Original Article

Enhancing Clinical Psychology Practice through Data-driven Machine Learning Monitoring Systems

Guillem Martínez[1]

Adriana Trujillo[3]

Johnny Núñez[2]

Julio C. S. Jacques[1,4]

Albert Clapés[1,4]

Beatriz Domínguez-Álvarez[2]

Laura Viñals [2]

Juan C. Onieva[3]

Sergio Escalera[1,4]

David Gallardo-Pujol[2]

[1]Dept. of Mathematics and Informatic, University of Barcelona, Barcelona, Catalonia, Spain

[2]Dept. of Clinical Psychology and Psychobiology, University of Barcelona, Barcelona, Catalonia, Spain

[3]MetrikaMind Health, S.L., Barcelona, Catalonia, Spain

[4]Computer Vision Center (CVC), Autonomous University of Barcelona, Bellaterra, Catalonia, Spain

Abstract:

Mental health is a critical global challenge, with significant societal impacts. Machine learning has emerged as a valuable and innovative approach to enhancing psychological assessment and intervention. This study focuses on leveraging mental health indicators for depression and anxiety to develop and test a data-driven alert system for a workplace-centred psychological monitoring system. Synthetic cases data (N=192) were generated through clinically-based criteria, while real evaluation data (N=489) were derived from evaluations provided by experts (N=46) across five mental health diagnostic fields. We adopt a multi-layer supervised machine learning approach (MLP) to address various predictive tasks, incorporating SHAP analysis (explainable AI) and Monte Carlo Dropout (uncertainty estimation) to enhance interpretability and reliability. The study demonstrated that MLPs achieved a commendable performance, with an alert system score of 0.68 on a scale of 0 to 1, and that agreement between evaluators is crucial to train supervised models. We showed that agreement between experts, although not strong, is higher than that of our MLP models and experts for medical discharge. These findings highlight the potential of these methods to improve mental health monitoring in workplace settings, while also underscoring the need for more extensive data collection in real-world environments to further validate and refine the system in future applications of Metrikamind's service.

Keywords: Psychological Assessment, Mental Health Monitoring, Machine Learning, Explainable AI, Uncertainty Estimation

Guillem Martínez Sánchez
Dept. of Mathematics and Informatic
University of Barcelona
Gran Via de les Corts Catalanes, 585
08007, Barcelona, Catalonia
Spain

guillem.martinez@ub.edu

Introduction

Depression and anxiety are among the leading causes of disability worldwide (WHO, 2017). There is substantial evidence that personalised approaches for managing individual clinical profiles, rather than broad diagnoses, can help to better understand not only the diagnoses, but the course of depression and anxiety symptoms over time (Deif & Salama, 2021; Schoevers et al., 2021; Walz et al., 2014). A personalised approach seems, thus, a prerequisite to enhance the assessment and treatment of these prevalent conditions. Despite growing commitment to precision psychiatry, methodological limitations such as data availability, data quality, or reproducibility and interpretability, have so far precluded the translational implementation of data-driven solutions to clinical settings.

Data-driven approaches are not new Psychology research. Yet, modern data-driven computational methods have begun to be widely used in research and clinical applications *only r ecently* (Fokkema et al., 2022). There is much to learn from Machine Learning (ML) approaches applied to the diagnosis, prognosis and treatment for common mental disorders (Dwyer et al., 2018). The related literature suggests that modern methods from ML may have potential to improve the well-being of psychiatric patients (Bzdok & Meyer-Lindenberg, 2018). In fact, ML algorithms have been used to predict the severity of major depressive disorder and anxiety symptoms over time (Kessler et al., 2016; Li et al., 2024). Evidence indicates, also, that a relatively small number of variables (e.g., constellations of no more than 15) could yield an efficient and highly scalable system for mental health assessment of anxiety symptoms (Bari et al., 2024).

The increasing prevalence of mental health issues, particularly in the workplace, has highlighted the need for scalable and effective solutions for monitoring and supporting mental well-being. Numerous studies have demonstrated the potential of machine learning across various contexts. Some approaches rely on physical devices for monitoring, such as those

used in tracking health metrics (Iyortsuun et al., 2023), although these methods can be intrusive and difficult to scale effectively, especially when considering a workplace setting. Other approaches focus on clinical data to predict therapeutic outcomes in clinical populations with depression (Lee et al., 2018), or depression and suicidal ideation among specific populations such as medical interns (Horwitz et al., 2023). The application of machine learning for workplace mental health monitoring has gained considerable attention, as improving mental well-being is increasingly recognized as vital to enhancing employee productivity and overall workplace morale (Fokkema et al., 2022). Despite this growing interest, most studies tend to focus on specific scenarios or settings (e.g., residency in medical training, Horwitz et al., 2023), leaving a gap in comprehensive solutions for general mental health monitoring in the workplace.

Here, we extend this approach by developing and testing a *data-driven alert system* aimed at guiding clinical decision making in the context of workplace mental health. The study we present is an example of ML applied to prediction research. By uncovering the features underlying the predictive models of depression and anxiety severity and course (e.g., medical discharge), we expect to gain valuable insights into the evolution of these problems over time in a cost-effective way. This work is intended to set the stage for additional investigations into how to better understand, assess and treat anxiety and depression using ML methods.

The environment where we aim to develop these methods is Metrikamind. It is a digital mental health ecosystem designed for work environments. It is focused on improving the assessment of anxiety and depression in the workforce through the use of advanced technology, psychometrics, and artificial intelligence (Gallardo-Pujol et al., 2024). The objective of Metrikamind is to provide a precise and effective monitoring system that improves early detection of mental health changes, supports interventions, and enhances

employee well-being, productivity, and organisational performance, aligning with SDGs (Gallardo-Pujol et al., 2022). Metrikamind's platform could facilitate innovations in health insurance, enabling the provision of personalised mental health coverage (Fokkema et al., 2022). Additionally, enhanced accuracy in assessments could facilitate the identification of novel therapeutic targets, thereby accelerating progress in personalised mental health care.

This study examines the potential of Metrikamind, and specifically, the data-driven alert system, as a transformative tool for workplace mental health monitoring by addressing key questions related to its predictive performance, alignment with expert assessments, and its role in enhancing clinical decision-making for anxiety and depression.

To achieve this, the following research questions are explored:

- What is the predictive capability of a data-driven alert system for personalized clinical psychology, and how confident is the model in its predictions?
- Can the model's feature importances provide meaningful insights into clinical decision-making?
- To what extent does the machine learning model's feature prioritization align with expert assessments, and how does this alignment vary across different pathologies?

Material and Methods

Participants

After reviewing proposals from different authors (Belton et al., 2019), it was expected that the sample would consist of at least 35 experts from different fields related to mental health and/or temporary work incapacity management. At the end of the platform evaluation phase, 45 experts had collaborated, the group including clinical psychologists, psychiatrists, occupational physicians, general practitioners and labour inspectors.

A number of criteria were taken into account in their selection, such as their area of specialization, clinical experience, publications in the field or nominations from other recognized experts in the field (Belton et al., 2019). Based on these criteria, a convenience sampling method was employed. Participants were recruited through announcements shared on various social networks and through word-of-mouth. The selected experts were compensated financially for their participation.

Table 1

Metrikamind's indicators and their categories

Indicators	Categories
Pathology	Depression; Anxiety; Depression&Anxiety
Initial Severity	Severe; Moderate-severe; Moderate; Low-moderate; Low; Normal
Current Severity	Severe; Moderate-severe; Moderate; Low-moderate; Low; Normal
Evolution (general, along two weeks)	Short time favourable/Long time favourable; Short time favourable/Long time negative; Short time negative/Long time favourable: Short time negative/Long time negative
Psychological risk	High; Low
Work stress (Burnout)	High; Low
Work performed	White-collar worker; Blue-collar worker
Engagement (to work)	High; Low
Resilience	High; Low
Evolving Honesty	High; Low
Current Honesty	High; Low
Evolving Responsibility	High; Low
Current Responsibility	High; Low
Evolving Adherence	High; Low
Current Adherence	High; Low
Medication Side Effects	Intense; Moderate; Low; None

Materials

Metrikamind's platform employs a range of measures to gather data on the characteristics of each user, encompassing socio-demographic information, responses to multiple questionnaires designed to assess clinical and personality profile aspects, and data pertaining to prescription follow-up, response rate during follow-ups, and other relevant variables. The aforementioned information is then aggregated into a specific set of categories and indicators.

Indicators

In the platform, a series of indicators were utilised, with their values graphically represented. These indicators subsequently served as a feature to facilitate decision-making regarding the patient in question. Table 1 describes the indicators and the possible categories in which they could be represented.

Severity indicator

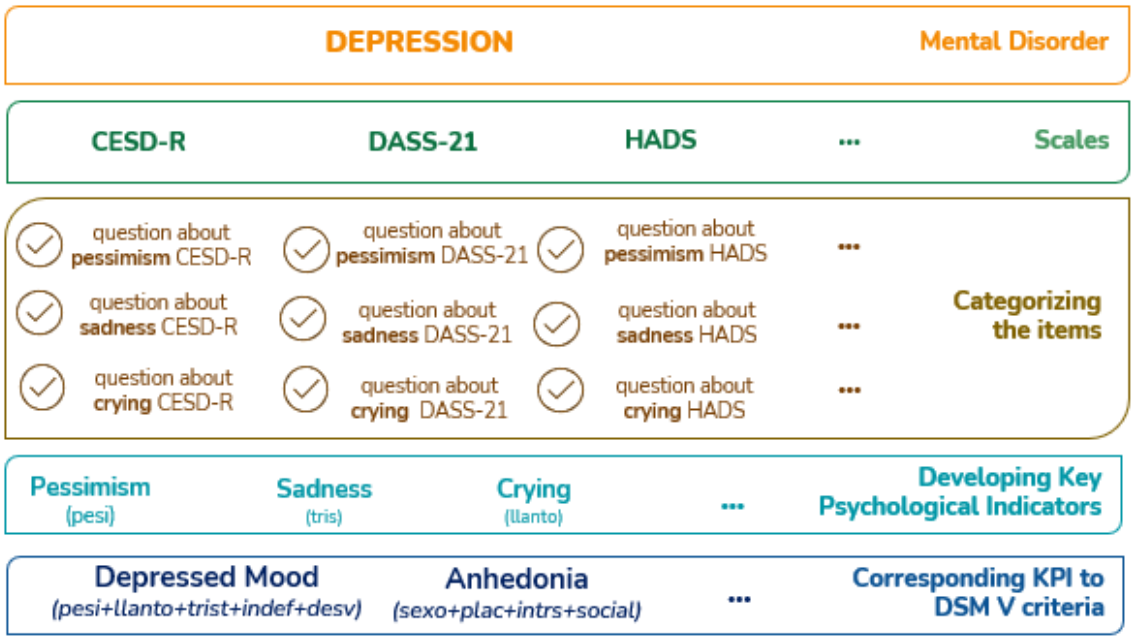
This category describes the course of depressive and anxiety symptoms according to the criteria described in the DSM-5, with a value of 0-10 depending on the fulfillment of the different criteria.

Key Psychological Indicators. In order to monitor compliance with these diagnostic criteria, a set of Key Psychological Indicators (16 KPIs for depression and 14 for anxiety) was developed from a review of the content of the most commonly used scales for assessing depression and anxiety. Using this criterion, 6 scales were identified and reviewed for depression: the Center for Epidemiological Studies Depression (CESD-R; Radloff, 1977), the Depression Anxiety Stress Scales (DASS-21; Lovibond & Lovibond, 1995), the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983), the Patient Health Questionnaire PHQ9; (Spitzer et al., 1999), the Patient-Reported Outcomes Measurement Information System (PROMIS; Pilkonis et al., 2011) and the Quick Inventory of Depressive

Symptomatology (QIDS-C; Rush et al., 2003), while 5 scales were identified and reviewed for anxiety: the Depression Anxiety Stress Scales (Lovibond & Lovibond, 1995), the Generalised Anxiety Disorder Scale (GAD-7; Spitzer et al., 2006), the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983), the Zung Self-Rating Anxiety Scale (Zung, 1971) and the Taylor Manifest Anxiety Scale (TMAS; Taylor, 1953). Figure 1 illustrates the interrelationship between scales, indicators and diagnostic criteria.

Figure 1

Mapping Scales, Item Categorization, and Indicators to DSM-5 Depression Criteria



Note. Illustrate the relationship between the scales, the categorisation of the items, the development of the indicators and their correspondence with the DSM-5 criteria for depression.

Alerts

An alert on Metrikamind platform is an automated notification that is triggered when significant changes in the evolution of a user's depression and anxiety symptoms are detected. These alerts are triggered by changes in the scores that could change the level of severity or changes in the evolution that are detected by applying the formula for identifying reliable and

clinically significant changes (Evans et al., 1998; Castillo, 2010) . In the context of sickness leave, those alerts will include information from different categories of indicators, beyond depressive or anxious symptomatology.

Procedure

Data collection

To adequately prepare the platform for expert assessment, a simulated database was created in which the performance of the indicators could be observed by combining different scoring profiles. As a result of these combinations, 192 simulated cases were obtained. The combination of variables included and simulated score categories are available in the supplementary material (https://osf.io/phmk9/?view_only=cd57a7f071994704a2fc484ec3000d3e).

Once the cases were uploaded to the web platform, the recruitment of experts started via social networks and voice-to-voice. Data for the study was collected exclusively from judges who gave informed consent. Subsequently, two 45-minute videocalls were scheduled for each expert. The initial session provided a theoretical overview of the scientific model underlying Metrikamind's platform and its operational principles. Additionally, it showed the procedure to be followed for the assessment of indicators and alerts.

The second videocall addressed any remaining queries regarding the procedure to be followed and provided illustrative examples of the manner in which experts should conduct the assessment. Each expert evaluated 11 cases, which were distributed to ensure diversity in terms of diagnosis, severity, and the categories presented in the different indicators (see supplementary material). To complete the assessment, each expert had to review the profile of the assigned patients and then navigate to the summary page. On this page, each expert could find a list of characteristics of each patient, some of which were related to personal or work

environment information, as later there was a list of indicators associated with a level (e.g. high or low) or category (e.g. severe, moderate, low) according to one of the possibilities previously explained in Table X. . Subsequently, participants were required to navigate a list of indicators , organised in order of importance (e.g. pathology, initial severity, burnout), in order to inform clinical, organisational and sick leave-related decisions for the simulated patients. The minimum number of indicators to be used by each expert was eight out of a total of 16 available.

At the end of the activity, experts were required to respond to a series of questions pertaining to the perceived level of severity of each simulated patient, the anticipated recovery time, and the necessity of issuing an alert based on the information and indicators presented. In the event of a positive response, the expert was required to evaluate whether to transmit an alert comprising clinical information or an alert pertaining to the patient's involvement in the assessment. Additionally, they were tasked with determining the urgency of the notification and selecting a timeframe within which the alert was expected to be resolved. Ultimately, the participant was required to determine whether further actions pertaining to the alert should be undertaken, such as contacting the relevant professional, contemplating a potential referral, or modifying the treatment regimen.

Upon completion of the assessment of the assigned cases, the participants were invited to complete a survey on the usability of the platform. In return for their participation, they were sent a voucher as a reward for their contribution.

Data Analysis and Processing

An important aspect of working with the Metrikamind database is addressing cases where multiple experts may evaluate the same instances. While not all cases are reviewed by multiple professionals, there is an overlap in evaluations that provides an opportunity to

analyze the level of agreement among experts. The primary goal of preprocessing is to ensure the dataset is clean, structured, and aligned with the requirements of the analysis while preserving its integrity and relevance to the problem domain. During preprocessing, columns and rows with excessive missing values were removed. Numerical features were standardized and in the case of categorical features: nominal ones were transformed to one-hot encoding, while ordinal categories were ordinally encoded.

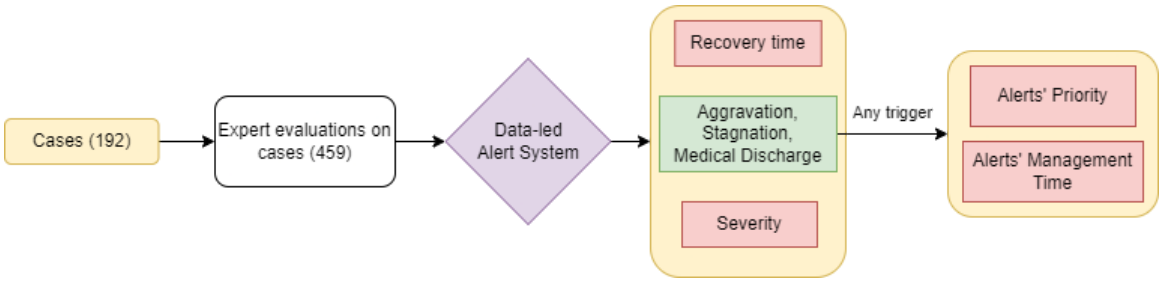
This study specifically focuses on medical alerts. As such, we did not aim to assess honesty or responsibility-related alerts, and cases' evaluations that exclusively trigger these types of alerts were excluded from the analysis.

Data-driven Alert System

For the predictive model, we used Multi-Layer Perceptrons (MLP, Rumelhart et al., 1986) as base classifiers for binary prediction tasks. This choice is motivated by the MLP's capability to perform Monte Carlo dropout (MC dropout; Gal & Ghahramani, 2016), enabling us to derive valuable insights into uncertainty measurements.

Figure 2

Training logic behind the data-driven alert system



Note. A set of synthetic cases evaluated by experts generate 459 data points. Tasks are color-coded: ordinal regression tasks in red and multilabel classification tasks in green. For cases that trigger alerts, the data is utilized to train two additional tasks.

As shown on Figure 2, the system comprises multiple tasks. For our approach, we opted to train each task independently. Ordinal regression tasks were reformulated as a series of binary classification problems, following the methodology outlined in Frank & Hall (2001). For the trigger of the alerts, since each task is a binary task, we optimize them within a multi-label classification framework instead.

For the first set of tasks—recovery time, severity, and alert trigger predictions—all data is used to train the models, ensuring comprehensive learning. However, for alert priority and management time, only the subset of data containing an alert trigger is employed, allowing the model to focus on task-relevant information. This task-specific approach enables each model to adopt representations best suited to its prediction goal, minimizing conflicts between tasks. Furthermore, it provides clearer insights into how features influence each task, avoiding the obscured relationships often seen in end-to-end systems with shared representations. While the models are trained separately, they are designed to function seamlessly in an end-to-end manner during prediction, ensuring smooth integration and consistent outputs.

Hyperparameter search

Hyperparameter search is essential for achieving optimal model performance by systematically exploring parameter configurations to improve predictive accuracy and generalization. The hyperparameters of the base MLP models are fine-tuned using the Optuna hyperparameter optimization framework, which efficiently explores the search space to identify optimal configurations (Akiba et al., 2019). To ensure robust and unbiased evaluation, each trial undergoes assessment through a nested cross-validation loop. This approach minimizes overfitting and provides reliable estimates of the model's performance on unseen data.

Explainability Algorithms

One of the most criticized aspects of machine learning methods has been the "black-box" problem, which refers to their lack of transparency in decision-making. To overcome this challenge, explainability methods are employed to shed light on the inner workings of these models. To address these issues, for the interpretability of our models, we utilized Shapley Additive Explanations (SHAP, Lundberg & Lee, 2017) and Monte Carlo (MC) Dropout. SHAP not only clarifies the model's decision-making process but also allows us to compare its outputs with expert criteria, helping to assess and understand any differences.

To further enrich the interpretability and performance insights of our models, we employed MC dropout. This technique estimates uncertainty by introducing variability in the model's predictions through random dropout of neurons during inference. By repeating this process multiple times, we capture the distribution of predictions, giving us a deeper understanding of the model's confidence. A model with low standard deviation obtained from the MC samples will be of higher confidence than one with high standard deviation.

Results

The final sample was composed of 24 women (53.3%) and 21 men (46.7%) (N=45). The participants included 12 clinical psychologists (26.6%), six psychiatrists (13.3%), 10 occupational physicians (22.2%), eight general practitioners (17.7%), and nine labour inspectors (20%). All of the participants were employed in a variety of mental health settings, where they were engaged in the provision of care and management of temporary incapacity due to common mental disorders.

Data Analysis

This section comprises the information on the preprocessing and basic statistics for this study database.

Preprocessing - After applying the aforementioned preprocessing we go from 492 evaluations to 407. Also, reducing the cases from 192 to 190. In this case, we do not drop any case or variable for having missing values, since the input data is synthetic and we have all required features for each case.

Dataset statistics - Out of the 190 cases, 38 were evaluated once, 96 were evaluated twice, 47 were evaluated three times, and 9 were evaluated nine times, with an average of 2.14 evaluations per case. A detailed summary of the cases and the corresponding evaluation combinations across occupational groups is provided in Table 2.

Table 2

Evaluation's occupational groups matrix and individual evaluations

Occupation	*Occupation*				
	Labor Inspectors	*Primary Care Physicians*	*Occupational Physicians*	*Psychologists*	*Psychiatrist*
Labor Inspectors	*14*	*4*	*14*	*10*	*10*
Primary Care Physicians	-	*11*	*10*	*11*	*12*
Occupational Physicians	-	-	*7*	*9*	*10*
Psychologists	-	-	-	*12*	*6*
Psychiatrist	-	-	-	-	*12*
Individual evals	*9*	*11*	*9*	*4*	*5*
Evals per occupation	*94*	*76*	*76*	*79*	*82*

Experts Disagreement Analysis

For the agreement analysis, we consider cases that have been evaluated multiple times and assess the level of agreement between occupational groups. The metric used for this purpose

is the Mean Absolute Difference (MAD). Since the variables in question are ordinal, a MAD value of 1 indicates that experts, on average, disagree by one step on the ordinal scale represented by the variable we are querying.

Figure 4

Mean Absolute Difference among occupational groups for medical discharge alert trigger

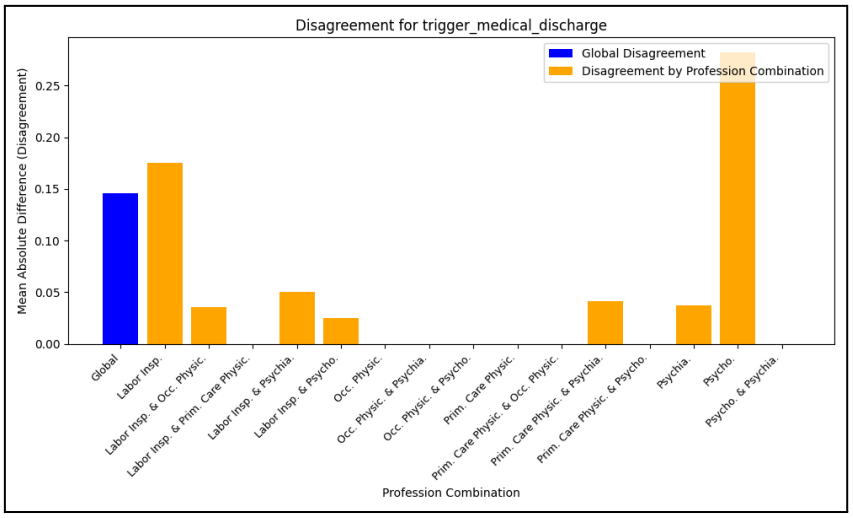
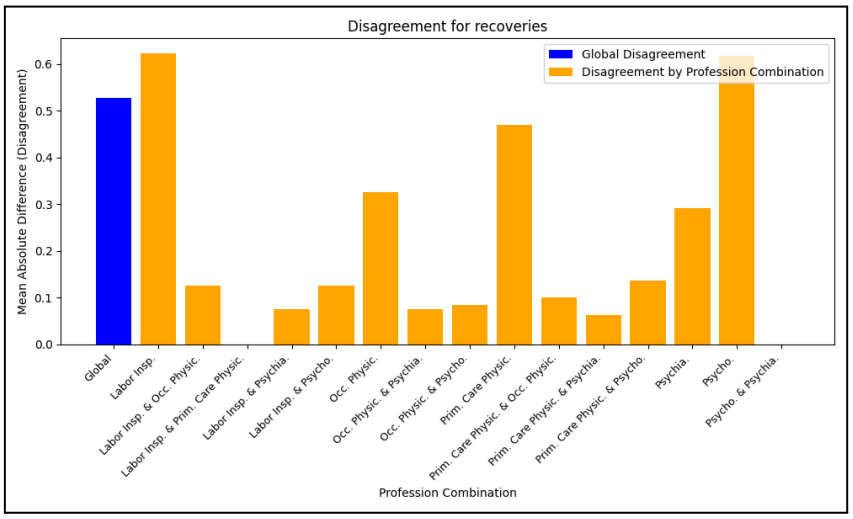


Figure 5

Mean Absolute Difference among occupational groups for providing a recovery time



Observing Figure 4, the agreement for the trigger of the medical discharge, we observe an overall acceptable agreement across expert groups, except from psychologists (0.283) and among labor inspectors (0.179). For recovery times prediction we visualize it in Figure 5, in

this case, agreement across professional categories is lower. Additional plots for all tasks can be found in the supplementary material.

System Performance

The alert system is composed of a set of ordinal regression and a multi-label classification task. Therefore, the model is evaluated on each task independently, an overview of all tasks and performance can be seen in Table 3. The Optuna's hyperparameter search framework (Akiba et al., 2019) runs and search space can be found at the supplementary material, together with additional information.

Table 4

Performance and Uncertainty Metrics Across Monte Carlo Samples

Data	*Task*	*Performance metrics*		*Uncertainty Metrics*	
		F1 Score	*NMAE*	*MC - mSD*	*MC - mKLD*
All Cases	Recovery Time	-	0.248	0.049	0.155
	Severity	-	0.165	0.037	0.149
	Aggravation Alert	0.533	-	0.039	0.031
	Stagnation Alert	0.571	-	0.049	0.040
	Medical Discharge	0.794	-	0.045	0.036
Cases with alert emission	Alert's Management Time	-	0.287	0.050	0.159
	Alert's Priority	-	0.243	0.059	0.144

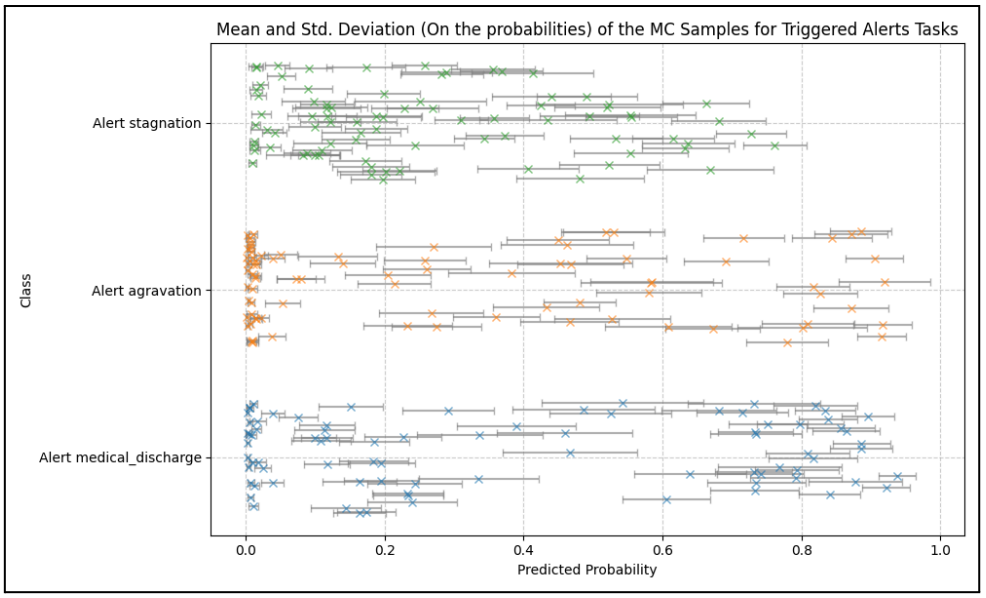
Note. mSD refers to the *mean Standard Deviation*, and *mKLD* represents the *mean KL Divergence*. Both metrics are calculated as the mean across the Monte Carlo (MC) samples

Table 4 shows that Severity performs the best among the ordinal tasks. For alert predictions, the model struggles with Aggravation and Stagnation alerts, but performs better on Medical Discharge.

Figure 6 displays the mean standard deviation of MC samples for all tasks. Higher variance indicates more uncertainty (around 0.5), while Medical Discharge predictions are concentrated near 0 or 1, suggesting better separation between patients requiring a discharge and those who do not.

Figure 6

Task-Specific Alert Probabilities and Monte Carlo Sampling Metrics



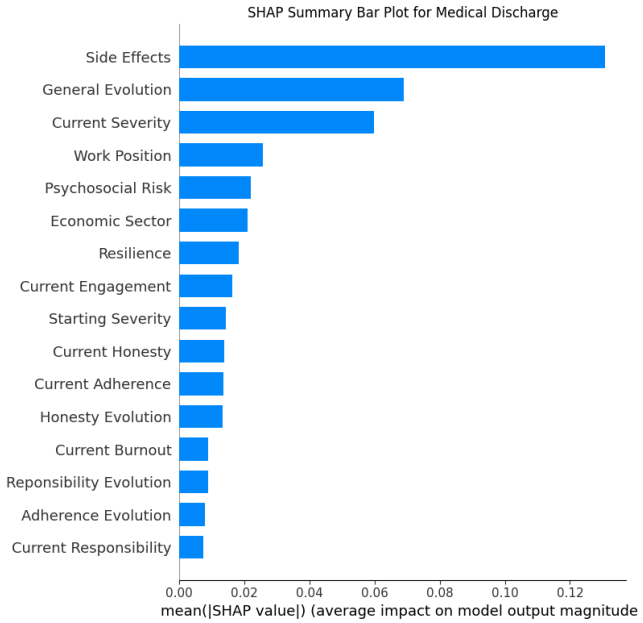
Note. Each alert task is represented by a distinct color. For each task, the probability of alarm activation (ranging from 0 for "no" to 1 for "yes") is displayed. Additionally, the Monte Carlo samples for each task are provided, along with the mean and standard deviation of the predicted probabilities obtained during dropout Monte Carlo sampling.

Explainability Analysis

In this section, we utilize SHAP and the feature importance rankings provided by experts regarding the factors influencing medical discharge. This allows us to compare the key features identified by the AI as most significant for prediction with those deemed most important by the experts. To gain deeper insights into the similarity between rankings, we use Kendall's Tau (Arndt et al., 1999) for pairwise comparisons (e.g. expert-expert and MLP-expert). As shown in Table 7, while expert ratings are not strongly correlated, they exhibit greater alignment compared to the ratings of the MLP model when evaluated against the experts, this shared among all pathologies.

Figure 7

SHAP Value Magnitudes for Key Indicators in Medical Discharge Prediction



Note. Secondary effects, general evolution and current severity are the three main discriminative features for the trigger of the alert.

Table 5

Average Kendall's Tau for Indicator Rankings in Medical Discharge Alerts by Pathology

Pathology	*Mean Kendall's Tau*	
	MLP vs Expert Rank(s)	Expert vs Expert Rank
Depression	0.202	0.310
Anxiety	0.176	0.280
Depressión & Anxiety	0.182	0.341

Discussion

The goal of this study was to design and evaluate a data-driven alert system within the workplace-focused digital mental health platform, Metrikamind, utilizing indicators for depression and anxiety. Specifically, the study aimed to assess the predictive capability of the system in providing personalized clinical psychology alerts and the confidence of the model in its predictions. Additionally, the study explored whether the feature importances identified

by the machine learning model could offer actionable insights for clinical decision-making. Furthermore, the research examined how the model's feature prioritization aligns with expert evaluations, particularly across various pathologies, shedding light on its applicability and reliability in diverse clinical contexts.

Predictive capability of the data-driven alert system

Overall, the predictive capability of the alert system is better than random on *aggravation alert*, *stagnation alert* (Above 0.50 F1-Score), *recovery time*, *alert's management time* and *alert's priority* predictions (All around 0.25 NMAE), and acceptable for *severity* prediction (0.165 NMAE) and *medical discharge alert (0.794 F1-Score)*. The low performance in stagnation and aggravation alerts is closely linked to data quality issues. This is largely due to greater disagreement among experts compared to the triggering of alerts. Additionally, the task itself poses challenges for accurate determination, as it involves time-based predictions that may contain initial inaccuracies or slightly deviate from actual outcomes, particularly when the ordinal time scale is non-linear. This challenge also extends to managing alert timing. Furthermore, prioritizing alerts and their managing time is another area marked by significant expert disagreement.

Model's feature importances and clinical decision-making

The Severity indicator demonstrates the greatest efficacy in relation to ordinal tasks.This can be attributed, at least in part, to the fact that it is based on robust and comprehensible criteria for the various experts who assessed the platform, despite their variety of background or role in relation to the administration of temporary incapacity for work. The Severity indicator is derived from compliance with the criteria set out in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), which, despite its limitations and criticisms, remains a widely accepted reference in mental health diagnosis (Buckley, 2014).

Additionally, the degree of compliance with these criteria is based on the assessment of different KPIs, based on validated psychological questionnaires, whose usefulness has been widely documented both in research and clinical practice, as identified in the review of the questionnaire that was conducted and previously described in the procedure.

However, when these same criteria are used to assess the patient's evolution, specifically to issue alerts related to Aggravation or Stagnation in their progress, the model seems to show limitations. In this context, the use of reliable and clinically significant change indicators could provide greater consistency (Evans et al., 1998; Fitzgerald, & Blampied, 2016; Wise, 2004). Although graphical representation of these indicators was available to experts, it was not used as an indicator in the assessment of alerts. The incorporation of more detailed graphical representations or the development of new indicators derived from the current criteria could strengthen the reliability of warnings of Aggravation or Stagnation. This refinement is crucial to ensure that alerts meet the platform's objectives and support evidence-based decisions.

A notable positive aspect is that the recognition of discharge appears to be clearer and more consistent among experts, despite their differences in role, training and professional experience. This finding is of particular relevance as it aligns with the platform's strategic objectives and underscores its practical utility in clinical settings

The alignment of machine learning model's feature prioritization with expert assessments

Disagreement among raters evaluating the same psychological construct (e.g., personality, psychopathology) is well-documented (Faherty et al., 2020; Way et al., 1998). Our findings of disagreement within and between professional groups align with this literature. However, the

moderate disagreement within some groups, particularly among psychologists and labor inspectors, may be problematic and difficult to interpret.

While multiple raters are generally preferred over one, substantial disagreement within the same professional group can introduce "noise" into the data, potentially hindering ML model performance. The consistently higher disagreement for psychologists and labor inspectors compared to other groups (e.g., psychiatrists) warrants further investigation, possibly related to rater training or other group-specific factors.

Interestingly, disagreement appears reduced when raters from different professional categories (e.g., psychologist and primary care physician) evaluate the same cases, compared to evaluations within a single category (e.g., psychologist and psychologist). These findings highlight the importance of using diverse rater groups to enhance assessment reliability and improve the quality of data for training ML models.

Limitations

This study has several limitations that should be acknowledged. Firstly, the patient data used in the analysis is synthetic and artificially balanced, which raises concerns about its ability to generalize effectively to real-world scenarios. While this issue can be addressed by re-training the alert system with real data, the current results must be interpreted with caution. Secondly, the small number of cases and evaluations limits the depth of analysis regarding professional agreement. To better understand the consistency and variability among professionals, a larger sample size and multiple raters from different fields evaluating the same cases would be required. Such an approach would allow for a more robust comparison and improve the reliability of an agreement analysis. Lastly, the limited sample size, coupled with high levels of disagreement among professionals, introduces noise that significantly impacts performance and certainty of the model.

Ethical implications

Artificial intelligence has not yet fully impacted fields such as psychological assessment (Fokkema et al., 2022), but we hope that works like this one will become widespread. Its use in mental health can transform how we approach these issues, but only if done with care and responsibility (MacIntyre et al., 2023). In an alert system like the one we propose in this paper, it is very important that there is a human responsible for each decision (Naik et al., 2022). An ethical commitment that puts people at the center of attention is essential. Systems must be fair, avoiding any bias or discrimination, especially towards the most vulnerable groups. Moreover, transparency is key: patients and professionals must understand how the systems work and how decisions are made, as a lack of clarity can generate unnecessary stress or even facilitate malpractices, in a very sensitive context such as occupational health (Martinez-Martin, 2021). If used correctly, these tools can help to detect significant problems in time, reducing delays in critical diagnoses, intervening in time is also a crucial advancement before a minor issue becomes more serious (Vollmer et al., 2020) , and this alert system works in this direction. In every data set, there is a person with their own narrative and a range of emotions that deserve to be treated with respect. Although technology plays a crucial role in current progress, the human essence remains fundamental and irreplaceable

Conclusions

This study presents an exploratory effort to develop an alert system within a mental health ecosystem designed to support professional decision-making. By creating a semi-synthetic database that combines synthetic diversity in patient data with real expert assessments, we established a foundation for testing the alert system's capabilities. The resulting models provide valuable information on key aspects such as severity, recovery time, and alerts for

stagnation, aggravation, and medical discharge, as well as guidance on managing alert timing and prioritization.

In addition, we performed explainability analysis using SHAP values, which allowed us to interpret the model's decision-making processes and assess alignment with expert criteria. Uncertainty measurements, conducted through Monte Carlo dropout, further provided insights into the reliability and confidence of the system's predictions. Together, these contributions glimpse on the potential for data-driven tools to transform mental health monitoring and decision-making in professional practice, upon a reasonable amount of curated real data is employed.

Conflict of Interest

Authors declare no conflict of interest

Authorship

All authors approved the final version of the article.

Guillem Martínez and Adriana Trujillo share first authorship.

Funding

This research is part of the CPP2021-008590 project, funded by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and the European Union-NextGenerationEU/PRTR, and received support from the Catalan Government Grant 2021SGR00709 (in both DG-P is the PI). In addition, this research is part of the PTQ2021-011931 project, also funded by MCIN/AEI/10.13039/501100011033, awarded to AT.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation*

Hyperparameter Optimization Framework (No. arXiv:1907.10902). arXiv. https://doi.org/10.48550/arXiv.1907.10902

Arndt, S., Turvey, C., & Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation. *Journal of Psychiatric Research*, *33*(2), 97–104. https://doi.org/10.1016/s0022-3956(98)90046-2

Bari, S., Kim, B.-W., Vike, N. L., Lalvani, S., Stefanopoulos, L., Maglaveras, N., Block, M., Strawn, J., Katsaggelos, A. K., & Breiter, H. C. (2024). A novel approach to anxiety level prediction using small sets of judgment and survey variables. *Npj Mental Health Research*, *3*(1), 1–16. https://doi.org/10.1038/s44184-024-00074-x

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting and Social Change*, *147*, 72–82. https://doi.org/10.1016/j.techfore.2019.07.002

Buckley, M. R. (2014, July 11). *Back to Basics: Using the DSM-5 to Benefit Clients - The Professional Counselor*. https://tpcjournal.nbcc.org/back-to-basics-using-the-dsm-5-to-benefit-clients/

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Castillo, I. I. (2010). Evaluación de resultados clínicos (y III): Índices de Cambio Fiable (ICF) como estimadores del cambio clínicamente significativo. *Norte de Salud Mental*, *8*(36), 105–122.

Deif, R., & Salama, M. (2021). Depression From a Precision Mental Health Perspective: Utilizing Personalized Conceptualizations to Guide Personalized Treatments. *Frontiers in Psychiatry*, *12*, 650318. https://doi.org/10.3389/fpsyt.2021.650318

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for

Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *BMJ Ment Health*, *1*(3), 70–72. https://doi.org/10.1136/ebmh.1.3.70

Faherty, A., Counihan, T., Kropmans, T., & Finn, Y. (2020). Inter-rater reliability in clinical assessments: Do examiner pairings influence candidate ratings? *BMC Medical Education*, *20*(1), 147. https://doi.org/10.1186/s12909-020-02009-4

Fokkema, M., Iliescu, D., Greiff, S., & Ziegler, M. (2022). Machine learning and prediction in psychological assessment: Some promises and pitfalls. *European Journal of Psychological Assessment*, *38*(3), 165–175. https://doi.org/10.1027/1015-5759/a000714

Frank, E., & Hall, M. (2001). A Simple Approach to Ordinal Classification. In L. De Raedt & P. Flach (Eds.), *Machine Learning: ECML 2001* (pp. 145–156). Springer. https://doi.org/10.1007/3-540-44795-4_13

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of The 33rd International Conference on Machine Learning*, 1050–1059. https://proceedings.mlr.press/v48/gal16.html

Gallardo-Pujol, D., Trujillo, A., Domínguez-Álvarez, B., Martínez, G., Clapés, A., & Escalera, S. (2024). *Developing a Digital Mental Health Ecosystem for Workplaces: Rationale, Objectives, and Methods of the MetrikaMind Project*. OSF. https://doi.org/10.31219/osf.io/yqra6

Gallardo-Pujol, D., Ziegler, M., & Iliescu, D. (2022). Can psychological assessment contribute to a better world? Our discipline's contribution to the sustainable development goals for 2030. *European Journal of Psychological Assessment*, *38*(5),

347–355. https://doi.org/10.1027/1015-5759/a000739

Horwitz, A. G., Kentopp, S. D., Cleary, J., Ross, K., Wu, Z., Sen, S., & Czyz, E. K. (2023). Using machine learning with intensive longitudinal data to predict depression and suicidal ideation among medical interns over time. *Psychological Medicine*, *53*(12), 5778–5785. https://doi.org/10.1017/S0033291722003014

Iyortsuun, N. K., Kim, S.-H., Jhon, M., Yang, H.-J., & Pant, S. (2023). A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare*, *11*(3), Article 3. https://doi.org/10.3390/healthcare11030285

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., Ebert, D. D., Hwang, I., Li, J., de Jonge, P., Nierenberg, A. A., Petukhova, M. V., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry*, *21*(10), 1366–1371. https://doi.org/10.1038/mp.2015.198

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Li, Y., Song, Y., Sui, J., Greiner, R., Li, X.-M., Greenshaw, A. J., Liu, Y. S., & Cao, B. (2024). Prospective prediction of anxiety onset in the Canadian longitudinal study on aging (CLSA): A machine learning study. *Journal of Affective Disorders*, *357*, 148–155. https://doi.org/10.1016/j.jad.2024.04.098

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck

Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-u

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (No. arXiv:1705.07874). arXiv. https://doi.org/10.48550/arXiv.1705.07874

MacIntyre, M. R., Cockerill, R. G., Mirza, O. F., & Appel, J. M. (2023). Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments. *Psychiatry Research*, *328*, 115466. https://doi.org/10.1016/j.psychres.2023.115466

Martinez-Martin, N. (2021). Minding the AI: Ethical Challenges and Practice for AI Mental Health Care Tools. In F. Jotterand & M. Ienca (Eds.), *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues* (pp. 111–125). Springer International Publishing. https://doi.org/10.1007/978-3-030-74188-4_8

Naik, N., Hameed, B. M. Z., Shetty, D. K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Rai, B. P., Chlosta, P., & Somani, B. K. (2022). Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Frontiers in Surgery*, *9*, 862322. https://doi.org/10.3389/fsurg.2022.862322

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R., Thase, M. E., Kocsis, J. H., & Keller, M. B. (2003). The 16-Item Quick Inventory of Depressive Symptomatology

(QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*(5), 573–583. https://doi.org/10.1016/s0006-3223(02)01866-8

Schoevers, R. A., van Borkulo, C. D., Lamers, F., Servaas, M. N., Bastiaansen, J. A., Beekman, A. T. F., van Hemert, A. M., Smit, J. H., Penninx, B. W. J. H., & Riese, H. (2021). Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. *Psychological Medicine*, *51*(11), 1906–1915. https://doi.org/10.1017/S0033291720000689

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA*, *282*(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Taylor, J. A. (1953). A personality scale of manifest anxiety. *The Journal of Abnormal and Social Psychology*, *48*(2), 285–290. https://doi.org/10.1037/h0056264

Vollmer, S., Ba, M., G, B., Fj, K., R, G., P, J., S, C., A, J., Ksl, M., P, M., D, G., M, B., R, B., Kgm, M., Gs, C., Jpa, I., C, H., & H, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, l6927–l6927.

Walz, L. C., Nauta, M. H., & Aan Het Rot, M. (2014). Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: A systematic review. *Journal of Anxiety Disorders*, *28*(8), 925–937. https://doi.org/10.1016/j.janxdis.2014.09.022

Way, B. B., Allen, M. H., Mumpower, J. L., Stewart, T. R., & Banks, S. M. (1998). Interrater

Agreement Among Psychiatrists in Psychiatric Emergency Assessments. *American Journal of Psychiatry*, *155*(10), 1423–1428. https://doi.org/10.1176/ajp.155.10.1423

WHO. (2017, January 3). *Depression and Other Common Mental Disorders*. https://www.who.int/publications/i/item/depression-global-health-estimates

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, *67*(6), 361–370. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

Zung, W. W. (1971). A rating instrument for anxiety disorders. *Psychosomatics: Journal of Consultation and Liaison Psychiatry*, *12*(6), 371–379. https://doi.org/10.1016/S0033-3182(71)71479-0