



Evaluating ChatGPT-generated psychoeducation for mood disorders: comparative insights from patients and mental health professionals

Francesco Attanasio^{a,b,*}, Valentina Fazio^c, Corinna Antonini^a, Nicola Lanzano^a,
Giulia Obumselu^a, Michele Prato^a, Emma Flutti^{a,b}, Federico Pacchioni^a,
Lorenzo Fregna^{b,c}, Linda Anna Marina Franchini^c, Cristina Colombo^{a,c}

^a Vita-Salute San Raffaele University, Milan, Italy

^b Department of Human Neuroscience, Sapienza University of Rome, Rome, Italy

^c Department of Clinical Neurosciences, Mood Disorder Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy

ARTICLE INFO

Keywords:

Psychoeducation
Mood disorders
Depression
AI
Digital mental health
ChatGPT
Large language models

ABSTRACT

Introduction: Psychoeducation is a key intervention in mood disorders. With the rise of artificial intelligence (AI) conversational agents, tools like ChatGPT are increasingly consulted by patients. Yet, empirical data on how AI-generated psychoeducational content is perceived by patients and professionals remain limited.

Methods: In this cross-sectional study, 30 depressed inpatients submitted five open-ended questions to ChatGPT-4o. Responses were rated by patients using 5-point Likert scales for relevance, comprehensibility, usefulness, empathy, and acceptance. Independent safety checks were applied to all outputs. The same responses were later blindly evaluated, in randomized order, by three psychiatrists and three psychiatric rehabilitation technicians (PRTs).

Results: All outputs passed safety review. Patients assigned higher total scores (mean $\pm$ SD = 22.43 $\pm$ 2.64) than PRTs (17.63 $\pm$ 3.39) and psychiatrists (15.42 $\pm$ 2.02). The largest gaps involved empathy and acceptance, whereas relevance, usefulness, and comprehensibility differed less. PRT ratings were intermediate: closer to patients on relevance, comprehensibility, and usefulness, but closer to psychiatrists on empathy and acceptance. Within patients, no associations emerged with age, education, depression severity, or prior psychoeducation.

Conclusions: Patients with mood disorders perceived ChatGPT-generated responses as more relevant, comprehensible, useful, accepting, and empathetic than health professionals did. With conversational agents entering psychoeducation, clinicians must develop strategies to critically integrate such tools, ensuring safety and quality while guiding patient use. The challenge is not resisting AI adoption, but framing it within safe, effective, and ethically sound psychoeducational care.

1. Introduction

Psychoeducation (PE) is an evidence-based adjunct in the treatment of psychiatric disorders, providing structured information to patients and their caregivers about the nature, course, and management of their illness with the goal of improving adherence, self-management, and long-term outcomes (Xia et al., 2011; I. R. H. Falloon, 2003; Rabelo et al., 2021).

Early conceptual foundations of PE emerged in the late 20th century with seminal work by Guerney (Guerney et al., 1971), Anderson (Anderson et al., 1980), and Falloon (I. R. Falloon et al., 1981) highlighting the benefits of involving family members in care, improving

quality of life and personal satisfaction, reduced hospitalization rates and better clinical outcomes of subjects with schizophrenia.

Subsequent research extended PE to mood disorders, with robust evidence in bipolar disorder showing reductions in relapse and recurrence, improved psychosocial functioning, and longer time to relapse, and growing support for its adjunctive role in major depressive disorder (Bond and Anderson, 2015; Prasko et al., 2013; Tursi et al., 2013; Rabelo et al., 2021; Levrat et al., 2023; Katsuki et al., 2022).

With the proliferation of digital health tools, there has been increasing interest in delivering PE via scalable, accessible formats such as mobile applications and conversational agents. These digital interventions have shown promise in augmenting traditional care for

* Corresponding author. Vita-Salute San Raffaele University, Milan, Italy.

E-mail addresses: attanasio.francesco@hsr.it, francesco.attanasio@uniroma1.it (F. Attanasio).

<https://doi.org/10.1016/j.jpsychires.2025.11.039>

Received 2 September 2025; Received in revised form 10 November 2025; Accepted 30 November 2025

Available online 1 December 2025

0022-3956/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

mental disorders, yet their designs are heterogeneous, long-term engagement is uncertain, and implementation barriers persist (de Azevedo et al., 2024; Moriya et al., 2023; Ho et al., 2025; Torous et al., 2025; Philippe et al., 2022).

Large language models (LLMs), including ChatGPT (OpenAI, 2025), represent a novel class of conversational agents capable of generating fluent, context-aware natural language responses. Their widespread availability and human-like interactivity position them as potential platforms for delivering PE content. However, their outputs vary in factual reliability, can include misleading “hallucinated” information, and lack genuine empathic understanding, raising concerns about clinical appropriateness, user trust, and differential reception by lay versus expert evaluators (Sorin et al., 2024; Haider et al., 2025; Lee et al., 2024; Li et al., 2023; Guo et al., 2024).

While conversational artificial intelligence (AI) and digital mental health agents have been explored in various capacities, to our knowledge, no published study has specifically evaluated ChatGPT as a PE tool for mood disorders while directly comparing how patients and mental health professionals perceive the same content.

1.1. Study objectives

This study sought to address that gap by empirically evaluating ChatGPT (GPT-4o) as a source of PE information for individuals with mood disorders, comparing its reception among patients, psychiatrists, and psychiatric rehabilitation technicians (PRTs). Specifically, we asked:

1. To assess how ChatGPT's responses are evaluated in terms of perceived relevance, comprehensibility, usefulness, empathy, and acceptance.
2. To determine whether, and to what extent, these evaluations differ between patients, psychiatrists, and PRTs.
3. To explore whether specific demographic and clinical characteristics of patients (namely age, years of education, severity of depressive symptoms, and prior participation in PE programs) are associated with their evaluations of ChatGPT's responses.

1.2. Hypotheses

H1. Psychiatrists and PRTs will rate ChatGPT's responses significantly lower than patients across most evaluation dimensions, with no difference between the two professional groups.

AI-based conversational agents have generally shown high acceptability among users seeking mental health information, with studies reporting high usability, perceived helpfulness, and a sense of being listened to (Aggarwal et al., 2023; Wutz et al., 2023; Li et al., 2023). In contrast, clinicians tend to be more cautious, expressing concerns about accuracy, safety, and accountability (Blease et al., 2024; Hipgrave et al., 2025; Vaidyam et al., 2021; Huo et al., 2025). Moreover, PE has historically positioned clinicians as authoritative guarantors of informational quality, reinforcing both users' expectations of rigor and professionals' responsibility for content accuracy (Wilson et al., 2018). Consequently, professionals are expected to appraise AI-delivered PE more critically than patients.

H2. Within the patient group, relevance and comprehensibility will receive higher ratings than empathy, acceptance and utility.

This hypothesis is based on evidence that AI-based conversational agents typically adopt a highly accommodating and emotionally congruent tone, even in the absence of genuine affective understanding (Sorin et al., 2024; Lee et al., 2024; Guo et al., 2024). At the same time, their safety policies systematically prevent them from providing specific clinical recommendations, prompting frequent disclaimers and referrals to medical professionals (Haider et al., 2025; Li et al., 2023; Huo et al.,

2025). This cautious communication style, though ethically appropriate, may result in content that users perceive as less directive or practically useful.

H3. Among patients, greater depressive symptom severity will be associated with lower overall ratings of ChatGPT's responses.

This hypothesis is grounded in evidence that depressive symptoms are associated with cognitive dysfunctions and impaired emotional responsiveness that may reduce engagement and positive evaluation of educational or supportive content (Beck et al., 2024).

2. Methods

2.1. Study design

This investigation was conducted as a single-center, cross-sectional, observational study.

2.2. Participants

We consecutively recruited 30 inpatients admitted to the Mood Disorder Unit at San Raffaele Hospital, Milan, who had been clinically referred to participate in an in-unit PE group and had no contraindications to this intervention. Recruitment took place between December 2024 and May 2025. Eligible participants were adults aged 18–75 years who met Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR, American Psychiatric Association, 2025) criteria for either Major Depressive Disorder or Bipolar Disorder, current depressive episode. Written informed consent was obtained from all participants prior to any study procedures. The in-unit PE group included patients regardless of whether their diagnosis was Major Depressive Disorder or Bipolar Disorder. Accordingly, in this study, they were recruited and analyzed as a single group, given that all were experiencing a current depressive episode and thus shared similar depressive phenomenology. Exclusion criteria were: diagnosis of a schizophrenia spectrum or other primary psychotic disorder, intellectual disability, neurocognitive disorder (e.g., dementia), acute substance intoxication or withdrawal, severe neurological disease, or any condition that would preclude reliable participation (e.g., significant language or sensory barriers). The study was approved by the Hospital's Ethical Committee and conducted in accordance with the latest version of the Declaration of Helsinki.

2.3. Procedure

Two trained psychiatrists collected sociodemographic and clinical data from each participant, including age, sex, years of education, prior exposure to PE interventions, and severity of depressive symptoms assessed with the 17-item Hamilton Depression Rating Scale (HDRS) (Hamilton, 1967).

Each participant was instructed to formulate five open-ended questions about their psychiatric condition, coping strategies, or illness-related concerns, phrasing them as if addressing a mental health professional unfamiliar with their history (see **Supplementary Materials** for the exact wording). Questions were submitted in Italian to ChatGPT (GPT-4o, base model, no system prompt modifications or external plugins) via the web interface (OpenAI, 2025). For each patient a single session was used, so that all five questions from that patient were entered within the same thread, allowing the model to retain contextual information across the questions. The exact prompt template is reported in **Supplementary Materials**, and all raw AI outputs were archived for reproducibility.

All ChatGPT-4o-generated responses underwent an independent safety and appropriateness review by a psychiatrist not involved in the evaluation phase. The reviewer applied a predefined checklist (**Supplementary Materials**) to screen for potentially harmful, misleading, or

anxiety-provoking content (e.g., encouragement of self-harm, diagnostic overreach, stigmatizing language, or exaggerated alarm). Responses failing to meet all safety criteria were regenerated once using the same prompt template; if unsafe elements persisted after a second attempt, the response would have been excluded with rationale documented. Approved responses were then provided to the originating patient for evaluation.

Participants rated each response on five separate items using a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree): relevance (extent to which the response addressed the original question), comprehensibility (clarity and ease of understanding), usefulness (perceived clinical utility), empathy (perceived emotional attunement), and acceptance (sense of being comforted and not judged). Each item was accompanied by explicit definitions on the evaluation form (see **Supplementary Materials**).

To minimise expectancy effects, patients were informed that the psychoeducational answers they would receive were produced by an *experimental, clinician-supervised digital tool*, without specifying whether the content originated from a human expert or an AI system. They were assured that every response had been independently reviewed by a psychiatrist for safety and accuracy. Full disclosure that the tool was ChatGPT-4o and that all answers were AI-generated was provided during a debriefing session immediately after completion of the rating task. This procedure was necessary to preserve the integrity of the evaluative comparisons. In order to reduce potential social desirability bias and evaluation apprehension, all participants were assigned a unique anonymized code. They were explicitly informed that their responses would be processed exclusively in fully anonymous form (Nederhof, 1985).

Subsequently, the same AI-generated responses were independently evaluated in a blinded fashion by three psychiatrists and three PRTs. Professional evaluators were blind to the source of the responses (i.e., they were not informed that content was produced by ChatGPT-4o) and had no access to patient ratings. The order of presentation of responses was randomized for each evaluator to mitigate order effects. Prior to rating, professionals received brief instructions about the study's general aim (evaluation of PE content) and the operational definitions of the five items; no additional training or calibration beyond the provided definitions was conducted to preserve independent professional judgment. The arithmetic mean of the three raters' scores yielded a single professional rating per item and response.

2.4. Data analysis

Group differences in total and domain-level scores were examined using repeated-measures ANOVA with evaluator group (patients, PRTs, psychiatrists) as a within-subject factor. Mauchly's test was used to assess sphericity, and Greenhouse-Geisser corrections were applied when violated. Bonferroni-adjusted pairwise comparisons followed significant omnibus effects. Effect sizes are reported as partial η^2 for omnibus tests and Cohen's d for pairwise contrasts. Associations between patient total scores and age, education, and HDRS were explored separately using Pearson correlations and multiple linear regression, as these covariates were only available for the patient group. All analyses were conducted using JASP version 0.18.3 Intel ("JASP Team (2024). JASP (Version 0.95.0)[Computer Software]," n. d.).

3. Results

All 30 enrolled participants completed the study; there were no dropouts, adverse events, or procedural issues, and none of the ChatGPT-4o-generated responses failed the independent safety review (i.e., no response was excluded or substantially modified). Clinical and sociodemographic characteristics of the sample are reported in **Table 1**.

Group-level descriptive means and variability for each evaluation dimension and the aggregated total score (we aggregated each patient's five question-level ratings into a single total score, range 5–25) are

Table 1

Demographic and clinical characteristics of the patient sample.

Variable	
Age, years (mean, SD)	50.4, 12.4
Education years (mean, SD)	13.1, 3.3
HDRS total score (mean, SD)	7.4, 3.9
Prior psychoeducational interventions (Yes, %)	16, 53.33 %
Gender (N female, %)	15, 50 %

presented in **Table 2**.

The aggregated total scores were compared across evaluator groups (patients, PRTs, and psychiatrists) using a one-way repeated-measures ANOVA. Mauchly's test indicated that the sphericity assumption was met for total scores ($p = 0.131$). There was a significant main effect of the evaluator group on total ratings, $F(2,58) = 44.30$, $p < 0.001$, partial $\eta^2 = 0.60$. Bonferroni-adjusted pairwise comparisons showed that patients rated responses significantly higher than PRTs and psychiatrists, and that PRTs rated responses higher than psychiatrists (**Table 3**).

Potential confounding variables (age, years of education, depressive severity as measured by the HDRS, and prior participation in PE interventions [yes/no]) were available only for the patient group; therefore, the primary between-group comparisons are presented unadjusted, and patient-specific characteristics were examined separately. Distributions of age, education, and HDRS scores did not deviate significantly from normality according to the Shapiro-Wilk test (all $p > 0.05$); accordingly, Pearson correlations were used to assess univariate relationships with the total score. Total score was not significantly correlated with age ($r = -0.115$, $p = 0.545$), years of education ($r = 0.053$, $p = 0.780$), or HDRS ($r = 0.014$, $p = 0.941$). The covariates themselves were also weakly interrelated, with only a non-significant trend between HDRS and education ($r = 0.324$, $p = 0.080$), and no appreciable correlation between age and either education ($r = 0.006$, $p = 0.974$) or HDRS ($r = 0.190$, $p = 0.314$). For the binary variable (prior PE) a point-biserial correlation likewise showed no association with total score ($r = -0.153$, $p = 0.420$). A multiple linear regression including age, years of education, HDRS and prior PE simultaneously as predictors of the total patient score confirmed minimal explanatory power: the model accounted for negligible variance ($R^2 = 0.047$, adjusted $R^2 = -0.106$) and was not statistically significant, $F(4,25) = 0.307$, $p = 0.870$. None of the predictors contributed meaningfully (standardized β : age = -0.147 , $p = 0.473$; HDRS = -0.013 , $p = 0.951$; education = 0.047 , $p = 0.823$; prior PE unstandardized $\beta = -0.938$, $p = 0.381$). These findings provide no evidence that age, education, or symptom severity confound the primary patient ratings in this sample.

To assess whether the overall group differences extended to specific response qualities, separate repeated-measures ANOVA were conducted for each of the five dimensions: relevance, comprehensibility, usefulness, empathy, and acceptance (**Table 4**, **Fig. 1**; **Table 5** post-hoc tests).

4. Discussion

Interest in the application of AI, particularly conversational agents

Table 2

Descriptive ratings assigned to ChatGPT-4o responses by patients, PRTs, and psychiatrists, for each individual evaluation dimension and for the aggregated total score. Higher values indicate more favorable evaluations. Total score is the sum of the five item-level ratings (range 5–25).

Dimension Score, mean (SD)	Patients	PRTs	Psychiatrists
Relevance	4.67 (0.48)	4.20 (0.92)	3.27 (0.60)
Comprehensibility	4.70 (0.53)	4.13 (0.94)	3.57 (0.54)
Usefulness	4.40 (0.81)	4.17 (0.91)	3.17 (0.62)
Empathy	4.33 (0.79)	2.78 (0.85)	2.68 (0.43)
Acceptance	4.40 (0.81)	2.63 (0.96)	2.73 (0.31)
Total score (sum of five 1–5 items, range 5–25)	22.43 (2.64)	17.63 (3.39)	15.42 (2.02)

Table 3
Post Hoc Comparisons
Bonferroni-adjusted pairwise comparisons of total response ratings across evaluator groups (A = patients; B = psychiatric rehabilitation technicians; C = psychiatrists).

			95 % CI for Mean Difference					
Mean Difference			Lower	Upper	SE	t	Cohen's d	p (Bonf.)
A	B	4.800	2.921	6.679	0.762	6.298	1.753	<0.001
A	C	7.017	5.138	8.896	0.762	9.207	2.562	<0.001
B	C	2.217	0.338	4.096	0.762	2.909	0.809	0.015

Table 4
Repeated-measures ANOVA results for the total score and each individual evaluation dimension. For dimensions with violated sphericity, Greenhouse-Geisser correction was applied.

Dimension	F (df)	p	Partial η^2	ϵ (Greenhouse-Geisser)
Total score	44.30 (2,58)	<0.001	0.60	–
Relevance	33.35 (2,58)	<0.001	0.54	–
Comprehensibility	23.54 (1.645, 47.692)	<0.001	0.45	0.822
Usefulness	20.05 (1.599, 46.372)	<0.001	0.41	0.800
Empathy	47.95 (2,58)	<0.001	0.62	–
Acceptance	42.58 (1.559, 45.208)	<0.001	0.60	0.779

based on LLMs, to PE is rapidly growing. This is primarily driven by two major considerations. On one hand, these tools offer the potential to expand and enhance the scalability and efficiency of PE interventions that would otherwise require significant human and logistical resources. The ability to reach patients in home-based settings, ensure continuity and immediacy of communication, and adapt to individual needs makes AI particularly appealing from a public health perspective. On the other hand, and perhaps more urgently in the present context, users, including psychiatric patients, are already spontaneously using tools such as ChatGPT to seek information about their mental health. This raises important questions regarding the quality, accuracy, and safety of the information provided, and how such responses are received and interpreted by patients.

Despite the increasing focus on digital mental health and the use of LLMs in educational or therapeutic contexts, there remains a significant lack of empirical studies directly comparing how patients and mental health professionals perceive PE content generated by AI systems. More broadly, very few studies on generative AI include standardized tools to assess response quality, blinded evaluations and provide clear descriptions of prompt construction, making most of the available literature poorly reproducible and difficult to compare across settings (Huo et al., 2025). This exploratory study, designed to be highly reproducible, aimed to contribute to this knowledge gap by investigating both the communicative potential of AI-generated PE responses and the differences in evaluation expressed by patients with mood disorders and mental health professionals.

Three key findings emerged and warrant particular discussion. First, as expected (see H1) patients rated ChatGPT's responses significantly more favorably than healthcare professionals. This discrepancy was most pronounced in empathy and acceptance, suggesting that the linguistic style adopted by AI systems may effectively resonate with patients' need for emotional support and reassurance, even in the absence of a genuine therapeutic relationship. However, contrary to our hypothesis (see H2), patients did not differentiate between the rated dimensions.

Second, contrary to the authors' initial expectations (see H1), PRTs gave intermediate ratings, showing closer alignment with patients on relevance, comprehensibility, usefulness, while more closely resembling psychiatrists in their lower scores for empathy and acceptance. This pattern may reflect their role in clinical practice, often mediating

between patients and physicians, which might sensitize them to both the informational validity and the emotional limitations of AI-generated content.

Third, contrary to our hypothesis (see H3) none of the patient demographic or clinical variables examined were significantly associated with their evaluations of ChatGPT's responses. This unexpected finding may suggest a generally favorable reception of the AI-generated content across broad individual characteristics. At the same time, the lack of association with prior experience of PE interventions may point to a limited critical capacity among patients in assessing PE content, highlighting the need for dedicated educational efforts to enhance users' ability to critically evaluate digital PE material, particularly when produced by automated sources.

A central aspect in interpreting these results lies in understanding why patients perceive ChatGPT as empathetic. It is crucial to clarify that the empathy provided by ChatGPT and similar systems is not based on emotional understanding but rather on linguistic simulation. These models are optimized to recognize common patterns in human dialogue and to produce responses that appear agreeable or comforting to users, without any intrinsic clinical or educational intent (Lahnala et al., 2022; Sorin et al., 2024). This simulated empathy may lead patients to misinterpret the interaction as deeply understanding or supportive, creating an illusion of therapeutic alliance that may lack factual accuracy or clinical appropriateness, thus undermining the efficacy of PE interventions if not critically contextualized.

In parallel, experts have raised serious concerns about the safety, quality, and reliability of automatically generated content. One of the most frequently reported risks is the phenomenon of "hallucination", plausible and convincing yet clinically incorrect or misleading responses (Guo et al., 2024). This risk is further exacerbated by the high level of trust that patients appear to place in AI-generated content, which could discourage help-seeking behavior or reduce engagement with professional care, especially in clinically sensitive or high-risk situations.

At the same time, patients may feel more comfortable turning to digital interlocutors, which are perceived as nonjudgmental and anonymous. This aspect, well-documented in qualitative research on patient-chatbot interactions, can promote greater openness in discussing needs, doubts, or sensitive topics that patients may feel too embarrassed or stigmatized to address with human professionals (Wutz et al., 2023; Galvão Gomes da Silva et al., 2018; Aggarwal et al., 2023). Such openness may, in turn, offer a valuable resource in the design of future digitally mediated PE interventions, if properly understood and harnessed.

Taken together, these findings have relevant clinical and operational implications, particularly if replicated in larger and more diverse samples. On one hand, there is a need to strengthen the educational role of clinicians by raising patients' awareness of the artificial and constrained nature of digital empathy. On the other hand, development efforts should aim toward more accurate and reliable AI models, possibly through hybrid approaches in which human professionals supervise, filter, or adapt automatically generated content. The authors argue that conversational AI tools will increasingly become part of the PE landscape, and that this inevitable integration calls for deliberate, evidence-based strategies for their safe and effective implementation in clinical practice.

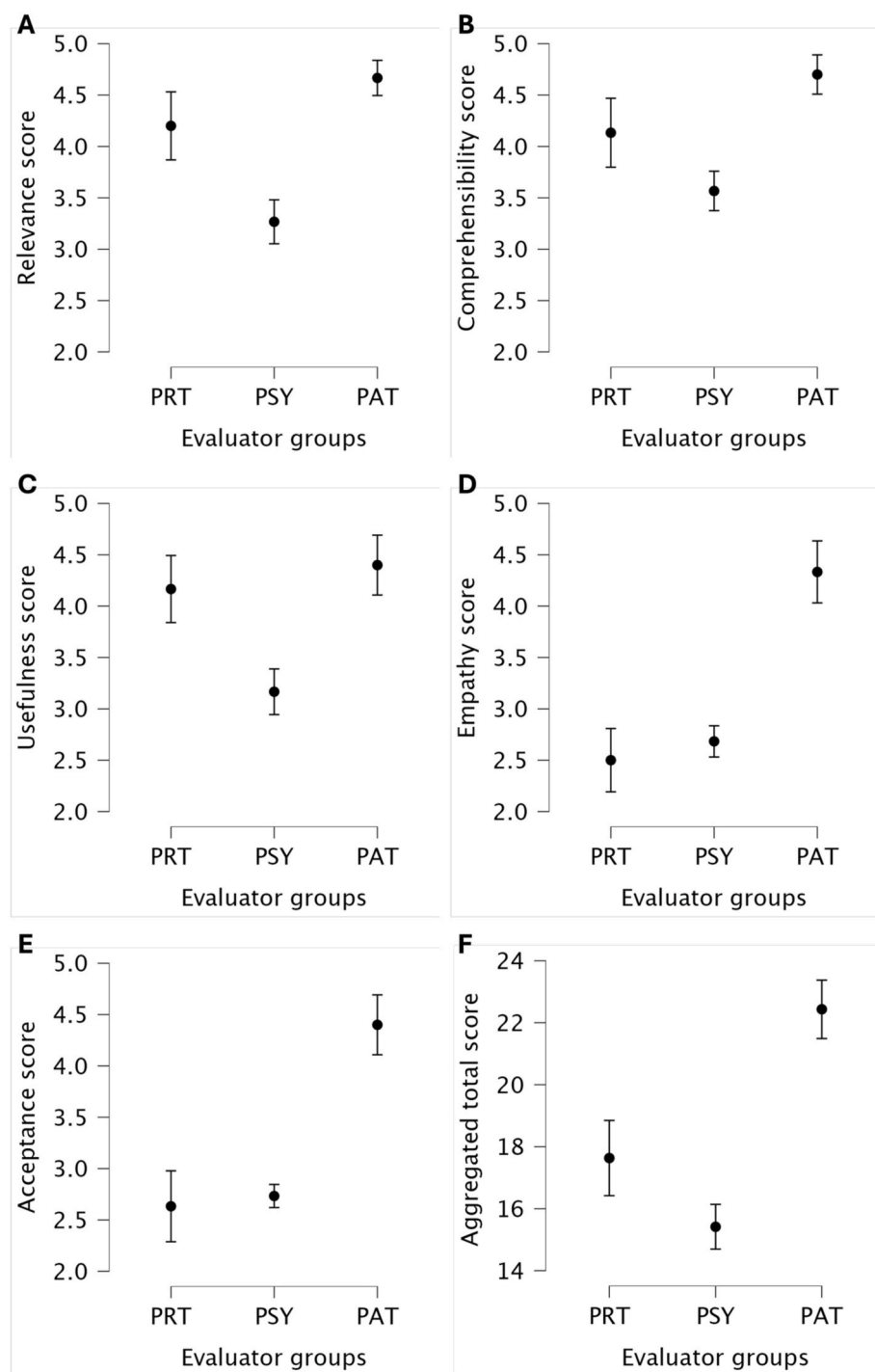


Fig. 1. Mean ratings ($\pm 95\%$ confidence intervals) assigned to ChatGPT-4o responses by psychiatric rehabilitation technicians (PRT), psychiatrists (PSY) and patients (PAT) for (A) relevance, (B) comprehensibility, (C) usefulness, (D) empathy, (E) acceptance, and (F) the aggregated total score.

4.1. Limitations

This study has several intrinsic limitations that should be acknowledged. First, as an initial exploratory investigation, our sample size was relatively small, limiting statistical power and the generalizability of findings; larger, multicenter trials will be necessary to confirm these preliminary results. Second, the research was conducted at a single psychiatric unit, raising the possibility of site-specific effects related to local clinical practice or patient demographics. Third, we focused exclusively on patients with mood disorders (major depressive disorder and bipolar depression), so our conclusions may not extend to other

psychiatric populations. Fourth, only one AI system (ChatGPT-4o) was evaluated. Fifth, our design captured only immediate ratings of AI-generated content and did not assess longer-term outcomes such as actual adherence to PE recommendations or clinical benefits. Sixth, reflecting the clinical composition of the in-unit PE group, we focused on patients experiencing a current depressive episode, without distinguishing between Major Depressive Disorder and Bipolar Disorder, as these patients shared similar depressive phenomenology. Seventh, no participant dropped out during the study procedures, which may reflect a highly compliant and motivated inpatient sample. While this strengthens the internal validity of the findings, it may also limit their

Table 5
Bonferroni-adjusted pairwise comparisons of evaluator group ratings (A = patients; B = psychiatric rehabilitation technicians; C = psychiatrists) for each individual dimension.

Dimension	Contrast	Mean Difference	95 % CI for Mean Difference	SE	t	p (Bonf.)	Cohen's d
Relevance	A vs B	0.467	[0.036, 0.897]	0.175	2.673	0.029	0.67
	A vs C	1.400	[0.970, 1.830]	0.175	8.020	<0.001	2.02
	B vs C	0.933	[0.503, 1.364]	0.175	5.347	<0.001	1.35
Comprehensibility	A vs B	0.567	[0.159, 0.974]	0.165	3.431	0.003	0.81
	A vs C	1.133	[0.726, 1.541]	0.165	6.862	<0.001	1.63
	B vs C	0.567	[0.159, 0.974]	0.165	3.431	0.003	0.81
Usefulness	A vs B	0.233	[−0.277, 0.743]	0.207	1.128	0.792	0.30
	A vs C	1.233	[0.723, 1.743]	0.207	5.961	<0.001	1.56
	B vs C	1.000	[0.490, 1.510]	0.207	4.833	<0.001	1.26
Empathy	A vs B	1.833	[1.325, 2.342]	0.206	8.891	<0.001	2.48
	B vs C	1.650	[1.142, 2.158]	0.206	8.002	<0.001	2.24
	B vs C	−0.183	[−0.692, 0.325]	0.206	−0.889	1.000	−0.25
Acceptance	A vs B	1.767	[1.236, 2.297]	0.215	8.214	<0.001	2.35
	B vs C	1.667	[1.136, 2.197]	0.215	7.749	<0.001	2.22
	B vs C	−0.100	[−0.630, 0.430]	0.215	−0.465	1.000	−0.13

generalizability to broader, less supervised clinical populations. Finally, all evaluations were based on subjective Likert-scale judgments, without objective measures of factual accuracy or behavioral change, and future work should integrate such metrics to provide a more comprehensive appraisal of AI-mediated PE.

5. Conclusions

Overall, the findings of this study suggest that patients with mood disorders tend to respond positively to PE content generated by ChatGPT, attributing to it a higher degree of empathy and acceptance than that expressed by mental health professionals. This discrepancy raises a crucial question for the future of PE: the growing accessibility of AI-based conversational agents is making it increasingly likely that patients will spontaneously turn to automated sources for sensitive mental health information. In this evolving landscape, it becomes essential to develop tools and strategies that not only enhance the reliability and quality of generated content, but also promote its responsible integration into clinical practice. Conversational agents are poised to become a stable component of future PE. The challenge is not to resist patients' spontaneous use of AI tools, but to understand how these technologies shape patient behavior and to guide their use toward becoming a truly complementary asset in PE care.

CRediT authorship contribution statement

Francesco Attanasio: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Valentina Fazio:** Writing – review & editing, Investigation, Formal analysis. **Corinna Antonini:** Writing – review & editing, Investigation. **Nicola Lanzano:** Writing – review & editing, Investigation. **Giulia Obumselu:** Writing – review & editing, Investigation. **Michele Prato:** Writing – review & editing, Investigation. **Emma Flutti:** Investigation. **Federico Pacchioni:** Conceptualization. **Lorenzo Fregna:** Investigation. **Linda Anna Marina Franchini:** Investigation. **Cristina Colombo:** Supervision, Project administration, Data curation.

Funding informations

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jpsychires.2025.11.039>.

References

Aggarwal, Abhishek, Tam, Cheuk Chi, Wu, Dezhi, Li, Xiaoming, Qiao, Shan, 2023. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. *J. Med. Internet Res.* 25 (February), e40789.

American Psychiatric Association, 2025. Diagnostic and statistical manual of mental disorders. In: Muskin, Philip R., Dickerman, Anna L., Drysdale, Andrew, Holderness, Claire C. (Eds.), *Text Revision (DSM-5-TR®)*, fifth ed. American Psychiatric Association Publishing, Arlington, TX.

Anderson, C.M., Hogarty, G.E., Reiss, D.J., 1980. Family treatment of adult schizophrenic patients: a psycho-educational approach. *Schizophr. Bull.* 6 (3), 490–505.

Beck, A.T., Rush, A.J., Shaw, B.F., Emery, G., DeRubeis, R.J., Hollon, S.D., 2024. *Cognitive Therapy of Depression*.

Blease, Charlotte, Worthen, Abigail, Torous, John, 2024. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. *Psychiatry Res.* 333 (115724), 115724.

Bond, Kirsten, Anderson, Ian M., 2015. Psychoeducation for relapse prevention in bipolar disorder: a systematic review of efficacy in randomized controlled trials. *Bipolar Disord.* 17 (4), 349–362.

de Azevedo, Cardoso, Taiane, Kochhar, Shruti, Torous, John, Morton, Emma, 2024. Digital tools to facilitate the detection and treatment of bipolar disorder: key developments and future directions. *JMIR Ment. Health* 11 (April), e58631.

Falloon, Ian R.H., 2003. Family interventions for mental disorders: efficacy and effectiveness. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)* 2 (1), 20–28.

Falloon, I.R., Liberman, R.P., Lillie, F.J., Vaughn, C.E., 1981. Family therapy of schizophrenics with high risk of relapse. *Fam. Process* 20 (2), 211–221.

Galvão Gomes da Silva, Joana, Kavanagh, David, J., Belpaeme, Tony, Taylor, Lloyd, Beeson, Konna, Andrade, Jackie, 2018. Experiences of a motivational interview delivered by a robot: qualitative study. *J. Med. Internet Res.* 20 (5), e116.

Guerney, Bernard, Stollak, Gary, Guerney, Louise, 1971. The practicing psychologist as educator: an alternative to the medical practitioner model. *Prof. Psychol.* 2 (3), 276–282.

Guo, Zhijun, Lai, Alvina, Thygesen, Johan H., Farrington, Joseph, Keen, Thomas, Li, Kezhi, 2024. Large language models for mental health applications: systematic review. *JMIR Ment. Health* 11 (October), e57400.

Haider, Syed Ali, Prabha, Srinivasagam, Gomez-Cabello, Cesar Abraham, Borna, Sahar, Genovese, Ariana, Trabilsy, Maissa, Collaco, Bernardo G., et al., 2025. Synthetic patient-physician conversations simulated by large language models: a multi-dimensional evaluation. *Sensors (Basel, Switzerland)* 25 (14), 4305.

Hamilton, M., 1967. Development of a rating scale for primary depressive illness. *Br. J. Soc. Clin. Psychol.* 6 (4), 278–296.

- Hipgrave, Lyndsey, Goldie, Jessie, Dennis, Simon, Coleman, Amanda, 2025. Balancing risks and benefits: clinicians' perspectives on the use of generative AI chatbots in mental healthcare. *Front. Digit. Health* 7 (May), 1606291.
- Ho, Thi Quynh Anh, Khanh-Dao Le, Long, Engel, Lidia, Ngoc, Le, Melvin, Glenn, Le, Ha N.D., Mihalopoulos, Cathrine, 2025. Barriers to and facilitators of user engagement with web-based mental health interventions in young people: a systematic review. *Eur. Child Adolesc. Psychiatr.* 34 (1), 83–100.
- Huo, Bright, Boyle, Amy, Marfo, Nana, Tangamornsuksan, Wimonchat, Steen, Jeremy P., Tyler, McKechnie, Lee, Yung, et al., 2025. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw. Open* 8 (2), e2457879.
- JASP Team, 2024. JASP (Version 0.95.0)[Computer Software] n.d.
- Katsuki, Fujika, Watanabe, Norio, Yamada, Atsurou, Hasegawa, Takaaki, 2022. Effectiveness of family psychoeducation for major depressive disorder: systematic review and meta-analysis. *BJPsych Open* 8 (5), e148.
- Lahnala, Allison, Welch, Charles, Jurgens, David, Flek, Lucie, 2022. A critical reflection and forward perspective on empathy and natural language processing. In: Goldberg, Yoav, Kozareva, Zornitsa, Zhang, Yue (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 2139–2158.
- Lee, Hyein S., Wright, Colton, Ferranto, Julia, Buttner, Jessica, Palmer, Clare E., Welchman, Andrew, Mazor, Kathleen M., et al., 2024. Artificial intelligence conversational agents in mental health: patients see potential, but prefer humans in the loop. *Front. Psychiatr.* 15, 1505024.
- Levrat, Vanessa, Favre, Sophie, Richard-Lepouriel, Hélène, 2023. Current practices of psychoeducation interventions with persons with bipolar disorders: a literature review. *Front. Psychiatr.* 14, 1320654.
- Li, Han, Zhang, Renwen, Lee, Yi-Chieh, Kraut, Robert E., Mohr, David C., 2023. Systematic review and meta-analysis of AI-Based conversational agents for promoting mental health and well-being. *npj Digit. Med.* 6 (1), 236.
- Moriya, Renato Mikio, Ragassi Urbano, Mariana, Vargas, Heber Odebrecht, Machado, Regina Celia Bueno, Zazula, Robson, Novaes de Oliveira Roldan, Ana Cecilia, Nunes, Sandra Odebrecht Vargas, 2023. Digital mental health interventions for anxiety and mood disorders patients: a 24-Week Follow-Up. *Clinical eHealth* 6 (December), 114–120.
- Nederhof, Anton J., 1985. Methods of coping with social desirability bias: a review. *Eur. J. Soc. Psychol.* 15 (3), 263–280.
- OpenAI. (2025). ChatGPT. <https://chat.openai.com/>.
- Philippe, Tristan J., Sikder, Naureen, Jackson, Anna, Koblanski, Maya E., Liow, Eric, Pilarinos, Andreas, Vasarhelyi, Krisztina, 2022. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. *JMIR Ment. Health* 9 (5), e35159.
- Prasko, Jan, Ociskova, Marie, Kamaradova, Dana, Sedlackova, Zuzana, Cerna, Monika, Mainerova, Barbora, Sandoval, Aneta, 2013. Bipolar affective disorder and psychoeducation. *Neuroendocrinol. Lett.* 34 (2), 83–96.
- Rabelo, Juliana Lemos, Cruz, Breno Fiuza, Ferreira, Jéssica Diniz Rodrigues, Viana, Bernardo de Mattos, Barbosa, Izabela Guimarães, 2021. Psychoeducation in bipolar disorder: a systematic review. *World J. Psychiatr.* 11 (12), 1407–1424.
- Sorin, Vera, Brin, Dana, Barash, Yiftach, Konen, Eli, Charney, Alexander, Nadkarni, Girish, Klang, Eyal, 2024. Large language models and empathy: systematic review. *J. Med. Internet Res.* 26 (December), e52597.
- Torous, John, Jake, Linardon, Goldberg, Simon B., Sun, Shufang, Imogen, Bell, Nicholas, Jennifer, Hassan, Lamie, Yining, Hua, Milton, Alyssa, Firth, Joseph, 2025. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)* 24 (2), 156–174.
- Tursi, Mariana Flávia de Souza, Baes, Cristiane von Werne, Camacho, Fabio Ribeiro de Barros, de Carvalho Tofoli, Sandra Marcia, Jurueña, Mario Francisco, 2013. Effectiveness of psychoeducation for depression: a systematic review. *Aust. N. Z. J. Psychiatr.* 47 (11), 1019–1031.
- Vaidyam, Aditya Nrusimha, Linggongoro, Danny, Torous, John, 2021. Changes to the Psychiatric Chatbot Landscape: a Systematic Review of Conversational Agents in Serious Mental Illness: Changements Du Paysage Psychiatrique Des Chatbots: une Revue Systématique Des Agents Conversationnels Dans La Maladie Mentale Sérieuse. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie* 66 (4), 339–348.
- Wilson, Lynere, Crowe, Marie, Scott, Anne, Cameron, Lacey, 2018. Psychoeducation for bipolar disorder: a discourse analysis. *Int. J. Ment. Health Nurs.* 27 (1), 349–357.
- Wutz, Maximilian, Hermes, Marius, Winter, Vera, Köberlein-Neu, Juliane, 2023. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. *J. Med. Internet Res.* 25 (September), e46548.
- Xia, Jun, Merinder, Lars Bertil, Belgamwar, Madhvi R., 2011. Psychoeducation for schizophrenia. *Cochrane Database Syst. Rev.* 2013 (6), CD002831.