Unlocking Empowerment: An Empirical Study on the Impact of Explainable AI in Mental Health Apps

Sven Bottesch
University of Ulm
sven.bottesch@uni-ulm.de

Yannik Terhorst
Ludwig Maximilian
University of Munich
yannik.terhorst@psy.lmu.de

Maximilian Förster
University of Ulm
maximilian.foerster@uni-ulm.de

Abstract

It is anticipated that apps based on artificial intelligence (AI) will be instrumental in mitigating the global shortage in mental healthcare. One important purpose of such apps is to encourage users' self-help. This study is dedicated to examining the potential role of explainable AI (XAI) for mental health apps. We build on mental health literature to conceptualize potential effects of explanations in terms of patient empowerment. We implement an online experiment with a fully instantiated mental health app based on a real-world dataset. The randomized between-subject experiment is conducted with 409 participants to test the effectiveness of feature importance and counterfactual explanations on patient empowerment, intention to use, and intention to act. Our results show that the provision of counterfactual explanations alongside AI-generated predictions of depression risk in a mental health app can significantly increase users' intention to use and empowerment.

Keywords: explainable artificial intelligence (XAI), mental health, smart sensing, digital phenotyping, online experiment

1. Introduction

Artificial intelligence (AI) holds significant potential to help address the increasing burden of mental health-related issues. Globally, more than one billion people are estimated to live with mental disorders (Global Burden of Disease Collaborative Network, 2024), and many suffer from insufficient treatment. Only every third patient receives specialized healthcare in high-income countries and less than every twelfth in low-income countries (Moitra et al., 2022). Key reasons include stigma and insufficient levels of awareness but also lengthy diagnostic processes and a shortage of trained medical personnel. Digital technology is considered an important lever to bridge the treatment gap by complementing and augmenting traditional care (World Health Organization, 2022). Digital phenotyping represents a promising approach and attracts substantial academic interest (Mendes et al., 2022). It refers to the usage of AI to analyze different types of data (e.g., physiological, behavioral, social) and to detect potential mental health disorders. Apart from traditionally collected clinical information, data from sensors now ubiquitously available in smart devices (e.g., mobile phones, smartwatches, other wearables) represent a valuable source of information (Liang et al., 2019). Smart sensing data could support healthcare professionals, facilitate remote intervention, and enable patient self-help.

However, despite the expected potential to cover a wide range of the diverse expressions of mental health disorders, "most studies still only scratch the surface" (Miralles et al., 2020, p. 13) of smart sensing capabilities. Digital phenotyping based on such data has not reached clinical maturity yet. Whilst academia has built early models for the symptomatic prediction of depression, anxiety, and schizophrenia (Moshe et al., 2021; Terhorst, Sander, et al., 2023; Wang et al., 2016), very few mental health applications utilizing smart sensing exist in practice so far. The issue that AI models are often a black box to their users represents a fundamental obstacle to potential widespread adoption. Health-related data and the decisions derived from it are of particularly sensitive nature. Therefore, the enablement of users to critically evaluate AI output is of high relevance (Ali, Akhlaq, et al., 2023). A lack thereof inhibits AI adoption (Loh et al., 2022) and hence impedes unlocking AI's potential for healthcare. This is particularly relevant for mental health, in which stigma and limited awareness intensify the sensitivity of the topic further.

Explainable artificial intelligence (XAI) can help solve this black-box issue. This research field investigates how to improve human-AI collaboration through automatically generated explanations, i.e., human-understandable lines of reasoning to help users interpret AI models' output (Adadi & Berrada, 2018; Arrieta et al., 2020). A growing body of research is concerned with designing XAI approaches and investigating their effectiveness in application domains such as business

URI: https://hdl.handle.net/10125/109010
978-0-9981331-8-8

management, industrial production, finance, and healthcare (Islam et al., 2022). With regards to the latter, existing research largely focuses on clinical decision support for medical personnel concerned with the treatment of physiological (versus psychological) conditions (Loh et al., 2022). There is a lack of understanding about the potential role of XAI for mental health applications, particularly in regard to lay users, i.e., citizens who aim to better understand and preserve or improve their mental health condition. Existing findings on the effectiveness of explanations in other application fields cannot directly be transferred to the mental health domain, given its high complexity, symptomatic diversity, and patient sensitivity (Antoniou et al., 2022).

Therefore, the aim of this paper is to explore the potential of XAI in addressing mental health issues as a supplement to the usage of AI. We build on mental health literature to conceptualize potential effects of explanations in terms of patient empowerment. Through an online experiment, we investigate explanations' effectiveness for the case of a mental health app, in which lay users receive AI-generated depression risk predictions based on smart sensing data and ecological-momentary-assessment (EMA) self-reports (Shiffman et al., 2008). Subjects in our study interact with a fully functional AI system predicting depression risk based on real-world data, while we manipulate the presence of two types of explanations provided by state-of-the-art XAI approaches: feature importance explanations generated by LIME (Ribeiro et al., 2016) and counterfactual explanations generated by CARE (Rasouli & Yu, 2024). Results suggest that counterfactual explanations can support the effectiveness of AI applications in mental health by increasing users' intention to use and their patient empowerment. Our contribution to research is twofold: first, we establish an initial understanding of the role of XAI in the mental health domain. Second, our study is the first to conceptualize and demonstrate the positive impact of explanations on users' empowerment. Thus, we contribute to XAI literature by introducing this important concept as a potential target variable.

The remainder of this paper is structured as follows: in Sections 2 and 3, an overview of related studies is provided and the theoretical background for our work is established. In Section 4, we describe our case setting, experimental procedure, participants, and measurement. In Sections 5, we present our analysis methods and results. We conclude the paper with a critical discussion of the implications as well as limitations of our research and provide directions for future research.

2. Related work

As the capabilities of AI models advance, their opacity increases, rendering functioning and output uninterpretable to humans (Berente et al., 2021). On the one hand, such opacity can lead to humans blindly relying on AI, substituting their own judgment with potentially false decisions. On the other hand, the lack of interpretability may lead to reluctance to use AI (Brasse et al., 2023). To overcome these challenges, XAI research aims to provide automatically generated explanations alongside AI models and their output addressing four main purposes: (1) explain to learn, i.e., help humans discover knowledge from the output of an AI model; (2) explain to justify, i.e., help humans understand and scrutinize AI output; (3) explain to evaluate, i.e., sharpen humans' understanding of potential unknown vulnerabilities and flaws in an AI model; and (4) explain to improve, i.e., help developers correct and refine an AI model (Adadi & Berrada, 2018; Meske et al., 2022). Various XAI methods have been established, which can be distinguished by their model dependency and scope of interpretability (Adadi & Berrada, 2018). Model-specific methods generate explanations for individual (classes of) AI models, while model-agnostic methods generate explanations for various (classes of) AI models. Methods with a global scope aim to explain how an entire model functions, while methods with a local scope seek to explain specific output. Over the past years, the focus of XAI research has widened to also explicitly consider the perspective of stakeholder groups such as users, developers, and regulators. Nevertheless, empirical studies that help understand the role and impact of XAI for users in different application domains are scarce (Gerlings et al., 2021). Within healthcare, key reasons to apply XAI include a better clinical understanding of diagnoses, enhanced patient care, improved healthcare system structures, and compliance with data privacy and ethics principles (Albahri et al., 2023). Much of the existing research aims to support diagnostic processes based on medical imaging (such as X-ray or MRI), patients' medical records, or other health-related data. It is hence primarily targeted at clinicians and medical practitioners (Ali, Akhlaq, et al., 2023).

Given the symptomatic complexity, multidimensionality, and subjectivity of psychological disorders, AI systems have not reached the same maturity in mental health as in some other medical fields yet (Yan et al., 2022). Meanwhile, given the omnipresence of smartphones and other sensor-bearing consumer electronics, the interest in applying AI methods on smart sensing data for the detection and management of health disorders has recently surged. One major focus area for application is to determine causes and consequences of mental disorders (Bhatt et al., 2022). While revolutionary potential is expected for mental health research and treatment (Mohr et al., 2017), the respective AI models have not yet reached maturity (Abd-Alrazaq et al., 2023). Furthermore, materializing the high aspirations

will, as in other domains, depend on end users' ability to leverage AI output for their benefit. In this regard, XAI could contribute to the further realization of AI's alleged potential in the mental health domain. However, XAI research in this field is rare. To date, XAI has mainly been used in mental health as a tool to enhance the AI models applied by researchers to identify correlations between people's mental health status and empirical observations from brain imaging, surveys, or online postings (Byeon, 2023; Joyce et al., 2023), and recently activity and sleep data from wearables (Misgar & Bhatia, 2024; Tsai et al., 2024). But the benefit of XAI for end users (persons affected by or at risk for mental health issues) has hardly been researched yet.

Initial studies have generally identified the need for targeted explanations. For example, Chatterjee et al. (2024) propose a deep-learning model for mood prediction and enhance it with an XAI component so that the AI output is easier to understand and use for clinicians. Shibuya et al. (2023), in a pilot study, added what-if explanations to their predictions of stress, which could potentially be used by individuals with high mental health risk. Jaber et al. (2022) developed a blood test-like explanatory report for stress predictions based on data from wearables, which is supposed to help doctors understand the AI output for individual patients. They tested their report with three psychiatry experts. These researchers are pioneers in that they address questions of target group-specific design for XAI in mental health. However, there has not been any empirical investigation yet of the impact of explanations on relevant user groups, particularly patients and other lay persons.

3. Theoretical background

We build on mental health literature to conceptualize how XAI might influence the effectiveness of mental health apps. Such apps provide self-directed or virtually facilitated mental health services with regards to communication, monitoring, diagnosis, and treatment (Koh et al., 2022). Besides purposes such as psychological education and medication tracking, the most common purpose of mental health apps is self-help, i.e., to assist users achieve symptom relief using self-help methods (Camacho et al., 2022; Radovic et al., 2016). The underlying mechanism is patient empowerment, which has become a priority concept in healthcare globally to enhance the impact and sustainability of care (Fumagalli et al., 2015). In the mental health domain, patient empowerment can be defined as people's ability to control their own mental health and be more involved in their care (Pekonen et al., 2020). Patient empowerment, in the context of digital technology usage, can be conceptualized through two lenses: process orientation (patient process) and outcome orientation (patient outcome) (François et al., 2024). The concept of patient process is rooted in self-determination theory and describes patients' competence building, i.e., the gathering of knowledge on and understanding of their mental health status before they are enabled to make active choices (Funnell et al., 1991). Patient outcome is built on the concept of self-efficacy and describes patients' sense of being able to impact their mental health condition for the better (McAllister et al., 2012).

We build on the body of existing XAI knowledge to identify the type of explanations that can be expected to be most suitable in the context of mental health apps used for self-help. Local explanations have been shown to be particularly well-suited for lay users with limited AI literacy, such as average citizens (Brasse et al., 2023). They can help users rationalize specific AI output, such as predictions and recommendations on their (individual) mental health condition. This is in line with the XAI purposes of explaining to justify and learn. A variety of local explanation types exist. According to Schwalbe and Finzel (2023), these include explanations by example, counterfactuals, feature importance attributions, rule-based explanations, prototypes, diagrams, and combinations of these. Feature importance attributions are among the most prominent and well-established types of explanations, as they describe the importance of input features for an AI output and are thus particularly easy to understand (Speith, 2022). Applied to mental health apps, feature importance attributions can help users identify the key factors influencing their mental health condition, thereby sharpening users' understanding and hence increasing their ability to preserve or improve their mental health. Counterfactual explanations are considered among the most user-friendly types of explanations, as they mimic how humans construct explanations (Byrne, 2019). They uncover what it would have taken for an AI model to come to a different output, i.e., to which degree the levels of specific input features would have had to be different for the AI output to change (Cheng et al., 2021). Applied to mental health apps, counterfactual explanations can support users not only in identifying the relevant factors influencing their mental health status, but also in understanding how these factors would have to change to preserve or improve users' mental health condition, thereby facilitating self-help. Hence, to initially explore the potential role of XAI for mental health apps, we investigate the impact of local explanations, specifically feature importance attributions and counterfactual explanations.

4. Research method

To empirically test the impact of XAI in mental health apps, we conducted a randomized between-subjects online experiment, in which we manipulated the

presence of explanations. We aimed to design an experimental setting in which participants embrace the scenario of a hypothetical user of a mental health app. An app was designed based on real-world data, a fully functioning AI model, and two fully functioning XAI methods. The goal of the app is to help users maintain or improve their mental health condition in the sense of self-help. In the control group, participants received an AI-generated prediction of depression risk. In two separate treatment groups, participants received not only the AI-generated prediction of depression risk, but also either a feature importance explanation or counterfactual explanation. We assessed the impact of these two types of explanations on patient empowerment and two further measures that depict whether users intend to use such a mental health app (intention to use) and whether they would take action based on such an app's output (intention to act). In the following, we provide details about the setting, dataset and algorithms, as well as experimental procedure, participants, and measurement.

4.1. Setting, dataset, and algorithms

We implemented an online experiment in which participants interacted with a digital mental health app designed to support users in maintaining or improving their mental health condition. We trained an AI model based on a real-world dataset to predict depression risk and instantiated two XAI methods to generate feature importance and counterfactual explanations alongside the AI output. The dataset, collected by Terhorst, Messner, et al. (2023), includes 190 instances with smart sensing data, such as phone usage, calls, or mobility, and EMA data on features such as quality of sleep, quantity and quality of social contacts, and quality of nutrition. Additionally, for each instance, the dataset contains each person's result from the Patient Health Questionnaire (PHQ), i.e. the PHQ-8 value, which is an established instruments to detect depression risk (Kroenke et al., 2009). A PHQ-8 value of 10 is considered the best clinical cut-off value to differentiate between depression (high risk, PHQ-8 >=10, 19% of cases in our dataset) and no depression (low risk, PHQ-8 <10, 81% of cases in our dataset) (Wu et al., 2020).

Accordingly, our AI model solves a classification task with two classes based on the smart sensing and EMA data included in the dataset. Recent studies have shown that more flexible models like XGBoost seem more capable of dealing with the non-linear and complex relationships in this kind of data and hence tend to outperform other models in our application field (Abd-Alrazaq et al., 2023; Opoku Asare et al., 2021). Based on this, we have chosen to train an XGBoost model. Parameter tuning was performed to improve model accuracy while avoiding overfitting. It resulted in a learning rate of 0.05, 100 trees in the ensemble, and a maximum depth of 4 in each tree. Using an 80/20 train-test-split, the trained AI model yielded an accuracy of 0.76 (further performance measures are displayed in Table 1). The performance of the AI model can be considered good but not perfect – as it would have to be expected in a real-world setting.

Table 1. Performance measures of the AI model.

Class	Performance measure	Value
Depression (high risk)	Precision	0.88
	Recall	0.68
	F1-score	0.77
No depression (low risk)	Precision	0.65
	Recall	0.87
	F1-score	0.74

We instantiated two model-agnostic XAI methods to generate feature importance explanations and counterfactual explanations, respectively, alongside the AI-generated depression risk predictions. Following the notion that explanations should be designed with a user focus (Ali, Abuhmed, et al., 2023), we selected the XAI methods with the needs of our target group in mind, i.e., lay users using a mental health app. To generate feature importance explanations, we chose the well-established XAI method LIME (Ribeiro et al., 2016). The resulting explanations are particularly simple to interpret compared to other XAI methods like SHAP – even if those methods generate feature importance explanations that are more theoretically sound. LIME generates perturbed data samples around a certain instance and constructs a simplified model, which approximates the AI model in the vicinity of the respective instance. We generated 5,000 samples around a certain instance using Gaussian sampling and Ridge regression as a simplified model to identify the (up to) 10 most important features for the prediction. To generate counterfactual explanations, we instantiated the state-of-the-art XAI method CARE. Its explanations are particularly actionable, in the sense that proposed changes are feasible according to an individual's situation (Rasouli & Yu, 2024) – which is relevant to enable self-help. For our case, CARE appears therefore superior to other XAI methods generating counterfactual explanations which focus, e.g., on sparsity (Wachter et al., 2018) or abnormality (Jahn et al., 2024). CARE generates a set of counterfactual explanations for a certain instance and selects one of them based on an optimization approach. We set the number of counterfactual explanations to be generated to 10. To determine the explanation alongside the prediction, we instantiated CARE to optimize for soundness, coherency, and actionability apart from the proximity to the original instance.

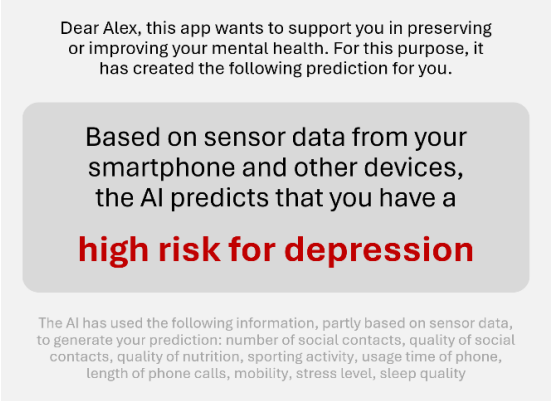

Figure 1. Exemplary screen of the app for a hypothetical user (group *control*).

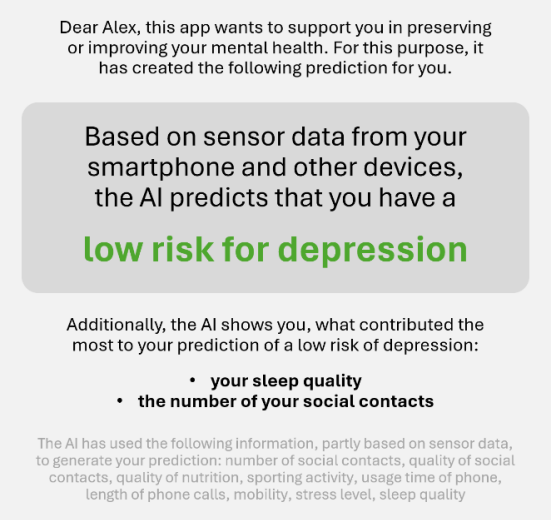

Figure 2. Exemplary screen of the app for a hypothetical user (group *feature importance*).

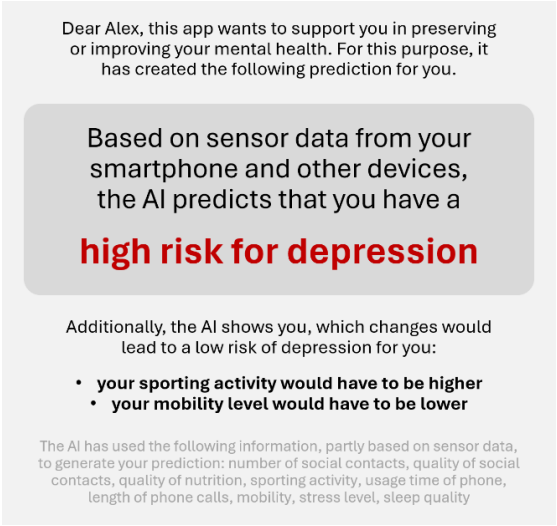

Figure 3. Exemplary screen of the app for a hypothetical user (group *counterfactual*).

4.2. Experimental procedure

For our experiment, participants were randomly assigned to the control group or one of the two treatment groups. The experiment consisted of three parts. The first part entailed an introduction to the experiment and a short survey to collect control variables. On the initial screen, the mental health app and its goals were introduced. Participants were informed about attention checks in all parts of the experiment.

The second part comprised the participants' interaction with the mental health app. They were familiarized with the scenario of a hypothetical user. Subsequently, participants were shown the information which the app had prepared for that hypothetical user. According to their random assignment, the information contained (1) an AI-generated prediction on depression risk (group *control*), (2) an AI-generated prediction on depression risk and a feature importance explanation (group *feature importance*), or (3) an AI-generated prediction on depression risk and a counterfactual explanation (group *counterfactual*). Figures 1-3 display examples for all three groups, *control* (Figure 1), *feature importance* (Figure 2) and *counterfactual* (Figure 3). Each participant was presented an individual scenario of a hypothetical user, including this user's AI-generated prediction of depression risk and the corresponding explanation (dependent on group assignment). To create each scenario using the instantiated algorithms, an underlying data instance was randomly drawn from the pool of instances in the dataset.

The third part of the experiment consisted of an online survey with items on subjective measures related to patient empowerment, intention to use, and intention to act, as well as two attention checks.

4.3. Participants

Participants were recruited via the platform Prolific, an established online platform providing high data quality for academic research (Eyal et al., 2022). Inclusion criteria were fluent German language skills and German, Austrian, or Swiss nationality, given that the experiment was conducted in German language and the underlying dataset was collected in Germany. A total of 437 participants took part in our experiment, of which 28 were excluded for not passing attention checks or having suspiciously short completion times. The final sample entailed 409 subjects with 135, 136, and 138 in the *control, feature importance*, and *counterfactual* group, respectively. Among the participants in the final sample, 146 stated to be female, 262 stated to be male, and one preferred not to state their gender. Participants' age ranged from 18 to 68.

4.4. Measurement

As dependent variables, we measured patient empowerment, intention to use, and intention to act. We adjusted existing item inventories for measurement. Patient empowerment was measured by four items for each sub-construct, patient process and patient outcome, adapted from François et al. (2024) and their reference papers (McAllister et al., 2012; Oh & Lee, 2012). To measure intention to use, three items were adapted from Henkel et al. (2023) and to measure intention to act, we relied upon four items adapted from Mackenzie et al. (2004). Additionally, we measured participants' pre-existing experience with mental-health apps as a control variable with three items (Venkatesh et al., 2003). All items were measured on a 7-point Likert scale ranging from *strongly disagree* to *strongly agree*.

5. Results

The aim of our empirical investigation was to test the impact of feature importance explanations and counterfactual explanations alongside AI output in mental health apps on patient empowerment, intention to use, and intention to act. We calculated construct scores for each participant as the mean of the respective item scores. For all constructs, internal consistency reliability was ensured: the Cronbach's alpha values exceeded the recommended threshold of 0.7 without having to remove any item. With regards to our three groups, the effectiveness of random assignment was confirmed through chi-square tests. No significant differences between the groups *control*, *feature importance*, and *counterfactual* were found in terms of pre-existing experience with mental health apps and socio-demographic variables relevant for our context (namely sex, age, nationality, and employment status, which we obtained from Prolific).

To compare the three experimental groups regarding each construct, we conducted a one-way analysis of variance (ANOVA) and least significant difference (LSD) post-hoc tests in R. As illustrated in Figure 4, ANOVA results indicate significant differences between groups for the constructs *intention to use*, *patient process*, and *patient outcome*. LSD results show that the participants receiving counterfactual explanations in addition to the AI-generated prediction of depression risk provide significantly higher values for *intention to use*, *patient process*, and *patient outcome* than participants receiving the AI output on depression risk alone.

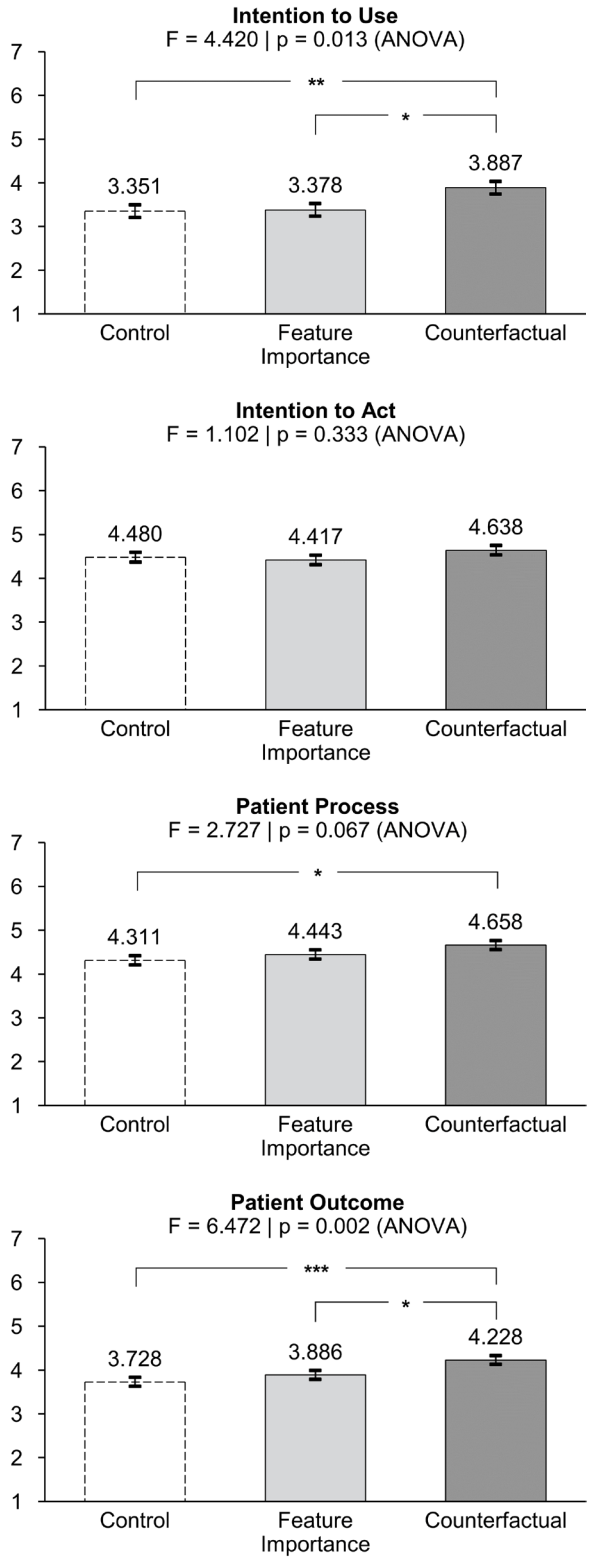

Figure 4. Overview of mean values per construct and explanation type (error bars indicate +/- 1 standard error); significance levels according to LSD tests: * p < 0.05, ** p < 0.01, *** p < 0.001

In contrast, feature importance explanations did not show the same impact. With regards to the construct *intention to act*, we did not find any significant differences. Altogether, our results show that only counterfactual explanations, not feature importance explanations, increased intention to use and patient empowerment (i.e., patient process and patient outcome).

In anticipation of the potential difficulty for participants to embrace the scenario of a hypothetical user (particularly when their own mental state does not comply with the example they are presented), we also compared the mean values for each construct between the participants who received depression (high risk) versus no depression (low risk) as an AI prediction in our experiment (cf. Table 2). The fact that participants receiving depression (high-risk) as an AI prediction reported a higher intention to act than those receiving no depression ($p < 0.05$ according to LSD) suggests that participants could relate to the hypothetical user.

Table 2. Overview of mean values per construct and exemplary case type

Construct	Depression		No depression	
	Mean	SD	Mean	SD
Intention to Use	3.645	1.620	3.429	1.770
Intention to Act	4.698	1.231	4.314	1.281
Patient Process	4.460	1.210	4.485	1.282
Patient Outcome	4.000	1.133	3.895	1.252

6. Conclusion, Implications, and Limitations

AI usage in mental health apps, especially based on sensor data, has attracted large academic interest. Given the sensitivity of this domain, interpretability of AI output is essential. The research field of XAI provides approaches to automatically generate explanations aiming to help users interpret AI output, thereby overcoming the black-box issue inherent to many AI models. However, the role of explanations for users of mental health apps has not yet been investigated. This study aimed to explore the effects of explanations on patient empowerment, intention to use, and intention to act. We empirically investigated the effectiveness of two types of explanations – feature importance and counterfactual explanations – through a randomized between-subject online experiment with 409 participants. The results yield three important contributions to IS research.

First, our results show that automatically generated explanations can increase patient empowerment in mental health apps. In our study, the provision of counterfactual explanations alongside AI-generated predictions on depression risk was effective in increasing knowledge on and understanding of mental health status (patient process) as well as the sense of being able to change it for the better (patient outcome). Existing XAI research has so far focused on constructs such as trust (Banovic et al., 2023), perception and understanding of AI (Elshawi et al., 2019), advice acceptance (Bayer et al., 2022), and task performance (van der Waa et al., 2021). To the best of our knowledge, our study is the first to conceptualize and demonstrate the positive impact of explanations on users' empowerment. Thus, we contribute to IS literature by introducing this important concept as a potential target variable in XAI research. Future XAI research might consider the impact of explanations on empowerment in other domains where personal AI-based recommendations have substantial impact on a user's life (e.g., education, job search, financial investments).

Second, our findings demonstrate that explanations in mental health apps can increase the intention to use. AI adoption is particularly relevant in this domain, because it will likely play a vital role in advancing towards "the vision of a digitalized, patient-centered, and data-driven mental health ecosystem" (Kalman et al., 2023, p. 1), which can help overcome the global shortage of care (Moitra et al., 2022). Previous XAI studies have already shown the positive impact of explanations on AI adoption in domains like education (Conati et al., 2021), insurance (Baroni et al., 2022), and recruiting (Fleiß et al., 2020). We expand this body of literature by demonstrating that the positive influence of explanations on AI adoption is also valid in the mental health domain. At the same time, our findings indicate potential differences in the effectiveness of XAI on lay users versus experts in healthcare, as previous studies have not found a significant impact of explanations on intention to use for doctors and nurses (Panigutti et al., 2022).

Third, our study suggests that it is not the mere presence of explanations, but rather their effective design that can help realize the potential of AI in the mental health domain. In our experiment, only counterfactual explanations were effective in increasing patient empowerment and intention to use while feature importance explanations did not make a difference as compared to the provision of AI output alone. Other than feature importance explanations, which merely state the input variables that were most relevant for the AI output, counterfactuals constitute a particularly user-centric type of explanation, which is similar to how humans explain (Byrne, 2019) and provides guidance for action (Rasouli & Yu, 2024). These nuanced findings might inform design researchers in developing further XAI approaches specifically for the mental health domain.

Apart from these academic contributions, our research is relevant for practitioners who aim to develop mental health apps. While a huge market has developed

for such products, the development of methods that are able to provide users with actual feedback on their mental health condition, e.g., based on smart sensing data, is in its infancy (Camacho et al., 2022). Developers should consider integrating counterfactual explanations in their apps to increase adoption and empower users to understand and improve their mental health condition.

While our study has revealed interesting initial insights on the impact of XAI in the mental health domain, it is subject to several limitations, which offer opportunities for future research. First, our experiment was conducted for one single use case incorporating one specific mental health app. Despite having been built on a real-world dataset and modeled to exhibit characteristics of real-life applications, the app is still artificial. The use case allowed us to recruit German-speaking participants for our experiment. These two aspects limit external validity. Therefore, future studies might replicate and adapt our experiment for different use cases and in different geographies. Second, our online experiment was scenario-based, which means that participants were not personally affected in that they did not see AI output and explanations that resulted from their own data. We encourage future researchers to conduct field experiments, in which users are personally invested in using mental health apps. Third, we investigated two types of local explanations which were expected to be particularly easy to interpret and facilitate self-help. Future studies might want to explore the impact of other types of explanations to examine requirements for successful XAI design in the mental health domain. For example, future research should consider different types of local explanations (such as explanations by examples) and global explanations, which may provide users with a more general understanding of factors impacting mental health. Fourth, our study has exploratory character and aims to provide a first step towards a better understanding of the role of explanations in the mental health domain. While our work provides empirical insights that counterfactual explanations increase patient empowerment and intention to use, we invite researchers to also investigate the underlying mechanisms for these effects (e.g., using structural equation modeling and considering contextual factors such as the user's psychological state).

7. References

Abd-Alrazaq, A., AlSaad, R., Shuweihdi, F., Ahmed, A., Aziz, S., & Sheikh, J. (2023). Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. *NPJ Digital Medicine*, *6*(1), Article 84.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O. S., Alamoodi, A. H., Bai, J. S., Salhi, A., Santamaria, J., Ouyang, C., Gupta, A., Gu, Y. T., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156–191.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*.

Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in Biology and Medicine*, *166*, Article 107555.

Antoniou, G., Papadakis, E., & Baryannis, G. (2022). Mental Health Diagnosis: A Case for Explainable Artificial Intelligence. *International Journal on Artificial Intelligence Tools*, *31*(3), Article 2241003.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Banovic, N., Yang, Z., Ramesh, A., & Liu, A. (2023). Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), Article 27.

Baroni, I., Re Calegari, G., Scandolari, D., & Celino, I. (2022). AI-TAM: a model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications. *Human Computation*, *9*(1), 1–21.

Bayer, S., Gimpel, H., & Markgraf, M. (2022). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, *32*(1), 110–138.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450.

Bhatt, P., Liu, J., Gong, Y., Wang, J., & Guo, Y. (2022). Emerging Artificial Intelligence-Empowered mHealth: Scoping Review. *JMIR MHealth and UHealth*, *10*(6), Article e35053.

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, *33*(1), Article 26.

Byeon, H. (2023). Advances in Machine Learning and Explainable Artificial Intelligence for Depression Prediction. *International Journal of Advanced Computer Science and Applications*, *14*(6), 520–526.

Byrne, R. M. J. (2019). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Camacho, E., Cohen, A., & Torous, J. (2022). Assessment of Mental Health Services Available Through Smartphone Apps. *JAMA Network Open*, *5*(12), Article e2248784.

Chatterjee, S., Mishra, J., Sundram, F., & Roop, P. (2024). Towards Personalised Mood Prediction and Explanation for Depression from Biophysical Data. *Sensors*, *24*(1), Article 164.

Cheng, F. R., Ming, Y., & Qu, H. M. (2021). Dece: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1438–1447.

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*, Article 103503.

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, *19*(1), Article 146.

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662.

Fleiß, J., Bäck, E., & Thalmann, S. (2020). Explainability and the intention to use AI-based conversational agents. In *Proceedings of the First International Workshop on Explainable and Interpretable Machine Learning*.

François, J., Audrain-Pontevia, A.-F., Boudhraâ, S., & Vial, S. (2024). Assessing the Influence of Patient Empowerment Gained Through Mental Health Apps on Patient Trust in the Health Care Provider and Patient Compliance With the Recommended Treatment: Cross-sectional Study. *Journal of Medical Internet Research*, *26*, Article e48182.

Fumagalli, L. P., Radaelli, G., Lettieri, E., Bertele, P., & Masella, C. (2015). Patient Empowerment and its neighbours: Clarifying the boundaries and their mutual relationships. *Health Policy*, *119*(3), 384–394.

Funnell, M. M., Anderson, R. M., Arnold, M. S., Barr, P. A., Donnelly, M., Johnson, P. D., Taylor-Moon, D., & White, N. H. (1991). Empowerment: An idea whose time has come in diabetes education. *The Diabetes Educator*, *17*(1), 37–41.

Gerlings, J., Shollo, A., & Constantiou, I. D. (2021). Reviewing the Need for Explainable Artificial Intelligence (xAI). In *Proceedings of the 54th Hawaii International Conference on System Sciences*.

Global Burden of Disease Collaborative Network. (2024). *Global Burden of Disease Study 2021*. Institute for Health Metrics and Evaluation (IHME).

Henkel, T., Linn, A. J., & van der Goot, M. J. (2023). Understanding the Intention to Use Mental Health Chatbots Among LGBTQIA+ Individuals: Testing and Extending the UTAUT. In *Proceedings of the 6th International Workshop on Chatbot Research and Design* (pp. 83–100). Springer International Publishing.

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences-Basel*, *12*(3), Article 1353.

Jaber, D., Hajj, H., Maalouf, F., & El-Hajj, W. (2022). Medically-oriented design for explainable AI for stress prediction from physiological measurements. *BMC Medical Informatics and Decision Making*, *22*(1), Article 38.

Jahn, T., Hühn, P., & Förster, M. (2024). Wasn't Expecting that – Using Abnormality as a Key to Design a Novel User-Centric Explainable AI Method. In *Design Science Research for a Resilient Future: 19th International Conference on Design Science Research in Information Systems and Technology* (pp. 66–80). Springer-Verlag.

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digital Medicine*, *6*(1), Article 6.

Kalman, J. L., Burkhardt, G., Samochowiec, J., Gebhard, C., Dom, G., John, M., Kilic, O., Kurimay, T., Lien, L., Schouler-Ocak, M., Vidal, D. P., Wiser, J., Gaebel, W., Volpe, U., & Falkai, P. (2023). Digitalising mental health care: Practical recommendations from the European Psychiatric Association. *European Psychiatry*, *67*(1), Article e4.

Koh, J., Tng, G. Y. Q., & Hartanto, A. (2022). Potential and Pitfalls of Mobile Mental Health Apps in Traditional Treatment: An Umbrella Review. *Journal of Personalized Medicine*, *12*(9), Article 1376.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1-3), 163–173.

Liang, Y. J., Zheng, X. L., & Zeng, D. D. (2019). A survey on big data-driven digital phenotyping of mental health. *Information Fusion*, *52*, 290–307.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). *Computer Methods and Programs in Biomedicine*, *226*, Article 107161.

Mackenzie, C. S., Knox, V. J., Gekoski, W. L., & Macaulay, H. L. (2004). An adaptation and extension of the attitudes toward seeking professional psychological help scale. *Journal of Applied Social Psychology*, *34*(11), 2410–2435.

McAllister, M., Dunn, G., Payne, K., Davies, L., & Todd, C. (2012). Patient empowerment: The need to consider it as a measurable patient-reported outcome for chronic conditions. *BMC Health Services Research*, *12*, Article 157.

Mendes, J. P. M., Moura, I. R., van de Ven, P., Viana, D., Silva, F. J. S., Coutinho, L. R., Teixeira, S., Rodrigues, J. J. P. C., & Teles, A. S. (2022). Sensing Apps and Public Data Sets for Digital Phenotyping of Mental Health: Systematic Review. *Journal of Medical Internet Research*, *24*(2), Article e28735.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1), 53–63.

Miralles, I., Granell, C., Díaz-Sanahuja, L., van Woensel, W., Bretón-López, J., Mira, A., Castilla, D., & Casteleyn, S. (2020). Smartphone Apps for the Treatment of Mental Disorders: Systematic Review. *JMIR MHealth and UHealth*, *8*(4), Article e14897.

Misgar, M. M., & Bhatia, M. P. S. (2024). Unveiling psychotic disorder patterns: A deep learning model analysing

motor activity time-series data with explainable AI. *Biomedical Signal Processing and Control*, *91*, Article 106000.

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, *13*, 23–47.

Moitra, M., Santomauro, D., Collins, P. Y., Vos, T., Whiteford, H., Saxena, S., & Ferrari, A. J. (2022). The global gap in treatment coverage for major depressive disorder in 84 countries from 2000-2019: A systematic review and Bayesian meta-regression analysis. *PLOS Medicine*, *19*(2), Article e1003901.

Moshe, I., Terhorst, Y., Asare, K. O., Sander, L. B., Ferreira, D., Baumeister, H., Mohr, D. C., & Pulkki-Råback, L. (2021). Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Frontiers in Psychiatry*, *12*, Article 625247.

Oh, H. J., & Lee, B. (2012). The effect of computer-mediated social support in online communities on patient empowerment and doctor-patient communication. *Health Communication*, *27*(1), 30–41.

Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021). Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study. *JMIR MHealth and UHealth*, *9*(7), e26540.

Panigutti, C., Beretta, A., Pedreschi, D., & Giannotti, F. (2022). Understanding the impact of explanations on advice-taking: A user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing (CHI'22)*.

Pekonen, A., Eloranta, S., Stolt, M., Virolainen, P., & Leino-Kilpi, H. (2020). Measuring patient empowerment - A systematic review. *Patient Education and Counseling*, *103*(4), 777–787.

Radovic, A., Vona, P. L., Santostefano, A. M., Ciaravino, S., Miller, E., & Stein, B. D. (2016). Smartphone Applications for Mental Health. *Cyberpsychology, Behavior and Social Networking*, *19*(7), 465–470.

Rasouli, P., & Yu, I. C. (2024). Care: Coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, *17*(1), 13–38.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*.

Shibuya, K., King, Z. D., Khalid, M., Yu, H., Shen, Y. F., Zanna, K., Brown, R. L., Majd, M., Fagunders, C. P., & Sano, A. (2023). Predicting Stress and Providing Counterfactual Explanations: A Pilot Study on Caregivers. In *Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*.

Shiffman, S., Stone, A. A., & Hufford (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32.

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2239–2250). Association for Computing Machinery.

Terhorst, Y., Messner, E. M., Asare, K. O., Montag, C., Kannen, C., & Baumeister, H. (2023). Which Smartphone-Based Sensing Features Matter in Depression Prediction? Results from an observation study. *JMIR Preprints*, Article 55308. https://preprints.jmir.org/preprint/55308

Terhorst, Y., Sander, L. B., Ebert, D. D., & Baumeister, H. (2023). Optimizing the predictive power of depression screenings using machine learning. *Digital Health*, *9*.

Tsai, C. H., Christian, M., Kuo, Y. Y., Lu, C. C., Lai, F. P., & Huang, W. L. (2024). Sleep, physical activity and panic attacks: A two-year prospective cohort study using smartwatches, deep learning and an explainable artificial intelligence model. *Sleep Medicine*, *114*, 55–63.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, Article 103404.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887.

Wang, R., Aung, M. S. H., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., Hauser, M., Kane, J., Merrill, M., Scherer, E. A., Tseng, V. W. S., & Ben-Zeev, D. (2016). Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *UBICOMP'16: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

World Health Organization. (2022). *World mental health report: transforming mental health for all*. World Health Organization.

Wu, Y., Levis, B., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., Rice, D. B., Boruff, J., Cuijpers, P., Gilbody, S., Ioannidis, J. P. A., Kloda, L. A., McMillan, D., Patten, S. B., Shrier, I., Ziegelstein, R. C., Akena, D. H., Arroll, B., Ayalon, L., . . . Thombs, B. D. (2020). Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: A systematic review and individual participant data meta-analysis. *Psychological Medicine*, *50*(8), 1368–1380.

Yan, W.-J., Ruan, Q.-N., & Jiang, K. (2022). Challenges for Artificial Intelligence in Recognizing Mental Disorders. *Diagnostics*, *13*(1).