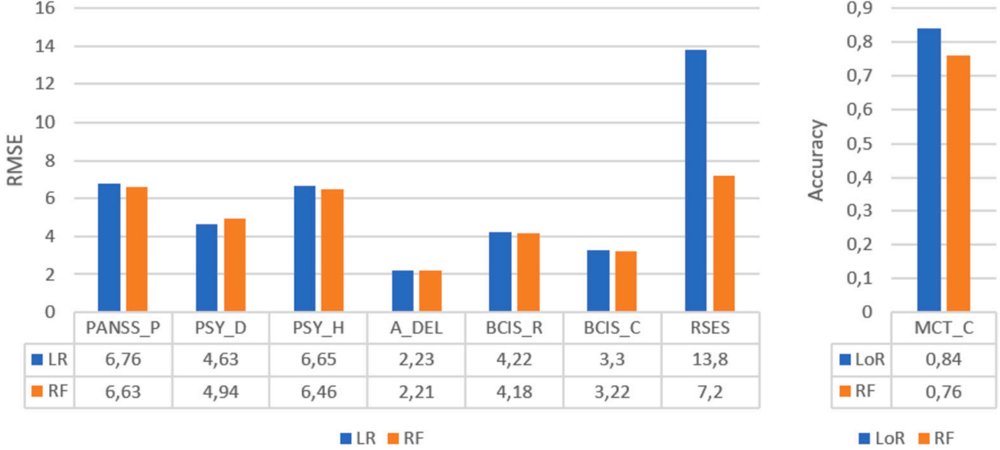


Fig. 4. Performance of LR and RF models for the target variables for feature relevance analysis measured as RMSE for regression (left) and accuracy for classification (right).

Table 1
Performance of different types of models in the prediction of the target variables, measured by RMSE.

Target	DT	LR	RF	XGB	LGBM
PANSS_P	9.77 ± 0.26	6.75 ± 0.16	**6.64 ± 0.07**	7.20 ± 0.16	6.95 ± 0.04
PSY_D	7.31 ± 0.01	5.05 ± 0.18	**4.61 ± 0.08**	5.16 ± 0.29	4.84 ± 0.32
PSY_H	8.58 ± 0.06	6.33 ± 0.81	6.75 ± 0.07	6.47 ± 0.63	**6.30 ± 0.72**
A_DEL	3.68 ± 0.06	2.70 ± 0.02	**2.52 ± 0.01**	2.70 ± 0.14	2.60 ± 0.12
BCIS_R	5.10 ± 0.07	**3.47 ± 0.03**	3.58 ± 0.01	3.79 ± 0.23	3.77 ± 0.13
BCIS_C	4.48 ± 0.02	2.93 ± 0.09	**2.74 ± 0.01**	3.15 ± 0.24	3.12 ± 0.15
RSES	7.70 ± 0.07	13.75 ± 2.9	**6.06 ± 0.05**	11.8 ± 5.21	12.46 ± 3.8

pre-evaluation mental health questionnaires often emerge as the most significant predictors. For MCT completion, variables like years of education, length of illness, and specific questionnaires (e.g., BCIS, TMT, PANSS) play an important role as well.

4.2. Definition of baseline variables

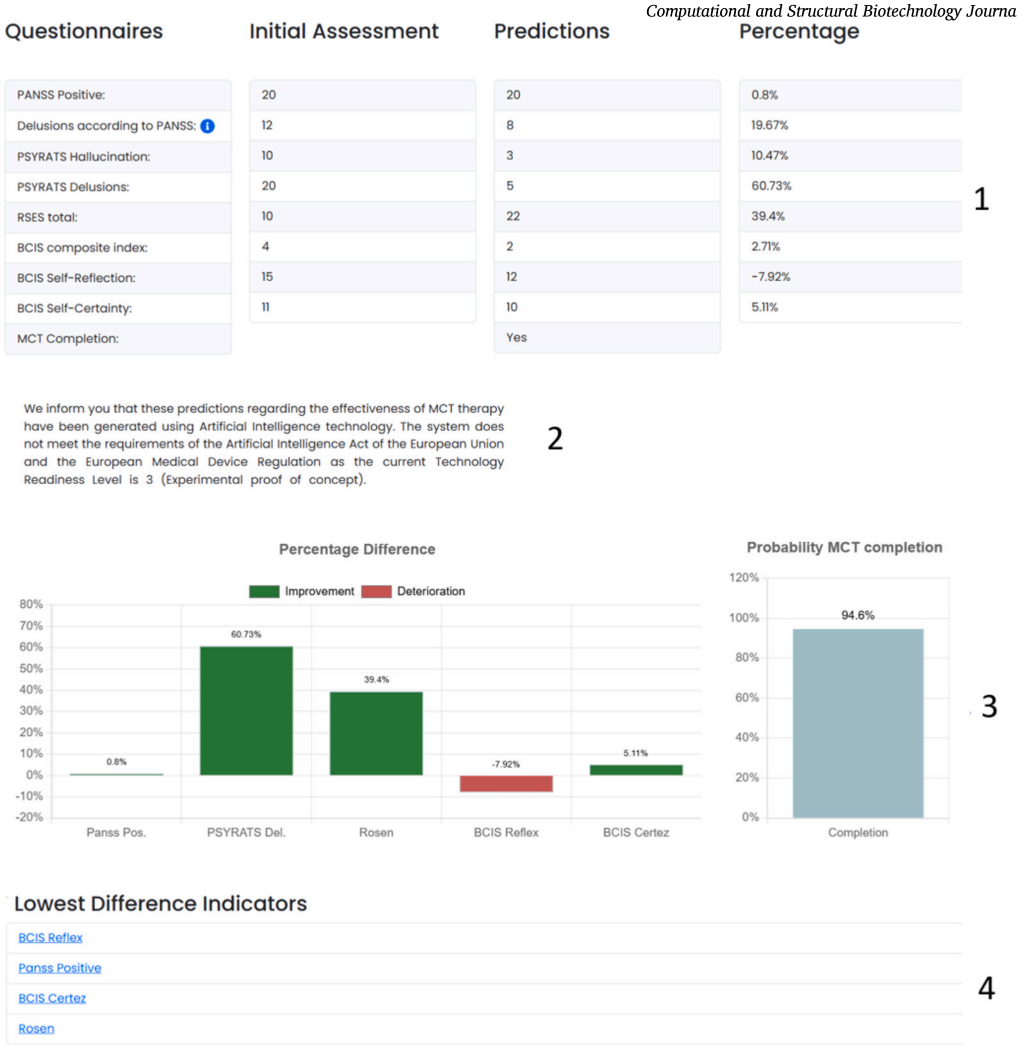
Based on the information on the relevant variables of the eight predictive models and the expert knowledge from the clinical partners in the PERMEPSY project, a common set of characteristics was selected. The selected features include 11 sociodemographic variables, 2 diagnostic-related variables (diagnosis and length of illness), and 8 psychological questionnaires from the MCT database [33]. These baseline variables are the foundation for training predictive models and are used as inputs for inference within the system. Several clinically relevant indicators derived from standardized assessments help define the cognitive, emotional, and functional profiles that a relevant for MCT engagement. These include subcomponents of the PANSS scale (positive, negative, and general scores) for symptom severity; BCIS scores (reflectiveness and certainty) for cognitive insight; the Rosenberg Self-Esteem Scale; Trail Making Test (TMT A and B) for executive functioning; the Global Assessment of Functioning (GAF) score; PSYRATS subscales for hallucinations and delusions; a Quality of Life (QoL) measure; and the Jumping to Conclusions (JTC) bias for reasoning style. The set of baseline variables also includes 11 sociodemographic and behavioral indicators that characterize patient profiles prior to MCT. These consist of demographic factors such as gender, age, education level, marital status, living situation, and employment status as well as lifestyle-related variables such as caffeine consumption, tobacco use, alcohol use, cannabis use, and illicit substance use. Together, these features provide information about the context for interpreting psychological assessments and may influence responsiveness to MCT as explained in the following. Psychological features such as *A_DEL*, *BCIS_R*, *BCIS_C*, *PANSS_P*, *PSY_D*, *PSY_H*, *RSES*, and *TMT* align with core mechanisms targeted by MCT, including belief inflexibility, reasoning biases, and emotional vulnerability. These variables reflect central aspects of psychosis such as delusional conviction, cognitive insight deficits, symptom severity, self-esteem, and executive dysfunction that all influence how patients engage with metacognitive content. Theoretically, individuals with more pronounced difficulties may benefit more from MCT modules tailored to those specific impairments, although such challenges may also pose obstacles to maintaining treatment adherence and completing the intervention. Sociodemographic variables such as age, gender, education, and living situation were selected based on availability and their potential impact on treatment access, engagement, and cognitive capabilities. For a detailed description of these features, the reader is referred to Appendix C in the supplementary material.

4.3. Predictive models

Several predictive models were trained from the baseline set of features. Information about the best-performing parameters found by hyperparameter tuning is detailed in Appendix E (supplementary material).

Table 1 shows the prediction error measured as RMSE of the different types of models for each numerical target variable. The findings indicate that ensemble models such as RF, XGB and LGBM show similar performance in almost all prediction targets compared to the LR model. Nevertheless, the RF model clearly outperforms other models for the RSES target. This result underscores the RF model's effectiveness in certain contexts, particularly for RSES prediction.

The RMSE values presented in Table 1 represent the prediction error in the original scale of each indicator. For instance, the PANSS_P model shows a prediction error of 6.6 PANSS score points on the PANSS positive scale. To facilitate a qualitative interpretation of these RMSE values, predictions are converted into percentage errors. These percentages are calculated based on the range of values for each indicator, as illustrated in Table 2. The prediction error of 6.6 points for the PANSS_P



Questionnaires	Initial Assessment	Predictions	Percentage
PANSS Positive:	20	20	0.8%
Delusions according to PANSS: ⓘ	12	8	19.67%
PSYRATS Hallucination:	10	3	10.47%
PSYRATS Delusions:	20	5	60.73%
RSES total:	10	22	39.4%
BCIS composite index:	4	2	2.71%
BCIS Self-Reflection:	15	12	-7.92%
BCIS Self-Certainty:	11	10	5.11%
MCT Completion:		Yes	

We inform you that these predictions regarding the effectiveness of MCT therapy have been generated using Artificial Intelligence technology. The system does not meet the requirements of the Artificial Intelligence Act of the European Union and the European Medical Device Regulation as the current Technology Readiness Level is 3 (Experimental proof of concept).

Lowest Difference Indicators

- BCIS Reflex
- Panss Positive
- BCIS Certez
- Rosen

Fig. 5. GUI of the prediction result page. The upper part (1) shows the predictions in a tabular format. Part 2 (middle) displays the AI ACT related disclaimer. Part 3 is a graphical representation of the ratio of change. The bottom part (4) highlights the indicators with the lowest improvement.

Table 2
Equivalence of RMSE and percentage error according to the range of values of the indicators.

	PANSS_P	PSY_D	PSY_H	A_DEL	BCIS_R	BCIS_C	RSES
RMSE	6.64	4.61	6.3	2.52	3.47	2.74	6.1
%	15.8	19.2	14.4	12	9.6	11.4	20.3

score corresponds to a percentage error of 15.8% in the indicators scale. These results indicate that the predictive models show a percentage error ranging between 10% and 20%, approximately.

Table 3 presents the classification performance of various models for the binary MCT completion variable. Among them, LoR stands out, achieving a recall of 0.76 for the negative class (No MCT completion), outperforming other models. Other models struggle to recognize this class accurately. For the positive class (MCT completion), LoR achieves a notably high recall of 0.95. A MCC value of 0.73 further confirms the strong classification performance. These results underscore LoR's ability to effectively predict both positive and negative classes for the MCT completion variable.

To facilitate the integration of AI-based technology into the prototype of the web-based predictive platform, RF models were uniformly selected for all psychological indicators. Their performance was deemed sufficiently robust across all target variables. For MCT completion, LoR was chosen for deployment on the platform.

4.4. Web-based predictive platform

The user-friendly web-based platform, built with Python and the Django framework, allows users to interactively engage with predictive models. This functionality enables the analysis of MCT effectiveness based on individual patient profiles. The platform includes a form view designed for users to introduce the patient's baseline information, i.e., the features defined for prediction in Appendix C of the supplementary material. The form ensures that all necessary data are collected accurately, serving as the foundation for generating predictions. The system displays a prediction result page, detailing the outcomes generated by the eight predictive models, as illustrated in Fig. 5. The prediction result page includes the following key sections:

1. **Predicted Values**: Presents predictions for psychological indicators and the percentage change compared to pre-evaluation values. For the binary MCT completion variable, predictions are displayed as 'Yes' or 'No', along with the model's confidence level for each prediction.
2. **Legal Disclaimer**: Provides compliance information related to AI technology regulations, ensuring transparency as required by relevant laws.
3. **Graphical Representation**: Visualizes the rate of change for psychological indicators and confidence levels for predictions. This al-

Table 3
Performance of different types of models for the prediction of the MCT completion.

Metric	DT	LoR	RF	XGB	LGBM
Accuracy	0.67 ± 0.08	**0.85 ± 0.05**	0.60 ± 0.026	0.69 ± 0.04	0.65 ± 0.03
Specificity	0.41 ± 0.16	**0.76 ± 0.12**	0.21 ± 0.05	0.41 ± 0.1	0.32 ± 0.05
Sensitivity	0.93 ± 0.01	**0.95 ± 0.02**	0.99 ± 0.00	0.98 ± 0.01	0.98 ± 0.01
MCC	0.4 ± 0.13	**0.73 ± 0.11**	0.32 ± 0.05	0.48 ± 0.06	0.39 ± 0.06

lows clinicians to quickly assess improvements or deterioration in patient profiles.

4. **Indicators with Lowest Improvement**: Highlights psychological indicators showing the least progress, helping clinicians to tailor MCT treatments for better patient outcomes.

These sections collectively ensure that clinicians can effectively interpret patient outcomes and make informed decisions about treatment adjustments. End users can generate and download a prediction report, as detailed in Appendix D of the supplementary material. This report includes the patient's baseline data, the model predictions, and the post-hoc explanations of the predictions based on SHAP explainability models. These features enhance transparency and assist clinicians in their interpretation of the predictions effectively for better-informed decision-making.

4.5. *Ethical concerns*

Several ethical concerns related to the use of AI technology in medical applications need to be addressed, including:

- **Transparency and Explainability**: Ensuring that AI models provide clear and interpretable predictions so clinicians can understand the rationale behind a decision.
- **Data Privacy and Security**: Protecting patients' data to comply with privacy laws and preventing unauthorized access or misuse.
- **Bias and Fairness**: Identifying and mitigating biases in datasets or algorithms that may lead to unfair treatment of certain patient groups. In the present work, the bias analysis focuses on gender. However, a comprehensive analysis would be required to examine all variables thoroughly.
- **Regulatory Compliance**: Adhering to standards like the EU AI Act or similar regulations to ensure ethical deployment and validation of medical AI tools.
- **Clinical Applicability**: Integrating AI-based tools in clinical applications poses challenges for all stakeholders, especially for clinicians and patients.

4.5.1. *Transparency and explainability*

The system provides human oversight of the system and transparency about the logic behind predictive models, as required by Art.14 of the EU AI ACT, through post-hoc explanations. These explanations are generated using a surrogate SHAP model, which explains how the final prediction is calculated from the individual feature's contributions in the prediction.

Fig. 6 provides an example of the explanation for the PSYRATS delusions prediction, displayed as a waterfall plot of SHAP values for each feature. In the plot, positive contributions to the predicted value are shown in red, while negative contributions are shown in blue. The waterfall plot highlights the 15 highest contributions based on SHAP values.

The predicted value is calculated as the cumulative sum of SHAP values, starting from the base predicted value of the SHAP model. In the example shown in Fig. 6, the base predicted value for *PSY_D* is 4.57. The total net contribution of all features is −1.2, resulting in a final predicted value of 3.37. While the PSYRAYS delusions variable in pre-evaluation (value: 13), the BCIS certainty score (value: 18), and an education level of 11 years have a positive contribution to the predicted value, the PSYRATS hallucinations score (value: 0) has a large negative contribution.

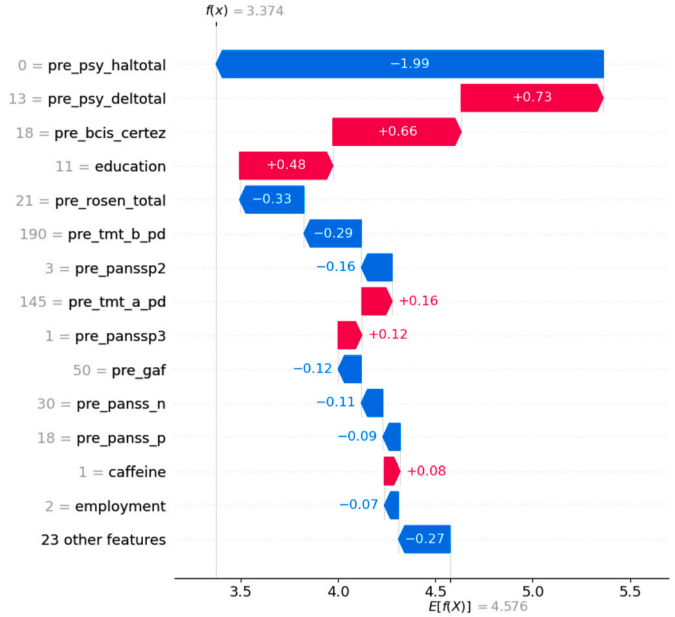


Fig. 6. Waterfall plot of the SHAP values to explain the prediction of the PSYRATS delusions target.

4.5.2. *Data privacy and security*

The MCT database contains patient information, which is safeguarded to ensure privacy and prevent unauthorized access or misuse. Although the data are fully anonymized and, therefore, not subject to the General Data Protection Regulation (GDPR), the PERMEPSY consortium has developed a comprehensive data management plan to securely handle the information. Access to the data is strictly limited to a secure environment — the MyDisk platform of RDLab at UPC [60], which complies with various regulatory frameworks. These include GDPR, HIPAA, ISO/IEC 2700X standards, and Computer System Validation practices such as Good Automated Manufacturing Practice (GAMP5). Additionally, the predictive platform, hosted at the RDLab facilities as well, is available as part of the web server, which relies solely on ML-based models and does not use the patients' records. Therefore this approach effectively restricts data distribution and enhances privacy protection.

4.5.3. *Gender bias analysis*

The MCT database contains information from 273 male and 425 female patients. To assess gender fairness in the predictive models, their performance was analyzed through disparate impact analysis on the test dataset, which consisted of 20% of the total data (77 females and 63 males). The analysis aimed to ensure that the models provided equitable predictions across genders. The fairness analysis for the MCT completion target evaluates disparate impact by calculating the impact rate of favorable predictions by gender, following the 4/5th rule. Favorable predictions are classified as successful MCT completion. With a male SR of 0.87 and a female SR of 0.99, the calculated IR is 1.13. Since this ratio is inside the thresholds specified

Table 4
Analysis of disparate impact on predictive models. RMSE values according to gender (left) to evaluate disparate impact on the accuracy of predictions and the predicted values (right) to evaluate whether there are significant differences in the predictions. *p-value* provided for a t-test for statistical significance between male and female sample.

	Accuracy				Prediction			
Target	Male	Female	All	p-value	Male	Female	All	p-value
PANSS_P	6.32	6.98	6.63	0.23	15.1	15.6	15.4	0.52
PSY_D	4.17	5.23	4.68	0.07	4.5	5.1	4.8	0.47
PSY_H	6.14	7.46	6.77	0.40	2.6	3.2	2.9	0.45
A_DEL	2.53	2.48	2.5	0.89	7.3	7.3	7.3	0.99
BCIS_R	3.7	3.4	3.57	0.43	15.0	14.5	14.8	0.44
BCIS_C	2.83	2.68	2.77	0.76	10.9	10.6	10.8	0.59
RSES	6.87	5.26	6.2	0.10	25.3	22.4	24.0	**0.04**

by the 4/5th rule, the predictions of the MCT completion model are considered fair with respect to gender, as no disparate impact is observed.

For the regression models, disparate impact is evaluated by testing for significant differences in the accuracy of the models by evaluating the prediction error (RMSE) between genders. Table 4 at the left side presents the RMSE values for the male and female samples, along with the corresponding p-value from the t-test. The findings indicate no significant differences in the accuracy of the models by gender, as the p-value is always greater than 0.05. Table 4 shows, on the right side, the mean predicted values for each prediction target. Only for the RSES indicator, a significant difference in the prediction according to gender was found.

4.5.4. Regulatory compliance

The development of the predictive platform based on AI technology for medical applications requires adherence to several regulations and laws to ensure compliance, safety, and ethical use:

1. **General Data Protection Regulation (GDPR)**: Governs the processing and storage of personal data, guaranteeing privacy and security. It is especially important for managing sensitive patient information.
2. **EU AI Act**: A comprehensive regulation that categorizes AI systems by risk levels (e.g., high-risk systems such as medical AI). It mandates requirements for transparency, accountability, safety, human oversight, and bias mitigation.
3. **Medical Device Regulation (MDR)**: Applicable if the AI system is classified as a medical device. MDR enforces clinical evaluation, risk management protocols, and post-market surveillance requirements.
4. **Ethical Guidelines**: Published by the European Commission, these guidelines emphasize principles of transparency, fairness, and accountability, fostering trustworthy AI implementation.

4.5.5. Clinical applicability

The platform provides AI-generated predictions regarding treatment outcomes and engagement based on pre-treatment characteristics of the patient. Nevertheless the platform is not designed to automate clinical decisions. Instead, it functions as a supportive tool that enhances clinician insight during MCT personalization. All treatment choices remain under the clinician's authority. Therefore the platform does not initiate, modify, or discontinue MCT independently. Its role is purely informative, offering clinicians data-driven insights that may support personalized treatment planning. For instance, predicted low engagement or symptom specific benefit might prompt clinicians to emphasize particular MCT modules or explore complementary strategies, but these decisions remain entirely led by the clinician. The current disclaimer indicates that predictions are AI-generated and do not comply with clinical regulatory standards. To reinforce ethics, a visible disclaimer within the interface should also clarify that outputs are non-prescriptive and intended solely for professional interpretation. This not only safeguards clinical autonomy but also aligns the system with established decision-making frameworks in mental healthcare.

In addition to ensuring bias-free behavior of AI systems at the technical level, broader ethical challenges must be addressed to facilitate their integration into clinical practice. Regulatory and ethical standards require accountability, patient autonomy, and informed consent for AI based systems. Therefore a key consideration is transparent disclosure of AI use in the clinical application so that patients are informed when AI contributes to their diagnosis or treatment. The system's function should be described whether it is a supportive tool or operates autonomously. Furthermore, human oversight must be preserved, with a clear statement whether healthcare professionals will review or override AI generated recommendations. In addition, stakeholders should receive a clear and comprehensible explanation of the system's functionality, including how it analyzes data, predicts outcomes, and informs clinical decision-making. Clinical implementation must be supported by well-defined protocols, particularly concerning oversight and accountability. This aligns with the argument presented by [61] who emphasize that explainability and transparency are essential for fostering trust and ethical use of AI in healthcare.

Beyond regulatory and ethical compliance, several practical challenges remain regarding the integration of AI tools into healthcare applications. These include cost, accessibility, and the effort for collecting assessment data prior to the start of the therapy. As highlighted by [62], such requirements can hinder scalable and fair AI-use adoption, especially in low resource environments or in context with limited data interoperability. In addition, deeper systemic issues further complicate the adoption of AI in clinical care. These include clinician skepticism stemming from limited trust in AI tools or lack of training, constrained resources in health environments that hinder access to digital platforms as well as uneven adoption rates due to cultural, economic, and infrastructural disparities across regions. Together, these factors create important challenges that need to be overcome for an equitable and scalable implementation of AI in real-world clinical workflows.

Furthermore, potential clinical overreliance on AI generated suggestions must be addressed. Overreliance in clinical settings occurs when healthcare professionals rely excessively on AI generated outputs, allowing that algorithmic recommendations override their own medical evaluation. In such cases, clinicians may neglect to critically evaluate the system's suggestions resulting in a transfer of the responsibility of decision-making from the clinicians to the AI system. This effect has been identified as automation bias, a significant concern in sensitive clinical applications that may compromise accountability and care quality [63]. However, implementing intuitive and context-sensitive AI explanations have been highlighted as a solution to reduce this risk by providing deeper clinical insight and more responsible decision-making [64].

4.6. MCT personalization

4.6.1. Present functional design

From a functional perspective, the system aims to act as decision support tool for personalizing MCT treatment. Using patient-specific information, it provides insights into the therapy's effectiveness, offering predictions regarding improvements in a wide range of psychological symptoms. The predictive platform is designed to address two specific clinical challenges in the context of delivering MCT. First, it aims to reduce uncertainty around individual therapy response by estimating patient specific change in cognitive and symptomatic indicators. Second, it seeks to support clinicians in planning personalized interventions by identifying in which psychological domains they are most likely to improve. In addition, the system predicts the likelihood of therapy completion, which is useful for decisions about engagement strategies and resource allocation. Together, these predictive insights assist clinicians in tailoring MCT delivery according to the patient's profile and therapeutic goals without replacing their role in evaluating and personalizing the therapy.

The predictive system was designed by choosing psychological outcomes that mainly focus on symptoms and thinking patterns addressed in MCT. In particular, *A_DEL* measures overconfidence in beliefs; *BCIS_R* and *BCIS_C* reflect aspects of cognitive insight, such as self-monitoring and belief inflexibility, which are distortions addressed by MCT. Other indicators are *PANSS_P*, *PSY_D*, *PSY_H*, and *RSES*. These outcomes relate to positive symptoms, negative self-perceptions, hostile ways of interpreting others' behavior, and self-esteem. Each is connected to specific MCT modules that address stigma and social thinking. *MCT_C*, while not a psychological distortion measure, informs predictions about treatment engagement, which may depend on cognitive flexibility.

To operationalize these insights, each of the prediction targets is estimated using a dedicated ML model trained on baseline psychological and sociodemographic variables. This includes predictions for the individual's change in seven metacognitive and symptomatic indicators, as well as the likelihood of completing the 8–10 week MCT program. During the model design relevant baseline variables for the prediction were selected from the MCT database. Standard ML techniques were used to select features that improve model performance. At the same time variables were kept that are considered clinically relevant to MCT. Then the ML models were trained using a comprehensive set of baseline features comprising 8 psychological indicators, 11 sociodemographic and behavioral variables, and 2 diagnosis related variables. Psychological indicators were selected for their relevance to MCT treatment targets, including belief flexibility, emotional vulnerability, and reasoning patterns. These include cognitive insight scores from the Beck Cognitive Insight Scale (BCIS_R for self-reflectiveness and BCIS_C for self-certainty), delusional conviction (A_DEL), self-esteem (RSES), symptom severity captured by the positive subscale of the PANSS (PANSS_P), hostile attribution style (PSY_H), reasoning bias measured through the Jumping to Conclusions indicator (JTC), executive function via Trail Making Test scores (TMT A and B), overall quality of life (QoL), and Global Assessment of Functioning (GAF) scores, which provide a measure of overall psychological, social, and occupational functioning. In addition, sociodemographic and lifestyle features, namely gender, age, years of education, marital status, living situation, employment status, and behavioral indicators such as caffeine, tobacco, alcohol, cannabis, and illicit substance use were included to contextualize patient profiles. These variables may affect MCT engagement, symptom expression, and treatment response patterns. Finally, two diagnosis related variables, the formal psychiatric diagnosis and illness duration, were included to reflect clinical history and guide treatment planning. Together, these features form the baseline information for model inference, therefore enabling personalized predictions based on cognitive, emotional, functional, and contextual profiles prior to treatment. While these features may offer contextual relevance to treatment planning, their relationship to metacognitive processes remains less clearly defined in existing literature. We acknowledge this as a limitation and recommend that future iterations of the system incorporate more theory driven functional and contextual predictors to enhance clinical interpretability and support personalization at the module level.

4.6.2. Challenges and strategic improvements

From a clinical perspective, the current prototype offers valuable insights as a preliminary model, but remains still limited in scope. It focuses mainly on factors specific to the patient, like their symptoms, personal background, and psychological traits, which are important, but only part of the clinical reality and do not fully capture the complex clinical reality of psychotherapy.

Notably, important clinical predictors such as psychiatric comorbidities, treatment adherence, and social support described as relevant in shaping therapy outcomes and engagement [65], are not available from the MCT dataset. Similarly, while the MCT dataset includes basic sociodemographic indicators such as living situation and employment status, which may serve as proxies for autonomy and occupational functioning, they do not offer the granularity or standardization necessary for multidimensional functional assessment. Although GAF scores are available in the dataset and offer a clinician-rated global index of psychological, social, and occupational functioning [66], they do not capture broader functional domains such as interpersonal relationships, autonomy, and vocational engagement, which are recognized as key clinical endpoints in psychiatric rehabilitation [67]. These limitations could be addressed in future iterations of the system by incorporating standardized tools, such as comorbidity indexes and adherence scales for clinical predictors, as well as validated instruments like the WHO Disability Assessment Schedule (WHODAS) [68] and the Personal and Social Performance scale (PSP) [69], enabling more comprehensive personalization and long-term MCT planning.

The preliminary prototype also presents limitations in its ability to model group-based MCT dynamics, illness trajectories, and cognitive functioning. Group-level factors such as cohesion, group size, and facilitator style are known to influence engagement and outcomes and should be considered in future iterations. Currently, only data from individual MCT sessions were available in the dataset, and no group-level metrics could be modeled. To overcome this limitation, future versions of the system should collect and incorporate data from group-based MCT settings to better reflect real-world therapeutic formats and capture dynamics that shape intervention effectiveness. Additionally, while illness duration was used as a proxy for chronicity, schizophrenia presents diverse trajectories such as cyclical symptom patterns and gradual cognitive decline, which should be assessed using broader clinical and cognitive indicators [70]. Cognitive functioning, a critical factor for MCT engagement, is also only partially represented. While the dataset includes indicators such as Trail Making Test scores and Jumping to Conclusions bias, and provides a global functioning index through GAF scores, broader standardized assessments in memory, attention, and judgment domains are missing [71]. In summary, incorporating in future development structured measures of cognitive capacity and illness progression, together with group-level variables, would enhance the system's ability to reflect real-world therapeutic contexts and support more personalized clinical decision-making.

Moreover, psychotherapy success depends not only on patient characteristics but also on relational dynamics and therapist related factors, which are currently not represented in the model due to the absence of corresponding data in the MCT database. Therapist variables, such as clinical orientation, experience, and interpersonal style along with the quality of the therapeutic alliance, substantially influence treatment effectiveness [72,73]. To better capture these elements, future iterations should incorporate therapist profiles and monitor therapeutic relationships using standardized questionnaires, including the Working Alliance Inventory (WAI) [74] and Session Rating Scale (SRS) [75]. Additionally, matching patient-therapist pairs on compatibility criteria, such as cultural or linguistic background, may strengthen personalization and improve outcomes [76]. Technically, these enhancements require expanded data collection from both patients and therapists.

Furthermore to support clinical integration, future versions of the platform may evolve along several development lines that enhance its practical relevance and ease of implementation. First, a qualitative study exploring clinicians' perspectives including usability, clarity, and perceived impact on therapeutic decision-making would provide valuable insights for refining system design and interface [77]. Second, model validation should be extended to diverse clinical contexts such as inpatient and outpatient care, and across different illness stages including early-episode and chronic presentations, to assess generalizability and robustness [78]. Third, the development of structured resources, such as a user guide or decision-support recommendations could facilitate integration into routine care, helping clinicians interpret model outputs effectively. This would reinforce transparency, support clinical autonomy, and ensure that the system remains a supportive tool for informed decision-making rather than a prescriptive algorithm.

5. Discussion and conclusions

The PERMEPSY project aims to develop a predictive platform as a technical prototype in a proof of concept to personalize the delivery of MCT. The purpose of the system is the prediction of MCT effectiveness on a personalized basis from the sociodemographic and mental health information of a patient. By analyzing psychological baseline data, the platform helps clinicians identify patients most likely to benefit from MCT. While it does not automate personalization or MCT module selection, it empowers therapists to adapt the intervention, for example, by emphasizing modules focused on self-esteem, cognitive biases, or attribution styles. This approach strengthens precision psychiatry by guiding MCT delivery according to each patient's unique symptom profile and therapeutic goals. As such, the system enhances treatment engagement and clinical relevance without replacing professional judgment as the platform's function is a decision support tool for the clinician without prescriptive character.

From a technical standpoint, the project has successfully achieved its development objectives. The development of the web-based predictive platform has integrated data science and software engineering through a structured sequence of phases, from requirement analysis, definition of prediction targets, construction of ML models to the implementation of system functionalities. This process has resulted in a functional prototype that allows the user to input baseline patient data and receive predictions regarding post-treatment outcomes and the likelihood of completing MCT. The prototype platform has been finalized and delivered to the project's clinical partners, facilitating testing and evaluation in a controlled environment, such as a statistical post-hoc validation. This milestone marks a significant step toward validating the system's functionality and reliability.

Nonetheless, the technical readiness level of the development, estimated as TRL3, a status prior to validation in a controlled environment according to ISO 16290:2013 definition [79], does not yet meet all the requirements outlined in the respective legislative regulations, as detailed in Section 4.5.4. Further developments will be necessary to ensure full compliance with the requirements of the medical device regulation and that of a high-risk AI system as described in [24]. These requirements constitute a comprehensive and demanding framework, covering aspects such as accuracy, robustness, transparency, human oversight, risk mitigation, and bias prevention, among many others. For instance, the requirements for accuracy and robustness outlined in Article 15 of the AI Act regarding defence against data/model poisoning, adversarial examples, and model flaws are closely tied to data quality. As described in Article 10, AI systems must be trained, validated, and tested using representative, relevant, and error-free data, with clearly defined quality criteria for these datasets.

Nevertheless, certain requirements of the AI Act have already been addressed. One key example is the emphasis on transparency, which has been tackled through the development of explainability models utilizing SHAP analysis. These models provide insights into the contributions of individual features to predictions, enhancing the interpretability and accountability of the system. This demonstrates the platform's commitment to aligning with ethical guidelines and fostering trust in AI applications.

Furthermore, models were assessed for gender bias, and the results confirmed that the system has gender-neutral behavior. MCT completion has been identified as a potentially sensitive decision, as the model's predictions might influence a clinician's decision to admit or exclude a patient from MCT treatment. However, the disparate impact analysis indicates that the gender bias within the model falls within the acceptable range defined by the four-fifths rule, suggesting that the model does not exhibit discriminatory behavior based on gender. For the models assessing MCT effectiveness, the system evaluated prediction accuracy differences across genders and variations in the predicted values themselves. The bias analysis on accuracy confirmed that the models perform equally well for both genders, ensuring fairness in predictive performance. Regarding the predicted values of MCT effectiveness, the disparate impact analysis identified significant differences only in the RSES model's predictions. However, prior research has demonstrated gender-specific differences in responsiveness to MCT treatment [18], suggesting that the observed bias in the RSES model reflects biological or clinical differences rather than discriminatory behavior. To ensure equitable system behavior, future fairness and bias analyses should be expanded to include additional sensitive variables, such as age and disability, as well as potential proxy attributes, following the criteria outlined in [29]. A comprehensive evaluation across these dimensions is necessary to validate the platform's fairness.

During this study, certain limitations were identified, primarily related to data quality. The MCT database, derived from 22 independent studies, presents challenges due to a substantial number of missing values and data collection under varying protocols and possible contextual conditions. To address missing information, extensive data preprocessing and imputation techniques were applied to the source data in an effort to complete the available information and mitigate data inconsistencies. However, residual biases stemming from study-specific characteristics may still persist. Data quality was found to be limited by heterogeneity across the 22 contributing sources, as evidenced by the results of Chi-square and Kruskal–Wallis tests. Nearly all variables exhibited statistically distinct distributions, indicating that they do not originate from a common population.

Additionally, assessing whether the prediction error of the models falls within acceptable limits is key to deciding on their practical application in medical decision-making. As described in the study, the prediction error for measuring MCT effectiveness was around 10% - 20% in the respective scales for the best-performing models. It is widely recognized that meeting data quality standards presents significant challenges when working with real-world, multi-study datasets. As described in [36], the integration of data from methodologically and temporally heterogeneous sources can complicate model stability and limit interpretability across populations. In conclusion, high-quality data remains paramount for the accurate and reliable functioning of predictive models.

To enhance model generalizability and clinical relevance, future improvements should prioritize the collection of representative patient data under a unified MCT treatment protocol. Additionally, data collection should encompass a broader range of clinically meaningful aspects, such as therapeutic engagement, illness trajectory, participation in group therapy, co-occurring psychiatric conditions, and functional outcomes in autonomy and social relationships, as outlined in Section 4.6.2. Capturing these dimensions would enable the platform to better reflect real-world therapeutic processes and support a more individualized and context-aware approach to MCT personalization. Addressing these data-related limitations will be fundamental to improving model performance, ensuring consistent results, and meeting the requirements for robustness and reliability outlined in the AI Act for medical applications and the medical device regulation.

Furthermore, successful clinical integration of AI tools requires overcoming key challenges as outlined in Section 4.5.5, such as generalizability across diverse clinical contexts, cost-effectiveness, and implementation feasibility. As well it is important to safeguard clinician autonomy and prevent overreliance on automated outputs by ensuring transparency and interpretability of the AI generated predictions. To support clinically sound use, future development should include clinician feedback on usability, validation across varied settings and illness stages, and structured resources to guide decision-making. These efforts must be supported by comprehensive data collection before and during treatment to strengthen therapeutic relevance and real-world applicability.

CRediT authorship contribution statement

Caroline König: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pedro Copado:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alfredo Vellido:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Conceptualization. **Àngela Nebot:** Writing – review & editing, Funding acquisition, Conceptualization. **Cecilio Angulo:** Writing – review & editing, Funding acquisition, Conceptualization. **Maria Lamarca:** Writing – review & editing, Validation, Project administration, Formal analysis, Data curation, Conceptualization. **Vanessa Acuña:** Writing – review & editing, Conceptualization. **Fabrice Berna:** Writing – review & editing, Conceptualization. **Steffen Moritz:** Writing – review & editing, Conceptualization. **Łukasz Gawęda:** Writing – review & editing, Conceptualization. **Susana Ochoa:** Writing – review & editing, Funding acquisition, Conceptualization.

Funding

This work is part of the European ERAPERMED 2022-292 research project entitled 'Towards a Personalized Medicine Approach to Psychological Treatment of Psychosis' (PERMEPSY) funded by European Union - NextGenerationEU, supported as grant AC22/0053 by the Instituto de Salud Carlos III (ISCIII) in Spain. The PERMEPSY project was supported under the frame of ERA PerMed by: Instituto de Salud Carlos III (IS CIII), Spain; German Federal Ministry of Education and Research (BMBF), Germany; Agence Nationale de la Recherche (ANR), France; National Centre for Research and Development (NCBR), Poland; Agencia Nacional de Investigación y Desarrollo (ANID), Chile.

Declaration of competing interest

The authors declare that they have no conflicts of interest related to the research, authorship, and publication of this article.

Acknowledgements

The authors would like to acknowledge the contributions of all PERMEPSY members: Coordinator center: Belen Ramos Josemaria, Judith Usall, Regina Vila Badia, Raquel Lopez Carrilero, Trini Pelaez Martinez, Irene Birulés, Claudia Requejo; Partner 1 members Rabea Fischer, Merle Schlechte, Annika Schmueser, Antonia Meinhart, Jakob Scheunemann; Partner 2 members Adrien Goncalves; Partner 3 members Marytna Krezolek, Hania Gelner, Adrianna Aleksandrowicz, Justyna Piwinska; Partner 4 members Marc Larroda; Partner 5 members Alvaro Cavieres.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2025.07.051.

References

[1] Zhang S, Bamakan SMH, Qu Q, Li S. Learning for personalized medicine: a comprehensive review from a deep learning perspective. IEEE Rev Biomed Eng 2018;12:194–208.

[2] Egger J, Gsaxner C, Pepe A, Pomykala KL, Jonske F, Kurz M, et al. Medical deep learning—a systematic meta-review. Comput Methods Programs Biomed 2022;221:106874.

[3] Fusar-Poli P, Schultze-Lutter F. Predicting the onset of psychosis in patients at clinical high risk: practical guide to probabilistic prognostic reasoning. BMJ Ment Health 2016;19:10–5.

[4] Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018;14:91–118.

[5] Stumpp NE, Sauer-Zavala S. Evidence-based strategies for treatment personalization: a review. Cogn Behav Pract 2022;29:902–13.

[6] Fisher AJ. Toward a dynamic model of psychological assessment: implications for personalized care. J Consult Clin Psychol 2015;83:825.

[7] Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022;54.

[8] Mancuso A, Specchia M, Lovato E, Capizzi S, Cadeddu C, et al. Economic burden of schizophrenia: the European situation. A scientific literature review. Eur J Public Health 2014;24:cku166–29. https://doi.org/10.1093/eurpub/cku166.129.

[9] Pares-Badell O, Barbaglia G, Jerinic P, Gustavsson A, Salvador-Carulla L, et al. Cost of disorders of the brain in Spain. PLoS ONE 2014;9:e105471. https://doi.org/10.1371/journal.pone.0105471.

[10] Morrison AP, Turkington D, Pyle M, Spencer H, Brabban A, et al. Cognitive therapy for people with schizophrenia spectrum disorders not taking antipsychotic drugs: a single-blind randomised controlled trial. Lancet 2014;383:1395–403.

[11] Moritz S, Woodward TS, Balzan R. Is metacognitive training for psychosis effective? Expert Rev Neurother 2016;16:105–7.

[12] Moritz S, Menon M, Balzan R, Woodward TS. Metacognitive training for psychosis (MCT): past, present, and future. Eur Arch Psychiatry Clin Neurosci 2023;273:811–7.

[13] Schlechte M, Moritz S, Veckenstedt R, König C, Berna F, et al. Metakognitives training für psychose. Psychother 2025:1–6.

[14] Lamarca M, Espinosa V, Acuña V, Vila-Badia R, Balsells-Mejia S, et al. Reducing self-stigma in psychosis: a systematic review and meta-analysis of psychological interventions. Psychiatry Res 2024;342:116262.

[15] Meinhart A, Sauvé G, Schmueser A, Penney D, Berna F, et al. Metacognitive training for psychosis (MCT): a systematic meta-review of its effectiveness. Transl Psychiatry 2025;15:156.

[16] Acuña V, Otto A, Cavieres A, Villalobos H. Efficacy of metacognitive training in a Chilean sample of people with schizophrenia. Rev Colomb Psiquiatr (Engl ed) 2022;51:301–8.

[17] Acuña V, Cavieres Á, Arancibia M, Escobar C, Moritz S, et al. Assessing patient satisfaction with metacognitive training (mct) for psychosis: a systematic review of randomized clinical trials. Clin Psychol Psychother 2024;31:e3065.

[18] Salas-Sender M, López-Carrilero R, Barajas A, Lorente-Rovira E, Pousa E, et al. Gender differences in response to metacognitive training in people with first-episode psychosis. J Consult Clin Psychol 2020;88:516.

[19] Moritz S, Menon M, Andersen D, Woodward TS, Gallinat J. Moderators of symptomatic outcome in metacognitive training for psychosis (MCT). Who benefits and who does not? Cogn Ther Res 2018;42:80–91.

[20] Leanza L, Studerus E, Bozikas VP, Moritz S, Andreou C. Moderators of treatment efficacy in individualized metacognitive training for psychosis (MCT+). J Behav Ther Exp Psychiatry 2020;68:101547.

[21] González-Blanch C, Birulés I, Pousa E, Barrigon ML, López-Carrilero R, et al. Moderators of cognitive insight outcome in metacognitive training for first-episode psychosis. J Psychiatr Res 2021;141:104–10.

[22] Moritz S, Woodward TS. Metacognitive training in schizophrenia: from basic research to knowledge translation and intervention. Curr Opin Psychiatry 2007;20:619–25. https://doi.org/10.1097/YCO.0b013e3282f0b8ed.

[23] European Union. Artificial intelligence act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, 2024. Adopted by the 27 EU member states in March 2024.

[24] Busch F, Kather JN, Johner C, Moser M, Truhn D, et al. Navigating the European Union artificial intelligence act for healthcare. npj Digit Med 2024;7:210.

[25] Lisboa P, Saralajew S, Vellido A, Fernández-Domenech R, Villmann T. The coming of age of interpretable and explainable machine learning models. Neurocomputing 2023;535:25–39.

[26] Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res 2021;70:245–317.

[27] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135–44.

[28] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.

[29] European Union Agency for Fundamental Rights. Handbook on European non-discrimination law – 2018 edition. European Union Agency for Fundamental Rights. https://fra.europa.eu/en/publication/2018/handbook-european-non-discrimination-law-2018-edition, 2018.

[30] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54:1–35.

[31] Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digit Med 2020;3:81.

[32] Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining; 2015. p. 259–68.

[33] König C, Copado P, Lamarca M, Guendouz W, Fischer R, et al. Data harmonization for the analysis of personalized treatment of psychosis with metacognitive training. Sci Rep 2025;15:10159.

[34] König C, Guendouz W, Copado P, Angulo C, Nebot À, et al. Group discovery in a clinical database of patients with psychosis who have undergone metacognitive training. In: International meeting on computational intelligence methods for bioinformatics and biostatistics. Springer; 2024. p. 269–80.

[35] Kim J, Kim DH, Kwak SG. Comprehensive guidelines for appropriate statistical analysis methods in research. Korean J Anesthesiol 2024;77:503–17.

[36] Chow S-M, Nahum-Shani I, Baker JT, Spruijt-Metz D, Allen NB, et al. The ILHBN: challenges, opportunities, and solutions from harmonizing data under heterogeneous study designs, target populations, and measurement protocols. Transl Behav Med 2023;13:7–16.

[37] Maale FD, Okyere GA, Awe OO. Effects of imputation techniques on predictive performance of supervised machine learning algorithms. In: Practical statistical learning and data science methods: case studies from LISA 2020 global network. USA: Springer; 2024. p. 29–48.

[38] Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987;13:261–76.

[39] Haddock G, McCarron J, Tarrier N, Faragher E. Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (PSYRATS). Psychol Med 1999;29:879–89.

[40] Rosenberg M. Rosenberg self-esteem scale (RSE): acceptance and commitment therapy. Measures package, 61. Society and the adolescent self-image; 1965.

[41] Beck AT, Baruch E, Balter JM, Steer RA, Warman DM. A new instrument for measuring insight: the Beck cognitive insight scale. Schizophr Res 2004;68:319–29.

[42] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[43] Django Software Foundation. Django. https://djangoproject.com, 2019.

[44] Naidu G, Zuva T, Sibanda EM. A review of evaluation metrics in machine learning algorithms. In: Computer science on-line conference. Springer; 2023. p. 15–25.

[45] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 2020;21:1–13.

[46] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[47] Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26:1340–7.

[48] Ke G, Meng Q, Finley T, Wang T, Chen W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.

[49] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232.

[50] Sutton CD. Classification and regression trees, bagging, and boosting. Handbook of statistics, vol. 24. 2005. p. 303–29.

[51] Bertsimas D, King A. Logistic regression: from art to science. Stat Sci 2017:367–84.

[52] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. ´SMOTE: synthetic minority oversampling technique. J Artif Intell Res 2002;16:321–57.

[53] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 2014;41:647–65.

[54] Shapley LS, et al. A Value for N-person Games; 1953.

[55] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, et al. From local explanations to global understanding with explainable ai for trees. Nat Mach Intell 2020;2:2522–5839.

[56] Equal Employment Opportunity Commission. Uniform guidelines on employee selection procedures (1978). https://www.govinfo.gov/content/pkg/CFR-2014-title29-vol4/xml/CFR-2014-title29-vol4-part1607.xml, 2014. Code of Federal Regulations, Title 29, Part 1607.

[57] Wisniewski J, Biecek P. Fairmodels: a flexible tool for bias detection, visualization, and mitigation in binary classification models. R J 2022;14:227–43.

[58] Latif E, Zhai X, Liu L. Ai gender bias, disparities, and fairness: does training data matter? arXiv preprint. arXiv:2312.10833, 2023.

[59] Mowery BD. The paired t-test. Pediatr Nursing 2011;37:320–2.

[60] UPC Research Development Lab. Mydisk cloud platform. https://rdlab.cs.upc.edu/mydisk/, 2025. [Accessed 28 April 2025].

[61] Eke CI, Shuib L. The role of explainability and transparency in fostering trust in ai healthcare systems: a systematic literature review, open issues and potential solutions. Neural Comput Appl 2025;37:1999–2034.

[62] Okwor IA, Hitch G, Hakkim S, Akbar S, Sookhoo D, Kainesie J. Digital technologies impact on healthcare delivery: a systematic review of artificial intelligence (AI) and machine-learning (ML) adoption, challenges, and opportunities. AI 2024;5:1918–41.

[63] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digit Med 2021;4:31.

[64] Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations can reduce overreliance on ai systems during decision-making. Proc ACM Hum-Comput Interact 2023;7:1–38.

[65] Ţenea Cojan S-T, Dinescu V-C, Gheorman V, Dragne I-G, Gheorman V, et al. Exploring multidisciplinary approaches to comorbid psychiatric and medical disorders: a scoping review. Life 2025;15:251.

[66] Pedersen G, Urnes ø, Hummelen B, Wilberg T, Kvarstein E. Revised manual for the global assessment of functioning scale. Eur Psychiatry 2018;51:16–9.

[67] Frost BG, Tirupati S, Johnston S, Turrell M, Lewin TJ, et al. An Integrated Recovery-oriented Model (IRM) for mental health services: evolution and challenges. BMC Psychiatry 2017;17:22.

[68] Üstün TB, Kostanjsek N, Chatterji S, Rehm J. Measuring health and disability: manual for WHO disability assessment schedule (WHODAS 2.0). Geneva: World Health Organization; 2010.

[69] Morosini PL, Magliano L, Brambilla L, Ugolini S, Pioli R. Development, reliability and acceptability of a new version of the DSM-IV social and occupational functioning assessment scale (sofas) to assess routine social functioning. Acta Psychiatr Scand 2000;101:323–9.

[70] Watson AJ, Harrison L, Preti A, Wykes T, Cella M. Cognitive trajectories following onset of psychosis: a meta-analysis. Br J Psychiatry 2022;221:714–21.

[71] Vita A, Gaebel W, Mucci A, Sachs G, Erfurth A, et al. European Psychiatric Association guidance on assessment of cognitive impairment in schizophrenia. Eur Psychiatry 2022;65:e58.

[72] Horvath AO, Del Re A, Flückiger C, Symonds D. Alliance in individual psychotherapy. Psychotherapy 2011;48:9.

[73] Stefana A, Fusar-Poli P, Vieta E, Youngstrom EA. Patients' perspective on the therapeutic relationship and session quality: the central role of alliance. Front Psychol 2024;15:1367516.

[74] Horvath AO, Greenberg LS. Development and validation of the working alliance inventory. J Couns Psychol 1989;36:223.

[75] Duncan BL, Miller SD, Sparks JA, Claud DA, Reynolds LR, et al. The session rating scale: preliminary psychometric properties of a “working” alliance measure. J Brief Ther 2003;3:3–12.

[76] Bartholomew TT, Pérez-Rojas AE, Lockard AJ, Joy EE, Robbins KA, et al. Therapists' cultural comfort and clients' distress: an initial exploration. Psychotherapy 2021;58:275.

[77] Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res 2018;2:11.

[78] Pennestrì F, Cabitza F, Picerno N, Banfi G. Sharing reliable information worldwide: healthcare strategies based on artificial intelligence need external validation. Position paper. BMC Med Inform Decis Mak 2025;25:56.

[79] International Organization for Standardization. ISO 16290:2013 - space systems — definition of the Technology Readiness Levels (TRLs) and their criteria of assessment. https://www.iso.org/standard/56064.html, 2013.