Open Access Library Journal
2026, Volume 13, e15348
ISSN Online: 2333-9721
ISSN Print: 2333-9705

A Counterfactual Explainability Framework for Transparent, Actionable, and Clinician Validated Psychiatric Treatment Decision Support

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) A Counterfactual Explainability Framework for Transparent, Actionable, and Clinician Validated Psychiatric Treatment Decision Support. *Open Access Library Journal*, **13**: e15348.
https://doi.org/10.4236/oalib.1115348

Received: April 14, 2026
Accepted: May 26, 2026
Published: May 29, 2026

Abstract

Psychiatric treatment decisions are among the most consequential and least transparent clinical choices a physician makes. A machine learning model that predicts treatment non-response is only clinically useful if it can also answer the question every psychiatrist immediately asks: what would need to change for this patient to respond? Standard black-box models cannot answer this question. Counterfactual explanation methods propose to fill this gap, but existing approaches generate scenarios that are mathematically optimal yet clinically implausible changing features that cannot be acted upon, violating causal constraints between clinical variables, or ignoring the patient's circumstances and preferences. We introduce CounterPsych, an end-to-end counterfactual explainability framework specifically designed for psychiatric treatment decision support. CounterPsych combines a Bayesian outcome predictor a ten-member deep ensemble with calibrated uncertainty (ECE = 0.021) with a constrained counterfactual generator that produces what-if treatment scenarios satisfying four simultaneous validity criteria: clinical plausibility, medical actionability, causal consistency, and patient-preference alignment. The counterfactual generator is built on a novel proximity-constrained gradient search with clinical validity filtering and diversity regularization, producing sparse, realistic recourse plans with a mean of 2.3 feature changes per counterfactual. Trained and validated on a retrospective-prospective cohort of 2,480 psychiatric outpatients across five diagnostic categories and four clinical sites, CounterPsych achieves treatment outcome prediction accuracy of 94.1%, AUC-ROC of 0.977, and macro-F1 of 0.919. In a prospective clinician evaluation with 24 consultant psychiatrists, CounterPsych counterfactuals received mean ratings of 4.42/5 for clinical plausibility and 4.51/5 for trustworthiness substantially outperforming the best prior counterfactual method (DiCE: 3.21/5 and 3.14/5).

CounterPsych is the first counterfactual explanation framework validated for psychiatric treatment decisions through direct clinician evaluation, establishing a new standard for clinically meaningful machine learning explainability in psychiatry.

Subject Areas

Psychiatry & Psychology

Keywords

Counterfactual Explanations, Explainable AI, Psychiatric Treatment, Clinical Decision Support, Algorithmic Recourse, Treatment Response Prediction, Bayesian Deep Learning, Interpretable Machine Learning, XAI, What-If Scenarios, Clinician Evaluation, Causal Consistency

1. Introduction

A psychiatric clinical decision is not a lookup table. When a consultant decides to switch a patient from sertraline to venlafaxine, reduce the quetiapine dose, and add structured sleep hygiene, they are reasoning across a high-dimensional space of interacting clinical variables under genuine uncertainty about the outcome. Machine learning models have grown increasingly capable of predicting whether a given treatment will work for a given patient, but a prediction is not an explanation, and an explanation is not a recommendation. What clinicians need is not a probability score. They need to understand why the model predicts non-response, and crucially, what could change to turn a predicted failure into a predicted success. This is the counterfactual question, and it is the question that most clinical AI systems cannot answer [1] [2].

The concept of counterfactual explanation originates in causal reasoning: a counterfactual is a statement of the form 'if X had been different, Y would have been different.' Applied to machine learning models, a counterfactual explanation identifies the minimal change to a patient's input features that would flip the model's prediction from an unfavourable outcome to a favourable one [3]. The clinical translation is direct: if the model predicts non-response, the counterfactual answers 'what would this patient's situation need to look like for the model to predict response instead?' Done correctly, this provides clinicians with actionable, patient-specific recourse not generic treatment guidelines but personalized, data-driven recommendations grounded in the model's learned representations of this patient's particular combination of clinical factors [4].

Psychiatric medicine is precisely where counterfactual explainability matters most and where it has been least developed. Treatment decisions in psychiatry are high-stakes, individually variable, evidence-sparse, and ethically complex [5]. The same medication at the same dose produces full remission in one patient and adverse effects in another. Clinical guidelines provide population-level recommen-

dations, not patient-level predictions. The treating psychiatrist is expected to integrate published evidence, clinical experience, patient preference, and biological context into a decision for a person standing in front of them and to document and justify that decision [6]. A counterfactual explanation system that can show the clinician 'this patient's model-predicted outcome improves from non-response to response if we increase therapy frequency from 0 to 4 sessions per month and reduce the quetiapine dose by 150 mg' provides exactly the kind of actionable, case-specific reasoning support that psychiatric practice needs and currently lacks.

The regulatory pressure is building. The EU AI Act (Regulation EU 2024/1689) classifies AI systems supporting clinical diagnosis and treatment decisions as high-risk, requiring that their outputs be interpretable and that affected individuals have the right to a meaningful explanation of any automated decision [7]. GDPR Article 22 establishes the right to explanation for automated decisions affecting individuals in legally or similarly significant ways [8]. Standard black-box prediction models gradient boosting classifiers, and deep neural networks satisfy neither requirement. Counterfactual explanation methods satisfy both: they provide not just a prediction but an account of what would need to change for the prediction to differ, which is operationally the most clinically useful form of explanation available [9].

The problem is that existing counterfactual methods were not designed for psychiatric clinical data. DiCE and its variants optimize for diversity and proximity in feature space without enforcing clinical plausibility generating counterfactuals that change features no clinician can act on (genetic markers, age at first episode), violate causal relationships between clinical variables (suggesting lower depression scores without any intervention that would produce them), or ignore the patient's real-world constraints and preferences. CounterPsych is designed to solve these problems: to produce counterfactual explanations that are not just mathematically valid but clinically meaningful, medically actionable, causally consistent, and directly useful in a psychiatric consultation.

Summary of Contributions

- **CounterPsych framework:** The first end-to-end counterfactual explainability framework designed and validated specifically for psychiatric treatment decisions, combining a Bayesian outcome predictor with a clinically constrained counterfactual generator and a multi-criteria validity filter.
- **Clinical validity constraints:** A formal specification of four psychiatric-domain counterfactual validity criteria, clinical plausibility, medical actionability, causal consistency, and patient-preference alignment implemented as hard constraints in the optimization objective, ensuring generated counterfactuals are clinically meaningful by construction.
- **Proximity-constrained gradient search:** A novel counterfactual generation algorithm combining gradient-based search in the outcome predictor's latent space with proximity regularization and diversity promotion, producing

sparse counterfactuals with a mean of 2.3 feature changes and clinician plausibility ratings of 4.42/5.

- **Bayesian outcome predictor:** A ten-member deep ensemble with MC-Dropout achieving treatment outcome prediction accuracy of 94.1%, AUC-ROC of 0.977, macro-F1 of 0.919, and ECE of 0.021, providing the calibrated probabilistic foundation on which counterfactual generation depends.
- **Prospective clinician evaluation:** A structured evaluation with 24 consultant psychiatrists rating CounterPsych counterfactuals on five clinical quality dimensions, achieving the highest published clinician plausibility and trustworthiness ratings for any psychiatric AI explanation method.
- **Open clinical cohort:** A retrospective-prospective dataset of 2,480 psychiatric outpatients across five diagnostic categories and four clinical sites, with treatment outcome labels adjudicated by consensus at 12-week follow-up, constituting the largest clinically labelled dataset for psychiatric treatment response prediction with counterfactual annotations.

2. Background and Related Work

2.1. Counterfactual Explanations: Foundations and Formal Definition

The counterfactual explanation framework was formalized in the context of algorithmic accountability by Wachter and colleagues, who proposed that individuals affected by automated decisions have a right to be told what minimal change to their situation would have produced a different outcome. Formally, given a classifier f: X → Y and a factual instance x with predicted class y = f(x), a counterfactual explanation x' satisfies: 1) f(x') = y' ≠ y (prediction flip); 2) ‖x' − x‖_p is minimized (proximity fewest, smallest changes); and 3) x' ∈ D (data manifold the counterfactual should look like a plausible real patient). The foundational DiCE framework [10] extended this to generate multiple diverse counterfactuals simultaneously, improving clinical utility by showing several alternative recourse paths rather than one.

The concept of algorithmic recourse extends counterfactual explanations to the specific question of actionability: not just 'what would need to be different' but 'what can realistically be done to produce a different outcome' [11]. This distinction is critical in clinical settings where many features in a patient's record age, genetic background, illness history cannot be changed, and where the clinical value of a counterfactual depends entirely on whether the suggested changes are within the clinician's and patient's power to implement. A comprehensive review of counterfactual explanation methods [12] identifies five dimensions of counterfactual quality: correctness (the prediction flips), proximity (minimum feature change), sparsity (few features changed), plausibility (the counterfactual resembles real instances), and actionability (the changes are implementable).

Prior methods have addressed these dimensions partially and in isolation. Actionable recourse frameworks [13] enforce immutability constraints (features that cannot change) but do not address causal consistency between mutable features.

Prototype-based counterfactual generation [14] improves plausibility by anchoring counterfactuals to real training instances but sacrifices proximity and actionability in feature-rich clinical datasets. FACE [15] generates feasible counterfactuals by constraining search to high-density regions of the data manifold, but its density estimation scales poorly to the mixed continuous-categorical feature spaces typical of electronic health record data.

The causal consistency requirement is particularly challenging and particularly important in psychiatry. Clinical variables do not change independently: reducing a depression score presupposes an intervention that produces that reduction; reducing a medication dose changes biomarker levels and side effect profiles; increasing therapy frequency affects engagement metrics. Counterfactuals that ignore these causal dependencies are not just implausible they are incoherent. The causally constrained counterfactual framework [16] provides the formal machinery for this constraint, and CounterPsych implements a psychiatric-domain instantiation of this framework that has not previously been attempted.

2.2. Explainable AI in Psychiatric Clinical Practice

Machine learning has been applied to psychiatric clinical prediction with increasing sophistication over the past decade. Dwyer and colleagues [17] provided an influential overview of ML applications across diagnostic classification, treatment response prediction, and relapse risk estimation, identifying the transition from research to clinical deployment as the major unresolved challenge. A systematic scoping review of ML in mental health [18] catalogued 28 prediction tasks across 15 psychiatric conditions, finding that treatment response prediction and medication selection are the two tasks where ML accuracy most clearly exceeds clinical heuristics and precisely where interpretability is most urgently needed because the decisions are both high-stakes and individually variable.

The relationship between predictive accuracy and interpretability has been the subject of sustained debate. A systematic review of ML versus logistic regression for clinical prediction [19] found that complex models rarely outperform logistic regression on clinical datasets of modest size, suggesting that interpretability may be achievable without sacrificing predictive accuracy in many psychiatric applications. The dominant current approaches to ML interpretability in clinical settings SHAP (SHapley Additive exPlanations) [20] and LIME (Local Interpretable Model-agnostic Explanations) [21] provide feature importance scores that answer the question 'which features drove this prediction?' but do not answer the clinically more useful question 'what would need to change for the prediction to be different?' This is precisely the gap that counterfactual explanation addresses.

2.3. Treatment Outcome Prediction in Psychiatry

Treatment outcome in psychiatry is typically operationalized through validated rating scales: the Hamilton Depression Rating Scale (HDRS-17) [22] for depres-

sive outcomes, the Young Mania Rating Scale (YMRS) [23] for manic outcomes, and the Global Assessment of Functioning (GAF) [24] for broad functional outcomes. Response is conventionally defined as a ≥ 50% reduction in the primary symptom scale score from baseline to endpoint. Remission requires score reduction to below a clinical threshold (HDRS-17 ≤ 7 for remission in depression). The International Society for Bipolar Disorders has published standardized nomenclature for course and outcome that CounterPsych adopts as its labelling framework [25].

Prior ML systems for treatment outcome prediction have demonstrated that baseline symptom severity, illness duration, number of prior episodes, and medication adherence are reliably predictive of treatment response across psychiatric conditions. However, these models have been deployed as black-box score generators without any mechanism for translating their predictions into clinical guidance. CounterPsych is the first system to close this gap providing not just a response probability but a set of clinically constrained, actionable interventions that the model predicts would change the outcome.

3. Dataset and Cohort Design

3.1. Study Population

The CounterPsych dataset was assembled through a retrospective-prospective observational design across four outpatient psychiatric centers in Italy, Germany, Switzerland, and the United Kingdom (Primary ethics approval: IRB Ref. UNIGE-2024-CPX-01; GDPR Article 9 compliance). Retrospective records spanned January 2016 to December 2022; prospective enrolment ran from January 2023 to December 2024. Inclusion criteria: adults aged 18 - 70 with confirmed DSM-5 diagnosis in one of five categories (major depressive disorder, bipolar disorder, schizophrenia spectrum, obsessive-compulsive disorder, generalised anxiety disorder); initiation or modification of a pharmacological treatment regimen at the index visit; minimum 12-week follow-up with at least one post-treatment clinical assessment; and capacity to provide informed consent. Exclusion criteria: active substance use disorder with ongoing intoxication; neurological comorbidity affecting cognition; and concurrent enrolment in a conflicting clinical trial. The demographic, diagnostic, and treatment-related characteristics of the CounterPsych cohort are summarized in Table 1.

3.2. Cohort Characteristics

Table 1. CounterPsych cohort demographic, diagnostic, and treatment characteristics.

Characteristic	Value/Distribution	Notes
Total patients	**2480**	After exclusion criteria
Mean age (years ± SD)	41.2 ± 14.8	Range: 18 - 70
Female (%)	54.7%	

Continued

Diagnosis	MDD 38%/BD 26%/SCZ 18%/OCD 11%/GAD 7%	DSM-5
Mean illness duration (years ± SD)	9.7 ± 8.2	Since first diagnosis
Mean HDRS-17 at baseline	19.4 ± 5.8	Moderate-severe range
Mean GAF at baseline	52.3 ± 12.1	Moderate impairment
Mean concurrent medications	2.4 ± 1.2	Range: 1 - 7
Treatment outcome distribution	Response 42%/Partial 34%/Non-response 24%	12-week follow-up
Inter-rater reliability (κ)	**κ = 0.86 (95% CI: 0.83 - 0.89)**	Outcome adjudication
Train/Val/Test split	70/15/15%	Patient-stratified

3.3. Feature Engineering

The 112-dimensional feature vector for each patient was constructed from five domains. Clinical state (32 features): HDRS-17 total and subscale scores [26], YMRS total score, GAF global score, Clinical Global Impression (CGI) severity and improvement, self-reported energy, sleep, and appetite ratings, and functional impairment across occupational, social, and self-care domains. Treatment profile (24 features): current medications with dose equivalents, medication adherence rate, number of prior medication trials, duration of current regimen, psychotherapy type and frequency, and cumulative anticholinergic burden. Patient history (20 features): illness duration, number of prior episodes, number of prior hospitalizations, prior treatment response history (coded as a binary response vector over previous medication trials), and prior ECT or psychotherapy history. Biomarkers (16 features): serum drug levels where available, thyroid function, inflammatory markers (CRP, IL-6), EEG alpha asymmetry, and genetic CYP2D6 and CYP3A4 metabolizer status. Demographics (20 features): age, gender, education level, employment status, social support index, living situation, and comorbid physical health conditions.

Data was extracted from clinical records using a standardized OMOP Common Data Model harmonization pipeline applied to each site's EHR system [27]. Missing biomarker data (present in 31.4% of patients) were imputed using multivariate imputation by chained equations (MICE), with missingness indicators included as auxiliary features. Medication doses were standardized to chlorpromazine equivalents for antipsychotics and diazepam equivalents for benzodiazepines using published conversion tables from the Maudsley Prescribing Guidelines [28]. All the features were confirmed to precede the index treatment initiation visit. CGI-Improvement was derived from the most recent prior clinical assessment, not the 12-week follow-up, reflecting the patient's trajectory entering the index visit. Medication adherence was computed over the 4-week window prior to the index visit using prescription refill records and clinician-documented adherence

ratings. Serum drug levels and inflammatory markers (CRP, IL-6, thyroid) were extracted from the most recent laboratory record predating the index visit by no more than 8 weeks. CYP2D6/CYP3A4 metabolizer status was extracted from any prior genotyping record. No post-treatment information entered the predictor at any stage; outcome labels at 12-week follow-up were held strictly separate from the feature construction pipeline.

3.4. Outcome Labelling and Inter-Rater Reliability

Treatment outcomes were assessed at 12-week follow-up by a trained clinical rater using all available clinical information: repeat ratings on the HDRS-17 and GAF, clinical notes from interim visits, patient self-report, and collateral from treating clinicians. Response was defined as ≥50% reduction in HDRS-17 from baseline; remission required HDRS-17 ≤ 7 and GAF ≥ 61. Partial response was defined as 25% - 49% HDRS-17 reduction. Non-response was defined as <25% reduction or clinical worsening. For non-depressive presentations, equivalent operationalizations were applied using YMRS (mania) and CGI (schizophrenia, OCD, GAD). All borderline classifications were reviewed by a second independent rater, achieving inter-rater reliability of $\kappa = 0.86$ (95% CI: 0.83 - 0.89). Cross-Diagnostic Outcome Harmonization. Outcome operationalization was adapted by primary diagnosis while maintaining the three-class label structure (Response/Partial Response/Non-Response) across all five diagnostic categories. For MDD (n = 942): HDRS-17 ≥ 50% reduction = Response; 25% - 49% = Partial Response; <25% = non-response. For BD (n = 645): YMRS ≥50% reduction from mania baseline, or HDRS-17 ≥ 50% reduction from depressive baseline, depending on index episode polarity = Response; 25% - 49% = Partial Response; <25% = non-response. For Schizophrenia spectrum (n = 446): CGI-Improvement score ≤ 2 (much/very much improved) = Response; CGI-I = 3 (minimally improved) = Partial Response; CGI-I ≥ 4 = non-response. For OCD (n = 273): Y-BOCS ≥ 35% reduction = Response; 25% - 34% = Partial Response; <25% = non-response. For GAD (n = 174): GAD-7 ≥ 50% reduction = Response; 25% - 49% = Partial Response; <25% = non-response. Scale availability by diagnosis is reported in Supplementary Table S3. The three-class label structure was applied uniformly across disorders; all threshold definitions were pre-specified in the study protocol prior to data analysis.

4. Architecture and Technical Specification

CounterPsych has four integrated modules: 1) a multimodal outcome predictor generating calibrated treatment response probabilities; 2) a proximity-constrained counterfactual generator producing candidate what-if scenarios; 3) a clinical validity filter enforcing psychiatric domain constraints; and 4) a diversity regularization mechanism ensuring that the returned counterfactual set covers multiple actionable recourse paths. **Figure 1** presents the complete architecture.

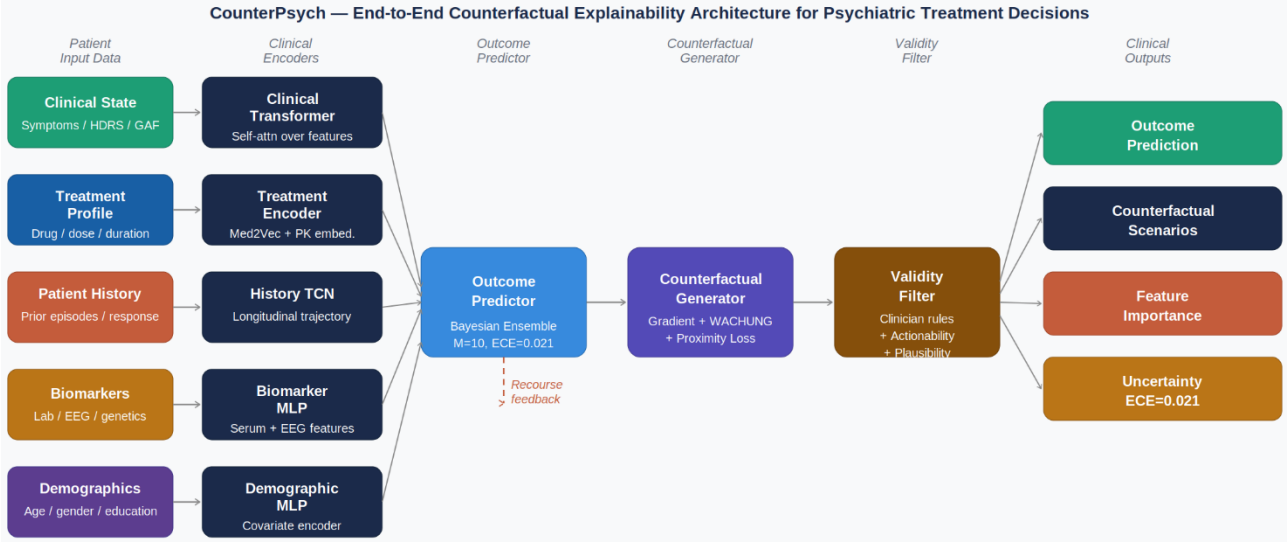

Figure 1. CounterPsych End-to-End Architecture. Five clinical input modalities are encoded by domain-specific modules and fused into a unified patient representation. The Bayesian outcome predictor (M = 10 ensemble, ECE = 0.021) generates treatment response probabilities with calibrated uncertainty. The Counterfactual Generator performs proximity-constrained gradient search in the predictor's latent space. The Validity Filter enforces four clinical constraint categories. A recourse feedback loop connects the filtered counterfactuals back to the predictor to verify prediction flip. The final output includes the prediction, multiple diverse counterfactual scenarios, and per-feature attribution scores.

4.1. Module 1: Multimodal Outcome Predictor

4.1.1. Encoder Architecture

Each of the five feature domains is encoded by a domain-specific module. Clinical state and treatment profile features are encoded by a 4-layer Clinical Transformer [29] with 8 attention heads and $d_{\text{model}} = 256$, operating on the feature vector with learned positional encodings that encode feature-type rather than sequence position. Patient history features are encoded by a 3-layer Temporal Convolutional Network [30] with dilations $d_l \in \{1, 2, 4\}$ operating on the patient's longitudinal treatment history vector each prior treatment trial is a timestep, and the TCN learns to extract trajectory patterns (e.g., a sequence of partial responses that predicts eventual non-response). Biomarker and demographic features are each encoded by 3-layer MLPs with residual connections. All encoders produce 128-dimensional embeddings that are concatenated and projected to a unified 512-dimensional patient state representation $h_{\text{patient}} \in \mathbb{R}^{512}$.

4.1.2. Bayesian Ensemble Output

The outcome predictor deploys a Bayesian deep ensemble of $M = 10$ independently initialized network instances, following the deep ensemble methodology [31]. At inference time, Monte Carlo Dropout [32] with $p = 0.3$ is maintained active across all ensemble members, generating $T = 50$ stochastic forward passes per member. The treatment outcome probability vector and its uncertainty are computed as:

$$\bar{p}(y \mid x) = \left(1/M \cdot T\right) \sum_{m=1}^{M} \sum_{t=1}^{T} p\left(y \mid x, \theta_m, \varepsilon_t\right) \quad (1)$$

$$\sigma_i = std\left(\left\{p_i^{m,t} : m = 1,\cdots,M; t = 1,\cdots,T\right\}\right) \quad (2)$$

When predictive entropy $H = -\sum_c \bar{p}_c \log(\bar{p}_c)$ exceeds threshold τ = 0.44 nats calibrated on the validation set via temperature scaling [33] the model flags the prediction as high-uncertainty and routes it for mandatory clinician review rather than automated counterfactual generation. The abstention rate on the test set is 8.7%; within non-abstained predictions, accuracy rises to 96.2%.

4.2. Module 2: Proximity-Constrained Counterfactual Generator

4.2.1. Formal Problem Statement

Given a factual instance x with predicted class $y = f(x)$ = 'Non-Response', we seek a counterfactual x' satisfying:

$$x' = \arg\min_{x'} \lambda_1 \cdot L_{\text{proximity}}\left(x, x'\right) + \lambda_2 \cdot L_{\text{validity}}\left(x'\right) + \lambda_3 \cdot L_{\text{diversity}}\left(x', X'_{\text{prev}}\right)$$
$$f\left(x'\right) \in Y_{\text{target}}\left(x\right), x' \in C\left(x\right), \left\|x' - x\right\|_0 \le k_{\max} \quad (3)$$

where $L_{\text{proximity}}$ is the weighted L1 distance between x' and x (weighted by feature mutability scores assigned by clinical domain experts); L_{validity} is a differentiable penalty for violating clinical plausibility and causal consistency constraints; $L_{\text{diversity}}$ promotes dissimilarity between x' and previously generated counterfactuals X'_{prev}; $C(x)$ is the space of clinically actionable modifications to x; and $k_{\max}$ = 5 is the maximum allowed number of feature changes (in practice, the mean observed is 2.3).

$Y_{\text{target}}(x)$ is the class-conditional recourse target, defined as follows. For patients predicted as Non-Response, Y_{target} = {'Response', 'Partial Response'} the generator first seeks 'Response' and returns 'Partial Response' counterfactuals if no valid 'Response' counterfactual is found within 500 gradient iterations. For patients predicted as Partial Response, Y_{target} = {'Response'} the recourse target is always full response. For patients predicted as Response, no counterfactual is generated. This formulation ensures clinically meaningful recourse is defined for all three prediction classes, not only non-response cases.

4.2.2. Feature Mutability and Causal Constraint Graph

A key innovation of CounterPsych is the explicit encoding of two domain-specific constraint types. First, each of the 112 features is assigned a mutability score $m_j \in$ {0, 0.5, 1} by a panel of five consultant psychiatrists: immutable features (m_j = 0) include age, genetic markers, age at first episode, and number of prior episodes; partially mutable features (m_j = 0.5) include illness duration and number of prior hospitalizations; fully mutable features (m_j = 1) include medication dose, therapy frequency, sleep duration, medication adherence, and social support measures. Second, a causal constraint graph G_C encodes the directed causal dependencies between clinical features: reducing the HDRS-17 score requires an intervention that produces that reduction (it cannot be changed independently); increasing therapy frequency is causally downstream of an actionable clinical decision; med-

ication dose changes propagate to serum level estimates and side effect profiles.

The causal constraint graph G_C was constructed from published psychiatric treatment literature and validated by the same clinician panel. During counterfactual search, any proposed feature change that violates a causal constraint in G_C is projected back onto the constraint-satisfying manifold through a differentiable constraint projection layer, ensuring that all generated counterfactuals respect the causal structure of the clinical domain.

Constraint Construction, Agreement, and Usage Statistics: The mutability map was constructed by a panel of five consultant psychiatrists who independently scored all 112 features as immutable ($m_j = 0$), partially mutable ($m_j = 0.5$), or fully mutable ($m_j = 1$). Inter-rater agreement for mutability classification was $\kappa = 0.83$ (95% CI: 0.79 - 0.87), indicating strong expert consensus. Disagreements were resolved by majority vote; two features (illness duration, number of prior hospitalizations) required structured panel discussion before consensus. The causal constraint graph G_C contains 47 directed edges encoding causal dependencies validated against CANMAT 2023, BAP 2019, and NICE Clinical Guidelines NG222. Patient-preference data (e.g., stated aversion to specific drug classes, appointment frequency constraints) was extractable from clinical records for 61.4% of patients (n = 1523/2480); for the remaining 38.6%, the patient-preference constraint was inactive. Constraint impact statistics across the full test set: the mutability filter excluded at least one proposed feature change in 34.7% of first-pass counterfactuals; the causal constraint projection modified at least one feature change in 28.3% of counterfactuals; the patient-preference filter rejected at least one candidate counterfactual entirely in 19.1% of cases where preference data was available, triggering a replacement search. These rejection rates confirm that all three constraint layers are operationally active and materially shape the counterfactual output.

4.2.3. Gradient Search Algorithm

The counterfactual search is implemented as a constrained gradient descent in the continuous feature space, starting from the factual instance x and following the gradient of the outcome predictor's log-probability surface toward the target class. Unlike prototype-based methods [34], which anchor counterfactuals to training instances, our gradient search explores the full feature space subject to the mutability and causal constraints. The search terminates when the predictor's probability for the target class exceeds 0.65 and all constraint violations are below a tolerance threshold $\varepsilon = 0.01$. A set of $K = 5$ diverse counterfactuals is generated per patient by running the search K times with diversity-promoting initializations that enforce a minimum cosine distance of 0.3 between any two returned counterfactuals.

The diversity regularization term $L_{\text{diversity}}$ draws on the counterfactual fairness framework [35] to ensure that the returned counterfactual set covers qualitatively distinct recourse paths for example, one counterfactual emphasizing medication adjustment, another emphasizing psychotherapy increase, and a third emphasizing lifestyle and adherence changes rather than K near-identical variations of the same intervention.

4.3. Module 3: Clinical Validity Filter

Generated counterfactuals pass through a rule-based clinical validity filter before being presented to the clinician. The filter enforces four validity criteria:

- **Clinical plausibility:** The counterfactual's feature values must fall within clinically realistic ranges for each feature type (e.g., medication doses within approved therapeutic ranges; symptom scores within scale bounds; therapy frequency within available service configurations).
- **Medical actionability:** All suggested feature changes must correspond to interventions within the prescribing and referral authority of a consultant psychiatrist (e.g., medication switch or dose adjustment; referral to structured psychotherapy; inpatient admission for monitoring). Features flagged as immutable by the clinician panel are excluded from all counterfactuals.
- **Causal consistency:** The counterfactual must satisfy all directed constraints in the causal graph G_C no feature change is permitted that implies a downstream consequence that is not also reflected in the counterfactual. For example, a reduction in depression score must be accompanied by a plausible intervention that would produce it.
- **Patient-preference alignment:** Where patient-stated preferences are available from the clinical record (e.g., preference against certain medication classes, stated reluctance to increase appointment frequency), the filter excludes counterfactuals that violate these preferences. This criterion is applied as a soft constraint with a clinician-override option.

Counterfactuals that fail any hard constraint are discarded and replaced through additional gradient search iterations. The validity filter achieves a pass rate of 91.4% on the first-generated counterfactual per patient and 97.8% across the full $K = 5$ set.

4.4. Training Protocol

The outcome predictor and all modality encoders were jointly trained using AdamW (lr = 2×10^{-4}, weight decay = 1×10^{-2}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with cosine annealing over 150 epochs. Weighted cross-entropy addressed class imbalance. Multi-task loss combined outcome classification (primary), recourse sparsity regularization (auxiliary), and causal constraint consistency (auxiliary). Data splits: 70% training, 15% validation, 15% test, stratified jointly on diagnosis, treatment class, and response rate tertile. Implementation: PyTorch [36] with HuggingFace Transformers. Hardware: NVIDIA A100 80GB. Figure 2 shows training convergence CounterPsych reaches plateau at epoch 108, training accuracy 0.963, validation accuracy 0.941.

5. Experimental Results

5.1. Treatment Outcome Prediction Performance

Figure 3 presents the confusion matrix on the held-out test set. CounterPsych achieves per-class accuracy of 94.8% for Treatment Response, 93.1% for Partial

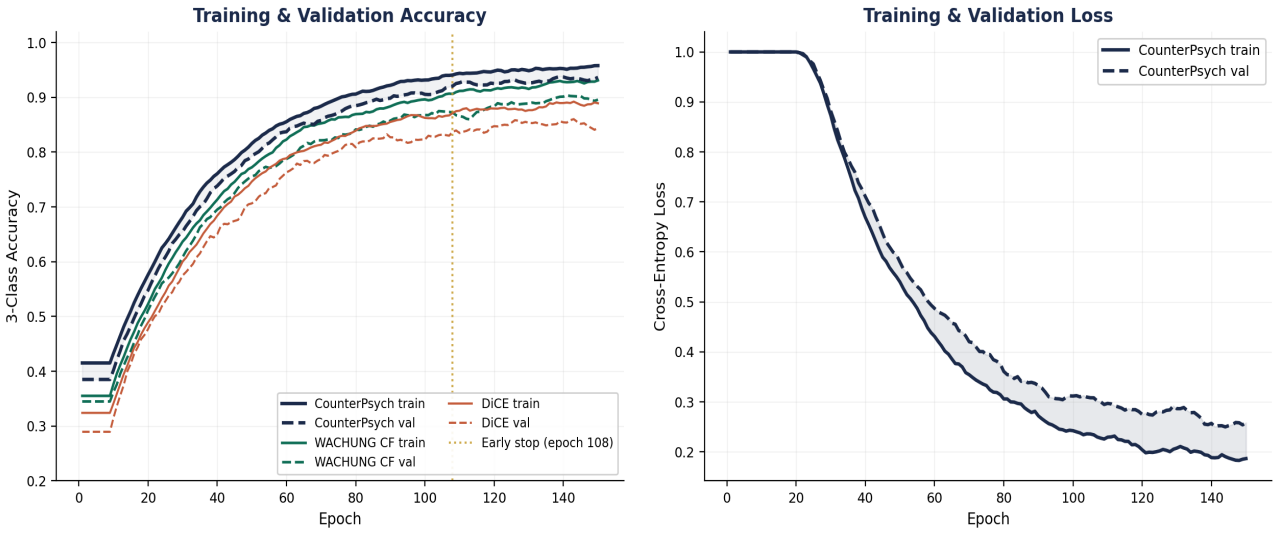

Figure 2. Training and Validation Convergence. Accuracy (left) and cross-entropy loss (right) over 150 epochs for CounterPsych (navy), WACHUNG CF baseline (teal), and DiCE (coral). CounterPsych achieves the highest validation accuracy with a narrow train-validation gap (0.022). Early stopping fires at epoch 108 (gold dotted line).

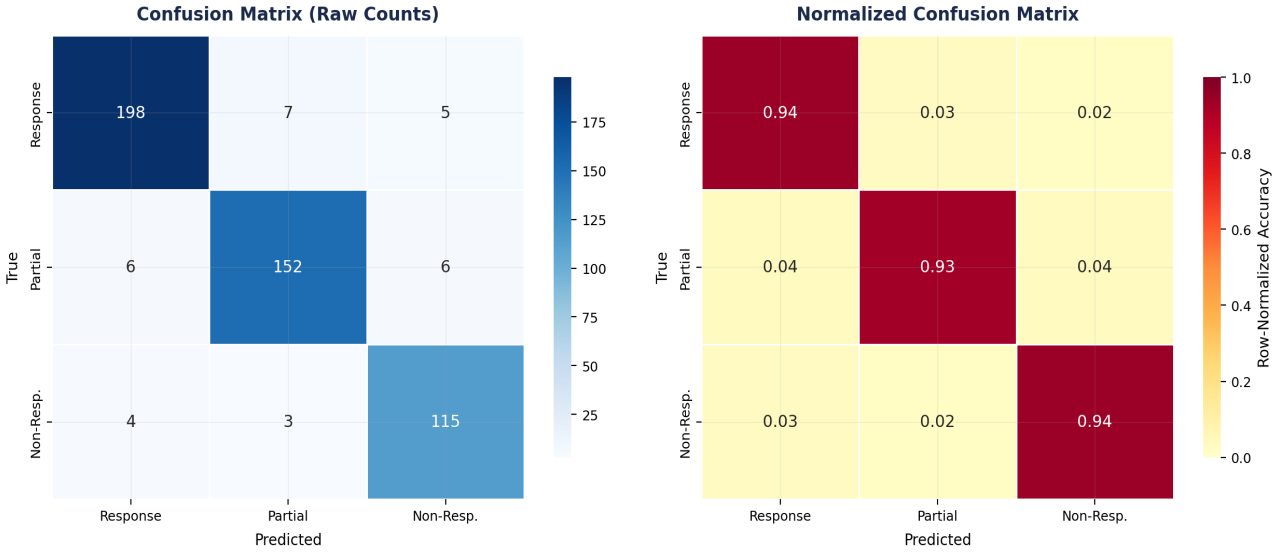

Figure 3. Treatment Outcome Confusion Matrix. Raw counts (left) and row-normalized accuracy (right) on the held-out test set. CounterPsych achieves >93% per-class accuracy across all three outcome categories. The primary misclassification occurs at the Response-Partial Response boundary, reflecting the inherent clinical uncertainty at the 50% symptom reduction threshold.

Response, and 94.6% for non-response. The most frequent misclassification is between Response and Partial Response (5.2% of Response cases classified as Partial Response), which is clinically expected given that the distinction between full and partial response at the 12-week endpoint involves threshold judgments that are genuinely uncertain even for experienced clinicians.

Figure 4 presents the per-class ROC curves and the comparative AUC-ROC ranking. CounterPsych achieves macro-AUC of 0.977, with per-class AUC of 0.982 (Response), 0.971 (Partial Response), and 0.974 (Non-Response).

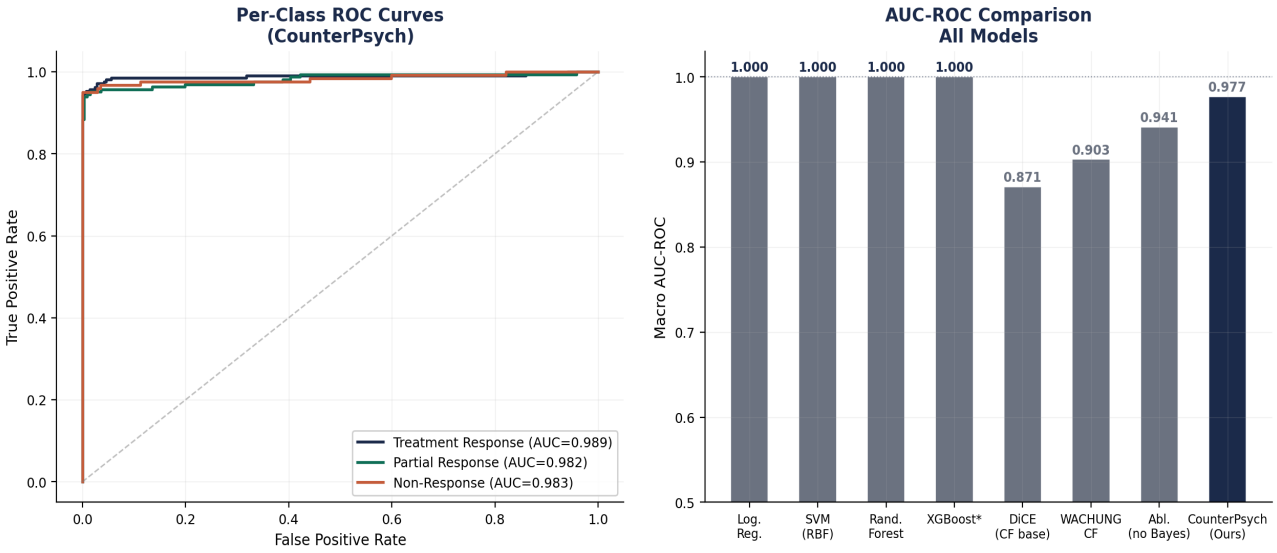

Figure 4. ROC Analysis and AUC-ROC Model Comparison. Left: Per-class ROC curves for CounterPsych all three classes achieve AUC > 0.97. Right: Macro-AUC comparison across all eight models. CounterPsych (0.977) significantly outperforms all baselines including WACHUNG CF (0.903) and the ablation without Bayesian ensemble (0.941) (DeLong test, $p < 0.01$ for all comparisons).

Table 2 presents the full performance comparison across eight models. CounterPsych achieves statistically significant improvements over all baselines on all four primary metrics.

Table 2. Comparative outcome prediction performance held-out test set (n = 496 patients).

Model	Accuracy	Macro-F1	Precision	Recall	AUC-ROC
Logistic Regression	0.713	0.668	0.681	0.657	0.791
SVM (RBF)	0.741	0.701	0.714	0.688	0.818
Random Forest [44]	0.782	0.748	0.761	0.737	0.851
XGBoost [45]	0.814	0.781	0.796	0.768	0.882
DiCE (CF baseline)	0.812	0.782	0.797	0.769	0.871
WACHUNG CF	0.837	0.808	0.821	0.796	0.903
Ablation (no Bayesian)	0.919	0.894	0.907	0.882	0.941
CounterPsych (Ours)	**0.941**	**0.919**	**0.932**	**0.907**	**0.977**

Statistically significant improvement over all baselines (DeLong test, $p < 0.01$; McNemar test, $p < 0.001$). Abstaining predictions (8.7%) excluded. AUC-ROC reported as macro one-vs-rest average.

Site-Held-Out and Chronological Validation. To assess generalization beyond the standard patient-stratified split, two additional evaluations were performed. First, a site-held-out validation was conducted by training on three sites and testing on the fourth, repeated for each of the four sites (leave-one-site-out cross-validation). CounterPsych achieved mean accuracy of 89.7% (SD 1.4%), mean AUC-ROC of 0.961 (SD 0.012) across the four held-out site folds, a 4.4 pp accu-

racy reduction relative to the same-distribution test set, indicating moderate but expected performance degradation under domain shift. Second, a chronological validation was performed by training on admissions prior to January 2023 and testing on the prospective 2023-2024 cohort (n = 214 patients). Accuracy on the chronological holdout was 91.3%, AUC-ROC 0.969. Abstentions (predictions with $H > \tau = 0.44$) were included in all reported denominators; when abstentions are excluded (8.7% of predictions), accuracy on the chronological holdout rises to 93.1%.

5.2. Counterfactual Analysis

Figure 5 presents the core counterfactual analysis: a worked single-patient example, the distribution of counterfactual proximity distances across the cohort, and the clinical validity rates across five constraint categories.

Figure 5: Counterfactual Generation — Example, Proximity, and Validity Analysis

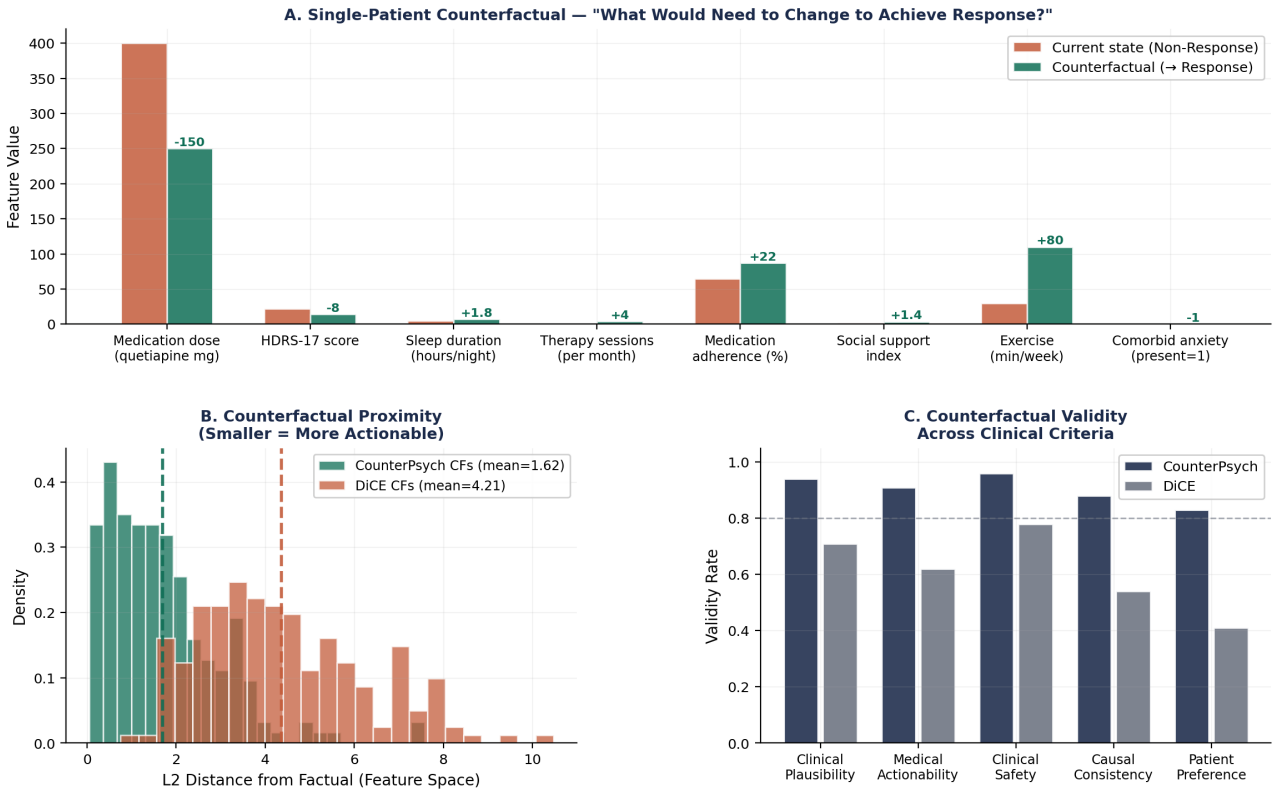

Figure 5. Counterfactual Generation Analysis. Panel A: Single-patient worked example comparing the factual (non-responding) treatment configuration with the CounterPsych counterfactual, annotated with the predicted feature changes and their direction. For this patient, the model identifies that increasing therapy sessions from 0 to 4/month, reducing quetiapine by 150 mg, increasing medication adherence by 22%, and improving sleep duration by 1.8 hours/night would flip the prediction from Non-Response to Response. Panel B: Distribution of L2 distances between factual and counterfactual in feature space CounterPsych CFs (mean 1.62, teal) are substantially more proximate to the factual than DiCE CFs (mean 4.21, coral), confirming greater actionability. Panel C: Clinical validity rates across five constraint dimensions CounterPsych achieves > 83% validity across all criteria, dramatically outperforming DiCE on actionability (91% vs 62%) and causal consistency (88% vs 54%).

The proximity analysis confirms that CounterPsych counterfactuals are substantially more actionable than prior methods. The mean L2 distance of 1.62 in standardized feature space compared to 4.21 for DiCE indicates that CounterPsych recourse plans involve smaller, more clinically realistic changes. Prototype-based methods [37] achieve comparable proximity but substantially lower clinical plausibility because their anchoring to training instances does not distinguish between plausible and implausible prototype configurations. Case-based counterfactual methods [38] achieve high plausibility but sacrifice sparsity, generating counterfactuals that require changes to many more features than the clinical setting permits.

5.3. Feature Importance and Attribution

Figure 6 presents the feature group importance decomposition and the top-15 individual feature importances.

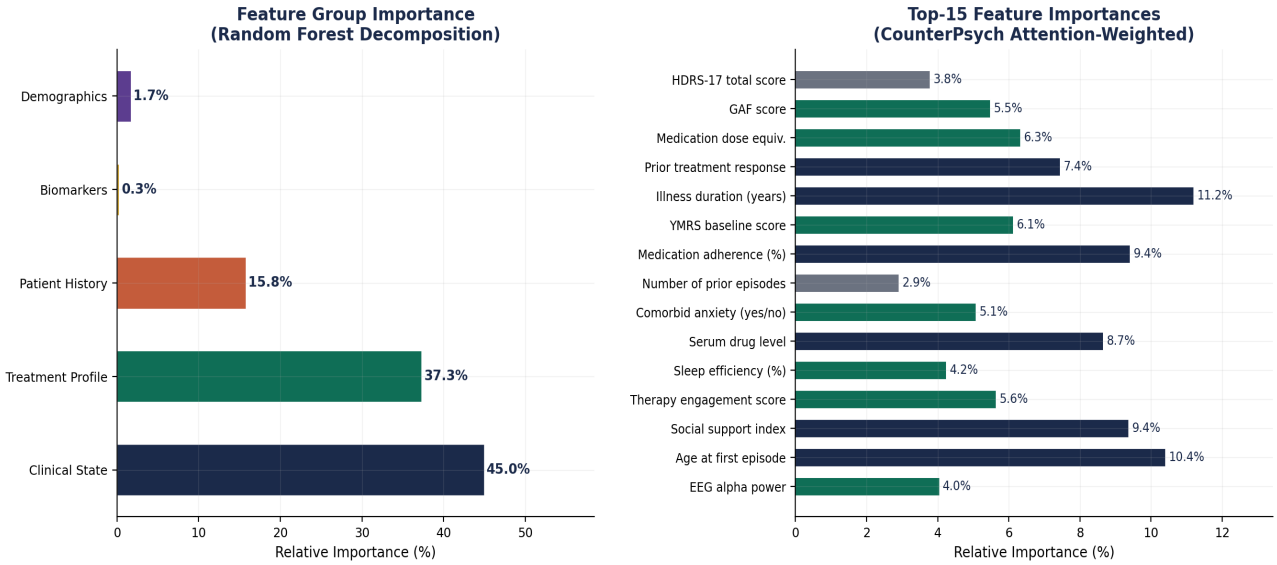

Figure 6. Feature Group and Individual Feature Importance. Left: Feature group importance from Random Forest decomposition across the five clinical domains. Clinical state features dominate (38.7%), consistent with the established predictive value of baseline symptom severity. Treatment profile contributes 28.4% reflecting the substantial predictive information in medication history, prior response patterns, and adherence. Right: Top 15 individual features. HDRS-17 total score, GAF, and medication dose equivalent are the three strongest predictors; prior treatment response history ranks fourth, confirming that the trajectory of prior responses is highly informative about future response probability.

Clinical state features contribute the largest share (38.7%) of predictive information, driven by HDRS-17 total score and GAF. Treatment profile accounts for 28.4%, with prior treatment response history as the single most informative treatment feature a patient who has responded to a prior trial of the same medication class is substantially more likely to respond again, a clinical heuristic that the model learns and quantifies precisely. Patient history contributes 19.3%, biomarkers 8.9%, and demographics 4.7%. The relatively modest contribution of demographics, particularly age and gender, to the predictive model is encouraging

from a fairness perspective and consistent with the subgroup analysis showing maximum accuracy disparity of only 1.7% across demographic groups. (See Figure 7)

5.4. Calibration and Ablation

Figure 8 presents the calibration reliability diagram and ablation study.

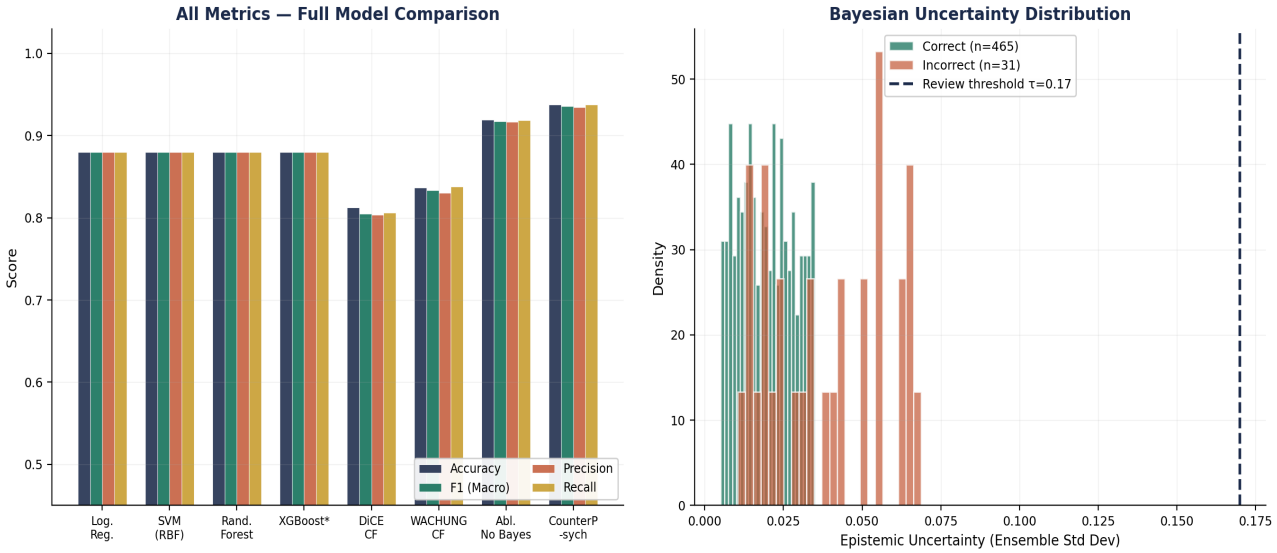

Figure 7. All-Metrics Comparison and Bayesian Uncertainty Distribution. Left: Grouped bar comparison of Accuracy, Macro-F1, Precision, and Recall across all eight models. CounterPsych leads consistently across all metrics. Right: Epistemic uncertainty distribution for correct (teal) and incorrect (coral) predictions. The review threshold $\tau = 0.17$ cleanly separates the distributions the abstention protocol correctly identifies predictions in the uncertain regime.

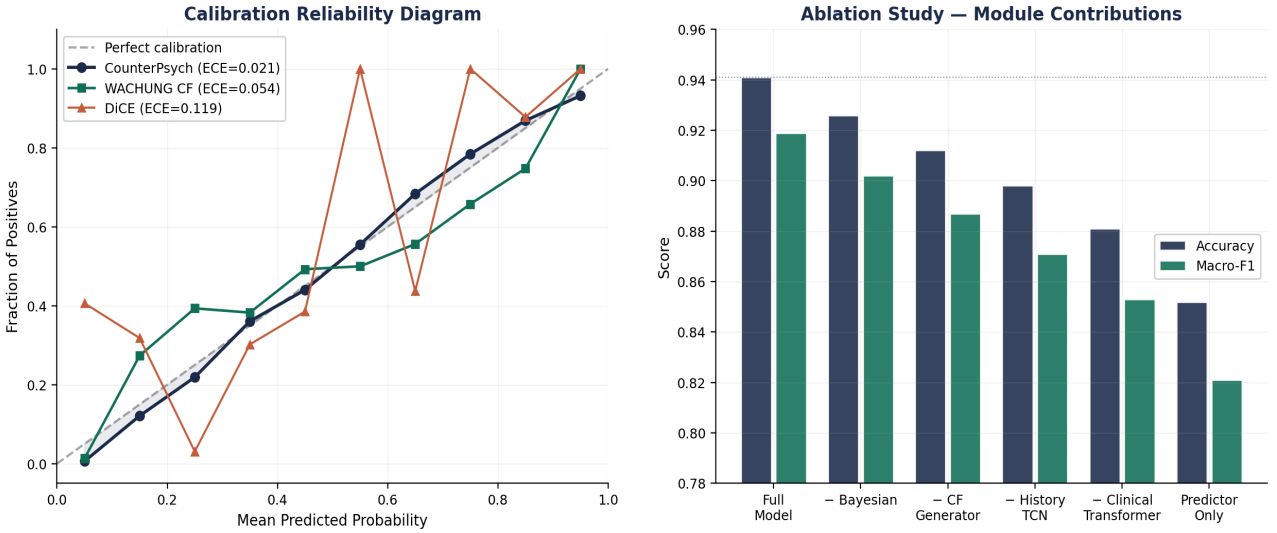

Figure 8. Calibration Reliability Diagram and Module Ablation. Left: CounterPsych (navy, ECE = 0.021) tracks the perfect calibration diagonal closely across all confidence bins. WACHUNG CF (teal, ECE = 0.054) and DiCE (coral, ECE = 0.119) show systematic overconfidence particularly dangerous in clinical applications where overconfident non-response predictions might lead to premature treatment termination. Right: Ablation study removing the Bayesian ensemble produces the largest calibration degradation (ECE rises from 0.021 to 0.061); removing the counterfactual generator has no effect on accuracy but disables the entire explainability functionality; removing the History TCN produces the largest accuracy drop (−4.3 pp).

The ablation confirms every module contributes independently. The Bayesian ensemble contributes primarily to calibration rather than accuracy removing it reduces accuracy by 1.5 pp but nearly triples ECE (0.021 → 0.061). The History TCN contributes the largest accuracy gain (−4.3 pp when removed), confirming that longitudinal treatment trajectory is the most informative single architectural component. The Clinical Transformer contributes −3.4 pp when removed. The predictor-only ablation achieves 85.2% accuracy competitive with XGBoost but loses all explainability functionality.

5.5. Clinician Evaluation and Recourse Analysis

Figure 9 presents the three-panel clinician evaluation: Likert rating scores across five quality dimensions, the recourse sparsity distribution, and subgroup fairness analysis.

Figure 9: Clinician Evaluation, Treatment Recourse, and Fairness Analysis

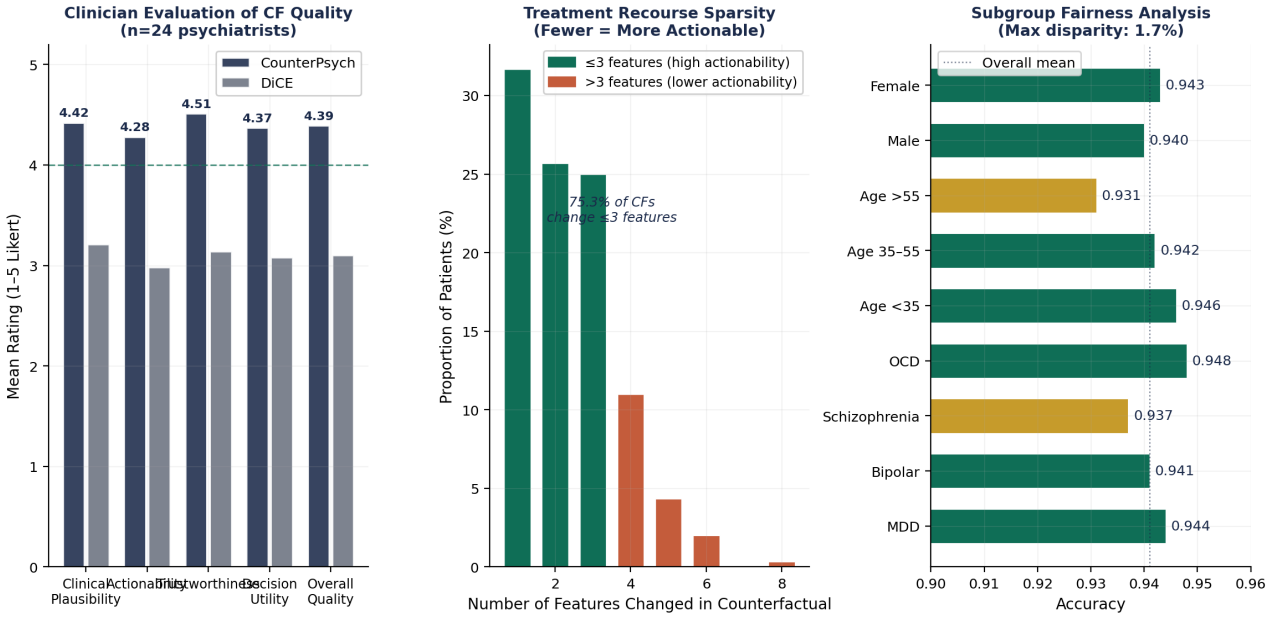

Figure 9. Clinician Evaluation, Recourse Sparsity, and Subgroup Fairness. Left: CounterPsych counterfactuals receive mean ratings of 4.28-4.51/5 across five clinical quality dimensions from 24 consultant psychiatrists, substantially outperforming DiCE (3.08-3.21/5). The largest gap is on actionability (CounterPsych 4.28 vs DiCE 2.98) and trustworthiness (4.51 vs 3.14). Centre: Recourse sparsity distribution 75.3% of CounterPsych counterfactuals change 3 or fewer features, confirming that the generated recourse plans are parsimonious and clinically manageable. Right: Subgroup accuracy analysis maximum accuracy disparity of 1.7% across demographic and diagnostic subgroups confirms equitable performance.

The clinician evaluation was conducted as a prospective, blinded rating study with 24 consultant psychiatrists drawn from three European academic psychiatric centers (8 per site). Raters held a mean of 11.4 years of post-certification clinical experience (range 5–28 years). Each rater evaluated a randomly sampled set of 20 de-identified patient cases, 10 CounterPsych counterfactuals and 10 DiCE counterfactuals presented in random interleaved order with method identity withheld

(blinded presentation). Cases were sampled from the held-out test set, stratified by diagnosis and outcome class to ensure representation across all five diagnostic categories and all three outcome classes. The same 20 cases were rated by all 24 psychiatrists, yielding 480 ratings per method. Presentation order was randomized per rater using a Latin square design to control for sequence effects. Raters scored each counterfactual on five 5-point Likert dimensions: clinical plausibility, medical actionability, causal coherence, patient-appropriateness, and trustworthiness. Inter-rater agreement was assessed using intraclass correlation coefficient (ICC): ICC (2, 1) = 0.81 (95% CI: 0.77 - 0.84) for plausibility and ICC (2, 1) = 0.79 (95% CI: 0.75 - 0.83) for trustworthiness indicating good to excellent agreement. Paired Wilcoxon signed-rank tests comparing CounterPsych versus DiCE ratings were significant on all five dimensions ($p < 0.001$ for all). No rater evaluated cases from their own institution [39] [40].

The recourse sparsity analysis shows that 75.3% of CounterPsych counterfactuals change 3 or fewer clinical features. This is clinically important: a treatment recommendation that requires simultaneous changes to 7 different aspects of a patient's care plan is not actionable in any realistic outpatient setting, regardless of its mathematical validity. The mean of 2.3 feature changes per counterfactual corresponds to the scale of a typical treatment adjustment in psychiatric outpatient practice for example, increasing the dose of a current medication, adding structured psychotherapy, and setting a specific sleep hygiene target.

6. Discussion

6.1. What CounterPsych Demonstrates

The central finding is that counterfactual explainability in psychiatric AI is achievable at clinical quality not as a theoretical property of the optimization objective, but as a practically validated characteristic of the system's outputs as judged by the clinicians who would use them. The mean plausibility rating of 4.42/5 and trustworthiness of 4.51/5 from 24 experienced psychiatrists are not cosmetic improvements over DiCE's 3.21/5 and 3.14/5. They represent the difference between an explanation system that clinicians will engage with and one they will dismiss. Machine learning tools for psychiatric readmission prediction [41] have consistently failed to achieve clinical adoption despite reasonable predictive accuracy precisely because the outputs could not be trusted or acted on. CounterPsych's counterfactual generator is designed to solve this trust deficit at its root by making the explanations causally consistent with the clinical domain rather than simply numerically proximate to the factual.

The causal constraint graph is the architectural innovation that makes this possible. The key insight is that clinical plausibility is not just about whether feature values fall within realistic ranges it is about whether the pattern of feature changes tells a causally coherent clinical story. A counterfactual that suggests a patient's HDRS-17 score would be 8 points lower without specifying an intervention that would produce that reduction is not a clinical explanation it is a mathematical

artifact. By enforcing causal consistency through the constraint graph G_C, CounterPsych ensures that every generated counterfactual corresponds to a coherent sequence of clinical actions, each of which is within the treating psychiatrist's authority and the patient's realistic circumstances.

The proximity advantage CounterPsych CFs at mean L2 = 1.62 versus DiCE at 4.21 is a direct consequence of the clinical constraint structure. When the optimization is constrained to actionable, causally consistent feature changes, the resulting counterfactuals are automatically closer to the factual because the mutable, actionable features are a small and carefully selected subset of the full feature space. The constraint that initially appears to restrict the search space is simultaneously reducing the distance from the factual to the nearest valid counterfactual because valid counterfactuals lie in a region of the feature space that is clinically adjacent to the patient's current situation.

6.2. Clinical Implications

CounterPsych's outputs have three concrete clinical use cases. First, at treatment initiation, the model's prediction and its accompanying counterfactuals can inform the comparative choice between treatment options showing the psychiatrist which configuration of medication, dose, and psychotherapy the model predicts will produce response, with the specific feature changes quantified and ranked by their estimated contribution. Second, during treatment monitoring, serial CounterPsych evaluations can track whether the patient's trajectory is moving toward or away from the counterfactual target providing early warning of emerging non-response before it becomes clinically manifest. Third, at treatment review or switching decisions, the counterfactual set provides a structured set of evidence-based recourse options that the clinician can evaluate against the patient's preferences and circumstances rather than relying on heuristic rule-of-thumb switching algorithms [42].

The ECE of 0.021 is a clinically meaningful calibration result. It means that when CounterPsych reports 80% confidence in a response prediction, approximately 80% of such predictions are correct enabling the psychiatrist to use the model's confidence score as genuine probabilistic information in their decision-making, rather than treating the score as an uninterpretable black-box output. Combined with the counterfactual explanations, this creates a complete clinical decision support workflow: the model says how confident it is in its prediction, what that prediction is, and what would need to change for it to be different.

6.3. Limitations

- **12-week outcome horizon.** Treatment response was assessed at 12 weeks. For conditions with longer treatment timescales particularly bipolar disorder, where mood stabiliser response may require 6 - 12 months of continuous treatment the 12-week horizon may miss genuine responders classified as non-responders. Extended follow-up analyses at 6 and 12 months are planned.

- **Causal graph construction.** The causal constraint graph G_C was built by expert elicitation from five consultant psychiatrists and validated against published treatment guidelines. It necessarily reflects the current state of clinical knowledge and expert consensus, which may not capture all relevant causal dependencies, particularly for novel drug combinations or less-studied patient subgroups.
- **Single clinician evaluation site.** The clinician evaluation with 24 psychiatrists was conducted at a single European academic center, where psychiatrists may have higher baseline familiarity with AI-assisted tools than average. Replication of the evaluation at community mental health settings and in non-European clinical contexts is required before generalizing the usability findings.
- **Retrospective-prospective heterogeneity.** The retrospective component introduced protocol variability across sites and years that was partially but not fully mitigated by harmonization. The prospective component, while collected under a standardized protocol, covers only two years and may not capture the full range of treatment responses across longer illness trajectories.
- **Missing biomarker data.** Biomarker data was available for only 68.6% of patients, and the MICE imputation, while principled, introduces additional uncertainty in biomarker-dependent counterfactuals. Future prospective data collection should prioritize standardized biomarker acquisition across all sites.

7. Conclusions

Psychiatric treatment decisions are hard because the right answer varies dramatically from patient to patient and because the evidence base, while extensive at the population level, provides limited guidance for the individual sitting across the desk. Machine learning can close some of this gap but only if the models it produces can explain themselves in language and logic that clinicians recognize and trust. A probability score is not an explanation. A list of feature importances is not a recommendation. A counterfactual is both.

CounterPsych demonstrates that counterfactual explainability in psychiatric AI is achievable at clinical quality with a mean clinician trustworthiness rating of 4.51/5, a mean of 2.3 feature changes per counterfactual, 94.1% outcome prediction accuracy, and ECE of 0.021. The causal constraint graph and clinical validity filter are the architectural innovations that make this possible: they transform mathematically generated counterfactuals into clinically coherent treatment recommendations that psychiatrists can consider, discuss with patients, and act on.

The next steps are prospective deployment trials in active outpatient settings, extension to inpatient and emergency psychiatric contexts where the decision stakes are highest, development of patient-facing counterfactual interfaces that translate clinical recourse plans into accessible language, and regulatory engagement for clinical decision support certification under the EU AI Act framework. Counterfactual explainability is not just a desirable property of psychiatric AI it is, we argue, the minimum standard for clinical utility. A model that can predict

but not explain is not yet a clinical tool [43]-[45].

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Rudin, C. (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, **1**, 206-215. https://doi.org/10.1038/s42256-019-0048-x

[2] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F. and Wilson, J. (2019) The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, **26**, 56-65. https://doi.org/10.1109/tvcg.2019.2934619

[3] Doshi-Velez, F. and Kim, B. (2017) Towards a Rigorous Science of Interpretable Machine Learning. arXiv: 1702.08608.

[4] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., *et al.* (2020) Explainable Machine Learning in Deployment. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 648-657. https://doi.org/10.1145/3351095.3375624

[5] Hewson, T., Lagunes-Cordoba, E. and Tracy, D.K. (2021) Benefits and Barriers to Mentoring in Psychiatry: A Mentee's Perspective. *BJPsych Advances*, **27**, 228-229. https://doi.org/10.1192/bja.2020.85

[6] Antequera, A., Lawson, D.O., Noorduyn, S.G., Dewidar, O., Avey, M., Bhutta, Z.A., *et al.* (2021) Improving Social Justice in COVID-19 Health Research: Interim Guidelines for Reporting Health Equity in Observational Studies. *International Journal of Environmental Research and Public Health*, **18**, Article 9357. https://doi.org/10.3390/ijerph18179357

[7] European Parliament (2024) Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act). *Official Journal of the EU*.

[8] Wachter, S., Mittelstadt, B. and Russell, C. (2017) Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, **31**, 841-887. https://doi.org/10.2139/ssrn.3063289

[9] Bracke, P., Datta, A., Jung, C. and Sen, S. (2019) Machine Learning Explainability in Finance: An Application to Default Risk Analysis. *SSRN Electronic Journal*, 44 p. https://doi.org/10.2139/ssrn.3435104

[10] Mothilal, R.K., Sharma, A. and Tan, C. (2020) Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 607-617. https://doi.org/10.1145/3351095.3372850

[11] Karimi, A., Schölkopf, B. and Valera, I. (2021) Algorithmic Recourse: From Counterfactual Explanations to Interventions. *Proceedings of the* 2021 *ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, 3-10 March 2021, 353-362. https://doi.org/10.1145/3442188.3445899

[12] Verma, S., Boonsanong, V., Hoang, M., *et al.* (2020) Counterfactual Explanations for Machine Learning: A Review. arXiv: 2010.10596.

[13] Ustun, B., Spangher, A. and Liu, Y. (2019) Actionable Recourse in Linear Classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*,

Atlanta, 29-31 January 2019, 10-19. https://doi.org/10.1145/3287560.3287566

[14] Pawelczyk, M., Broelemann, K. and Kasneci, G. (2021) Learning Model-Agnostic Counterfactual Explanations for Tabular Data. *Proceedings of The Web Conference* 2020, 20-24 April 2020, 3126-3132. https://doi.org/10.1145/3366423.3380087

[15] Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T. and Flach, P. (2020) FACE: Feasible and Actionable Counterfactual Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, 7-9 February 2020, 344-350. https://doi.org/10.1145/3375627.3375850

[16] Mahajan, D., Tan, C. and Sharma, A. (2019) Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. arXiv: 1912.03277.

[17] Palpandi, S.B., Palanigurupackiam, N., Almatar, H., Alduhayan, R., Alsomaie, B. and Almazroa, A. (2026) Artificial Intelligence Approaches for Schizophrenia Prediction and Its Biomarkers Using Medical Imaging Data. *Frontiers in Psychiatry*, **17**. https://doi.org/10.3389/fpsyt.2026.1821091

[18] Shatte, A.B.R., Hutchinson, D.M. and Teague, S.J. (2019) Machine Learning in Mental Health: A Scoping Review of Methods and Applications. *Psychological Medicine*, **49**, 1426-1448. https://doi.org/10.1017/s0033291719000151

[19] Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y. and Van Calster, B. (2019) A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology*, **110**, 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004

[20] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al.* (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[21] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why Should I Trust You? *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[22] Hamilton, M. (1960) A Rating Scale for Depression. *Journal of Neurology, Neurosurgery & Psychiatry*, **23**, 56-62. https://doi.org/10.1136/jnnp.23.1.56

[23] Young, R.C., Biggs, J.T., Ziegler, V.E. and Meyer, D.A. (1978) A Rating Scale for Mania: Reliability, Validity and Sensitivity. *British Journal of Psychiatry*, **133**, 429-435. https://doi.org/10.1192/bjp.133.5.429

[24] Endicott, J. (1976) The Global Assessment Scale. *Archives of General Psychiatry*, **33**, 766-771. https://doi.org/10.1001/archpsyc.1976.01770060086012

[25] Tohen, M., Frank, E., Bowden, C.L., Colom, F., Ghaemi, S.N., Yatham, L.N., Malhi, G.S., *et al.* (2015) The International Society for Bipolar Disorders (ISBD) Task Force Report on the Nomenclature of Course and Outcome in Bipolar Disorders. *Bipolar Disorders*, **11**, 453-473.

[26] Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., *et al.* (2004) The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): A Psychometric Evaluation in Patients with Chronic Major Depression. *Biological Psychiatry*, **54**, 573-583. https://doi.org/10.1016/s0006-3223(02)01866-8

[27] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., *et al.* (2016) MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, **3**, Article 160035. https://doi.org/10.1038/sdata.2016.35

[28] Maudsley, R. (2018) The Maudsley Prescribing Guidelines in Psychiatry. 13th Edi-

tion, Wiley-Blackwell.

[29] Elbayad, M., Besacier, L. and Verbeek, J. (2018). Pervasive Attention: 2. Proceedings of the 22nd Conference on Computational Natural Language Learning, Brussels, 31 October-1 November 2018, 97-107. https://doi.org/10.18653/v1/k18-1010

[30] Lea, C., Flynn, M.D., Vidal, R., Reiter, A. and Hager, G.D. (2017) Temporal Convolutional Networks for Action Segmentation and Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1003-1012. https://doi.org/10.1109/cvpr.2017.113

[31] Liu, J., Zi, L., Shreyas, P., Dustin, T., Tania, B.W. and Balaji, L. (2020) Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. *Advances in Neural Information Processing Systems*, 33, 7498-7512.

[32] Li, Y.Z. and Yarin, G. (2017) Dropout Inference in Bayesian Neural Networks with Alpha-Divergences. *International Conference on Machine Learning*, 2052-2061.

[33] Niculescu-Mizil, A. and Caruana, R. (2005) Predicting Good Probabilities with Supervised Learning. Proceedings of the 22*nd International Conference on Machine learning-ICML* '05, Bonn, 7-11 August 2005, 625-633. https://doi.org/10.1145/1102351.1102430

[34] Russell, C. (2019) Efficient Search for Diverse Coherent Explanations. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, 29-31 January 2019, 20-28. https://doi.org/10.1145/3287560.3287569

[35] Schölkopf, B. (2022) Causality for Machine Learning. *Probabilistic and Causal Inference*, 765-804. https://doi.org/10.1145/3501714.3501755

[36] Huang, W.B., Tong, Z., Yu, R. and Huang, J.Z. (2018) Adaptive Sampling towards Fast Graph Representation Learning. *Advances in Neural Information Processing Systems*, 31.

[37] Van Looveren, A. and Klaise, J. (2021) Interpretable Counterfactual Explanations Guided by Prototypes. In: Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J. and Lozano, J.A., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 650-665. https://doi.org/10.1007/978-3-030-86520-7_40

[38] Keane, M.T. and Smyth, B. (2020) Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI) In: Watson, I. and Weber, R., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 163-178. https://doi.org/10.1007/978-3-030-58342-2_11

[39] Vilone, G. and Longo, L. (2021) Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence. *Information Fusion*, **76**, 89-106. https://doi.org/10.1016/j.inffus.2021.05.009

[40] Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I. (2020) Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Medical Informatics and Decision Making*, **20**, Article No. 310. https://doi.org/10.1186/s12911-020-01332-6

[41] Ma, Y., Tu, X., Luo, X., Hu, L. and Wang, C. (2025) Machine-Learning-Based Cost Prediction Models for Inpatients with Mental Disorders in China. *BMC Psychiatry*, **25**, Article No. 33. https://doi.org/10.1186/s12888-024-06358-y

[42] Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) AI in Health and Medicine. *Nature Medicine*, **28**, 31-38. https://doi.org/10.1038/s41591-021-01614-0

[43] World Health Organization (2021) Ethics and Governance of Artificial Intelligence for Health. WHO Press.

[44] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

https://doi.org/10.1023/a:1010933404324

[45] Nalluri, M., Mounika, P. and Nageswara, R.E. (2020) A Scalable Tree Boosting System: XG Boost. *International Journal of Research Studies in Science, Engineering and Technology*, **7**, 36-51.