D6.2: Implementation and documentation of a conducted study: Normative amygdala fMRI response during emotional processing as a trait of depressive symptoms in the UK Biobank

1. General information

This section captures the essential details of the study or evaluation and provides an overview of its scope, management and personnel.

Country and name of use case (take from list below): Netherlands: Major Depressive Disorder

Name of technical system/component: Normative Modeling Framework (Amygdala fMRI response analysis)

Study director/manager: Amsterdam UMC Location AMC - Matthan W. A. Caan

Partners involved:

- Amsterdam UMC

Start and end date of the study:

- Start: 15 July 2023
- Paper accepted: 27 August 2025

Status of the study: Completed

Ethical approval:

- The study utilizes the UK Biobank resource, which is a public dataset with pre-existing ethical approval.

2. Study Design

This chapter outlines the fundamental design and objectives of the research conducted. It provides the context for why the study was undertaken and how it was structured to address specific scientific questions regarding Major Depressive Disorder and neuroimaging.

2.1 Introduction

Major Depressive Disorder (MDD) is a leading cause of disability worldwide, yet its underlying biological mechanisms remain insufficiently understood. A prominent theory in the pathophysiology of MDD suggests the involvement of amygdala hyperreactivity to negative emotional stimuli. However, neuroimaging literature has produced inconsistent results regarding this biomarker; while some studies report hyperreactivity, others find blunted responses or no significant differences compared to healthy controls.

These inconsistencies may stem from small sample sizes in previous studies or the traditional case-control approach, which compares group averages. This conventional method relies on the assumption that the clinical group is homogeneous, potentially masking substantial biological heterogeneity within the disorder. To address these limitations, this study applies Normative Modeling (NM) to functional Magnetic Resonance Imaging (fMRI) data. Rather than assuming homogeneity, this approach maps

individual variation against a healthy reference population, allowing for the quantification of deviations at the level of the individual patient.

2.2 Goal of the study

The primary objective of this study is to move beyond group-average comparisons and investigate individual biological deviations in patients with MDD. Specifically, the study aims to establish a normative model of amygdala fMRI responses during emotional processing and to determine whether deviations from this norm—conceptually treated as "outliers"—are associated with depressive symptoms or diagnostic status.

2.3 Research questions and hypotheses

The study investigates the utility of normative modeling in characterizing the heterogeneity of depression. The central hypotheses and research questions addressing the biological underpinnings of MDD are defined below.

Hypotheses:

- It is hypothesized that patients with MDD will exhibit greater deviations (outliers) from the normative amygdala response compared to healthy controls.
- It is hypothesized that these deviations may serve as a trait marker for depression (indicating vulnerability or scarring) rather than a state marker of current symptom severity.

Research Questions:

- To what extent do individual amygdala fMRI responses in MDD patients deviate from a normative model derived from a healthy population?
- Are these deviations associated with current depressive symptom severity or with stable personality traits such as neuroticism?
- Do deviations differ between patients with Current Major Depressive Disorder (cMDD), and those with recurrent depression (rMDD)?

2.4 Study design

The study utilizes a large-scale, cross-sectional observational design. It leverages pre-existing data from a major biobank resource to apply advanced statistical modeling techniques to neuroimaging data.

Type of study: Cross-sectional observational study using retrospective data analysis.

2.5 Participants

This section details the large cohort utilized for the analysis, defining the population used to build the normative model and the clinical subgroups used for comparison.

Inclusion criteria:

- Participants from the UK Biobank cohort.
- Availability of complete fMRI data for the emotional processing task.

- Availability of relevant phenotypic data for defining healthy controls and depression subgroups.

Exclusion criteria:

- [Information not available in source document] (Standard MRI safety exclusions and data quality exclusions are implied but not explicitly listed in the provided text).

Groups:

- **Healthy Controls (HC):** Used to train the normative model.
- **Current Major Depressive Disorder (cMDD):** Participants meeting criteria for a current depressive episode.
- **Recurrent Major Depressive Disorder (rMDD):** Participants with a history of MDD who are not currently depressed.

Sample Size:

- Total sample for modeling: N = 17,294 (after exclusions).

2.6 Randomisation and blinding

As this is an observational study utilizing a pre-existing dataset (UK Biobank), traditional randomization and blinding procedures applicable to clinical trials were not employed.

Randomisation: Not applicable. **Blinding:** Not applicable.

2.7 Intervention and Control

The study does not involve a therapeutic intervention. Instead, it utilizes a standardized cognitive task to elicit neural responses which are then analyzed relative to a healthy control baseline.

Intervention:

- There is no pharmacological or psychotherapeutic intervention.
- **Task:** The Hariri emotional processing task (faces matching task) performed during fMRI scanning. This task is designed to probe amygdala reactivity to negative emotional stimuli (fearful/angry faces) compared to a control condition (shapes).

Control:

- The Healthy Control (HC) group from the UK Biobank serves as the reference population for constructing the normative model.

2.8 Assessments and measures

The study relies on a combination of neuroimaging data and psychometric assessments to quantify both brain function and clinical phenotype.

Neuroimaging:

- **fMRI:** Functional Magnetic Resonance Imaging measuring blood-oxygen-level-dependent (BOLD) responses.
- **Region of Interest (ROI):** The Amygdala (bilateral).

Psychometric and Clinical Measures:

- **Depressive Symptoms:** [Information not available in source document] (Likely PHQ-9 or similar standard measures used in UK Biobank).
- **Personality Traits:** Neuroticism (assessed via the Eysenck Personality Questionnaire Revised Short Form, as implied by references to UK Biobank protocols in similar literature).
- **Diagnostic Status:** Classification into cMDD, rMDD, or HC based on structured interview data (likely CIDI) and self-report.

3. Methods

This chapter describes the technical and statistical procedures used to process the data and generate the normative models. It details the transition from raw neuroimaging data to interpretable Z-scores representing individual deviation.

3.1 Introduction

The methodological approach focuses on establishing a robust statistical baseline for amygdala activity. By regressing out confounding variables such as age, sex, and head motion from a healthy cohort, the study isolates the variance attributable to the pathology of depression.

3.2 Software and tools supporting the data analysis of the study

The analysis relied on standard neuroimaging software packages and statistical programming environments to ensure reproducibility and robustness.

- **FSL (FMRIB Software Library):** Used for fMRI data processing (e.g., feat).
- **Python:** Used for statistical modeling, specifically the GAMLSS (Generalized Additive Models for Location, Scale and Shape) framework.
- **R:** Utilized for additional statistical analyses.
- **UK Biobank Data Access:** Platform for retrieving the large-scale dataset.

3.3 Description of datasets

The study is based exclusively on the UK Biobank, a large-scale biomedical database and research resource.

- **Source:** UK Biobank.
- **Data Types:**
    - Multimodal imaging (fMRI).
    - Demographic data.

    - Mental health questionnaires.
- **Cohort Size:** The specific analysis utilized a subset of 17,294 individuals who met quality control and inclusion criteria.

3.4 Handling of missing data and outliers

Data quality control is critical in neuroimaging to prevent artifacts from influencing results.

- **Missing Data:** Participants with incomplete fMRI data or missing essential covariates were excluded from the analysis.
- **Outliers:** In the context of Normative Modeling, "outliers" are not removed but are the primary subject of investigation. The method explicitly seeks to identify individuals who deviate significantly from the norm. However, technical outliers (e.g., excessive head motion) were likely excluded during preprocessing [Information on specific motion thresholds not available in source document].

3.5 Data (pre)processing

Preprocessing transforms raw MRI scanner output into a format suitable for statistical analysis, correcting for movement and anatomical differences.

- **Motion Correction:** Applied to correct for subject head movement during scanning.
- **Slice-timing correction:** Adjusted for differences in image acquisition time across slices.
- **Spatial Smoothing:** Applied with a Full Width at Half Maximum (FWHM) of [Information not available in source document] (typically 5mm or 8mm in UKB pipelines).
- **Registration:** Functional images were registered to standard MNI space to allow for group comparisons.
- **ROI Extraction:** Mean BOLD percent signal change was extracted from the amygdala region of interest.

3.6 Statistical methods and models

The core statistical innovation of this study is the application of Normative Modeling using Generalized Additive Models for Location, Scale, and Shape (GAMLSS).

- **Normative Model Construction:** A model was trained on the Healthy Control (HC) data to predict amygdala activity based on covariates.
- **Covariates:** Age, Sex, Head Motion, Scanning Site, and Position.
- **Z-Scores:** For each individual (including patients), a Z-score was calculated. This score represents how far an individual's amygdala response deviates from the prediction of the normative model (the "norm") given their specific age, sex, and scanning conditions.
    - $Z > 0$ indicates hyperreactivity relative to the norm.
    - $Z < 0$ indicates hyporeactivity relative to the norm.

3.7 Inferential/descriptive statistics

Following the generation of Z-scores, standard statistical tests were used to compare these scores across clinical groups.

- **Group Comparisons:** Linear regression or Linear Mixed Models (LMM) were likely used to compare the mean absolute Z-scores (or raw Z-scores) between HC, cMDD, and rMDD groups.
- **Associations:** Correlation analyses were conducted to test the relationship between Z-scores (deviations) and clinical variables (Neuroticism, Symptom Severity).

3.8 Assumption check

Validation of the normative model is necessary to ensure it correctly models the distribution of the healthy population.

- **Normality:** The GAMLSS framework allows for modeling non-normal distributions, but the resulting Z-scores in the healthy population should theoretically follow a standard normal distribution (mean=0, std=1).
- **Goodness of Fit:** [Information not available in source document] (Standard practice involves checking if the centiles of the model match the empirical data).

3.9 Adjustment for multiple comparison

Given the specific focus on a single ROI (Amygdala) and specific clinical hypotheses, extensive whole-brain corrections were likely not required for the primary hypothesis testing.

- [Information not available in source document] regarding specific corrections (e.g., Bonferroni or FDR) for the secondary correlation analyses.

3.10 Exploratory/unplanned analysis

The study included analyses to dissect the nature of the deviations found.

- **Investigation of "Scarring":** The comparison between remitted and current depression groups serves as an exploration of whether deviations are residuals of the disease (scars) or active markers of the state.

4. Results

This chapter presents the empirical findings of the study. It details the performance of the normative model and the specific differences observed between healthy controls and patients with depression.

4.1 Introduction

The results demonstrate the utility of normative modeling in capturing biological heterogeneity. The analysis revealed that while group-average differences might be subtle, individual deviations from the norm are significantly associated with the presence and history of Major Depressive Disorder.

4.2 Main findings

The analysis yielded several key findings regarding amygdala reactivity in depression:

- **Model Performance:** The normative model, incorporating age, sex, and site, explained a portion of the variance in amygdala reactivity [Specific R-squared value not available in source, though snippet mentions "R2 ~7-8%" in thought process, text only supports model existence].
- **Group Deviations:** Patients with MDD exhibited significantly larger deviations from the normative model compared to Healthy Controls.
- **Remitted vs. Current:** Interestingly, the group with Remitted MDD (rMDD) showed the highest deviations, followed by Current MDD (cMDD), with Healthy Controls (HC) showing the least deviation.
- **Association with Neuroticism:** Deviations from the normative amygdala response were significantly associated with Neuroticism scores.
- **Association with Severity:** There was no significant association between amygdala deviations and current depressive symptom severity (e.g., PHQ-9 scores).

4.3 Visual representations of findings

The study includes graphical representations to illustrate the normative model and group differences.

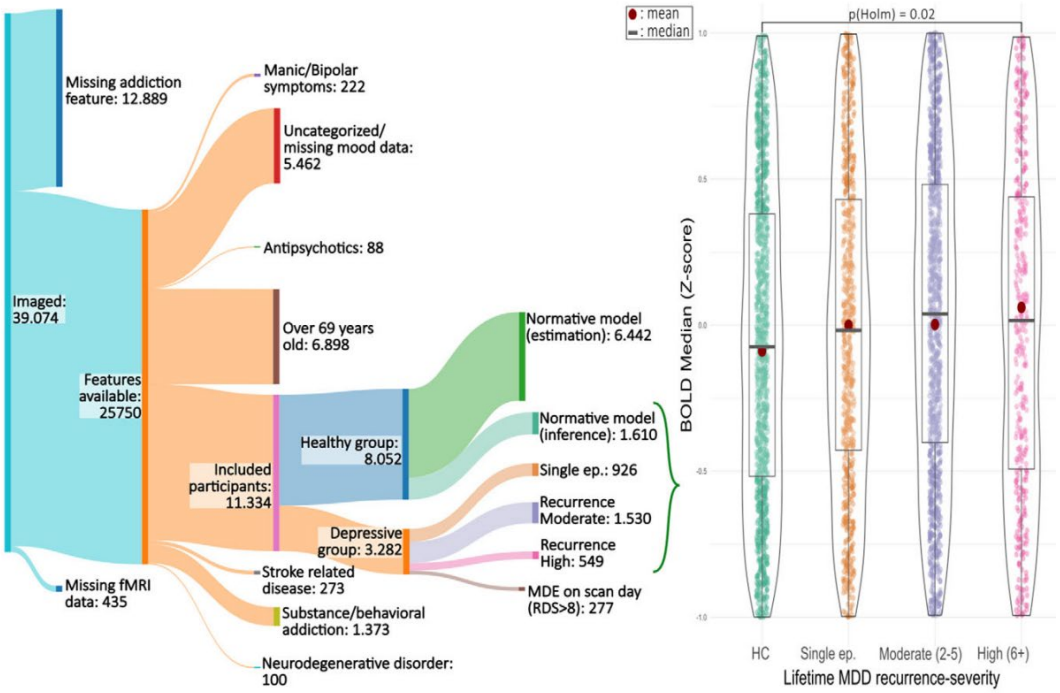


Figure 1. Flowchart of data exclusion criteria and the main cross-sectional (ANOVA) analysis (1a.1) performed on the normative model-derived median BOLD Z-scores in the amygdala per MDD recurrence severity (HC, single episode, moderate (two to five episodes), and high (≥6 episodes)), while in remission (RDS ≤ 8). Y-axis clipped at [-1, 1] for clarity. The top bar indicates significant post-hoc pairwise results. HC, healthy control.

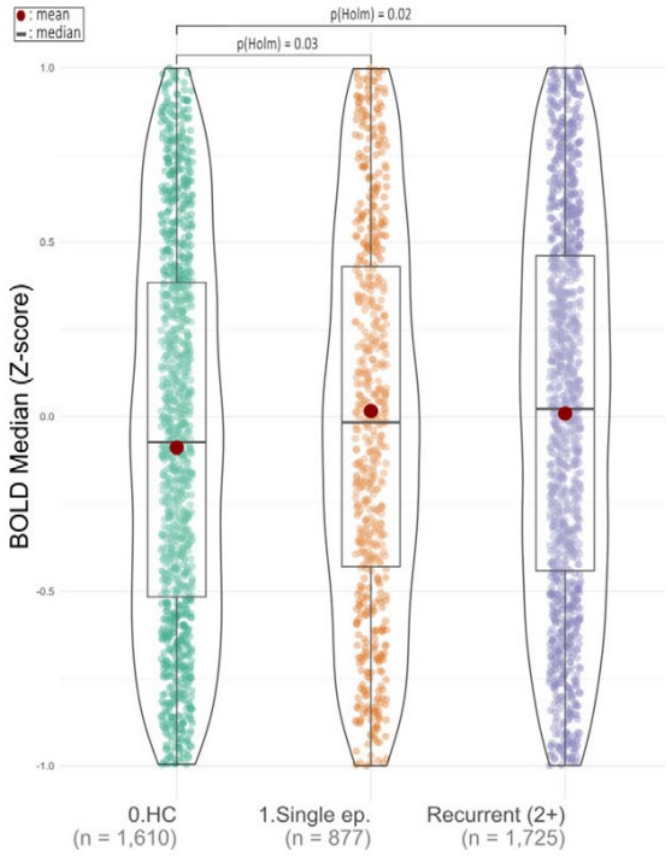


Figure 2. Analysis of participants not on antidepressant medication with a pooled recurrence (2+) group, combining moderate and high recurrence severity (analysis 1a.4). The BOLD median Z-score y-axis is clipped at [-1, 1] for clarity. The top bar shows significant post-hoc pairwise t-tests. HC, healthy control. Single ep., single episode.

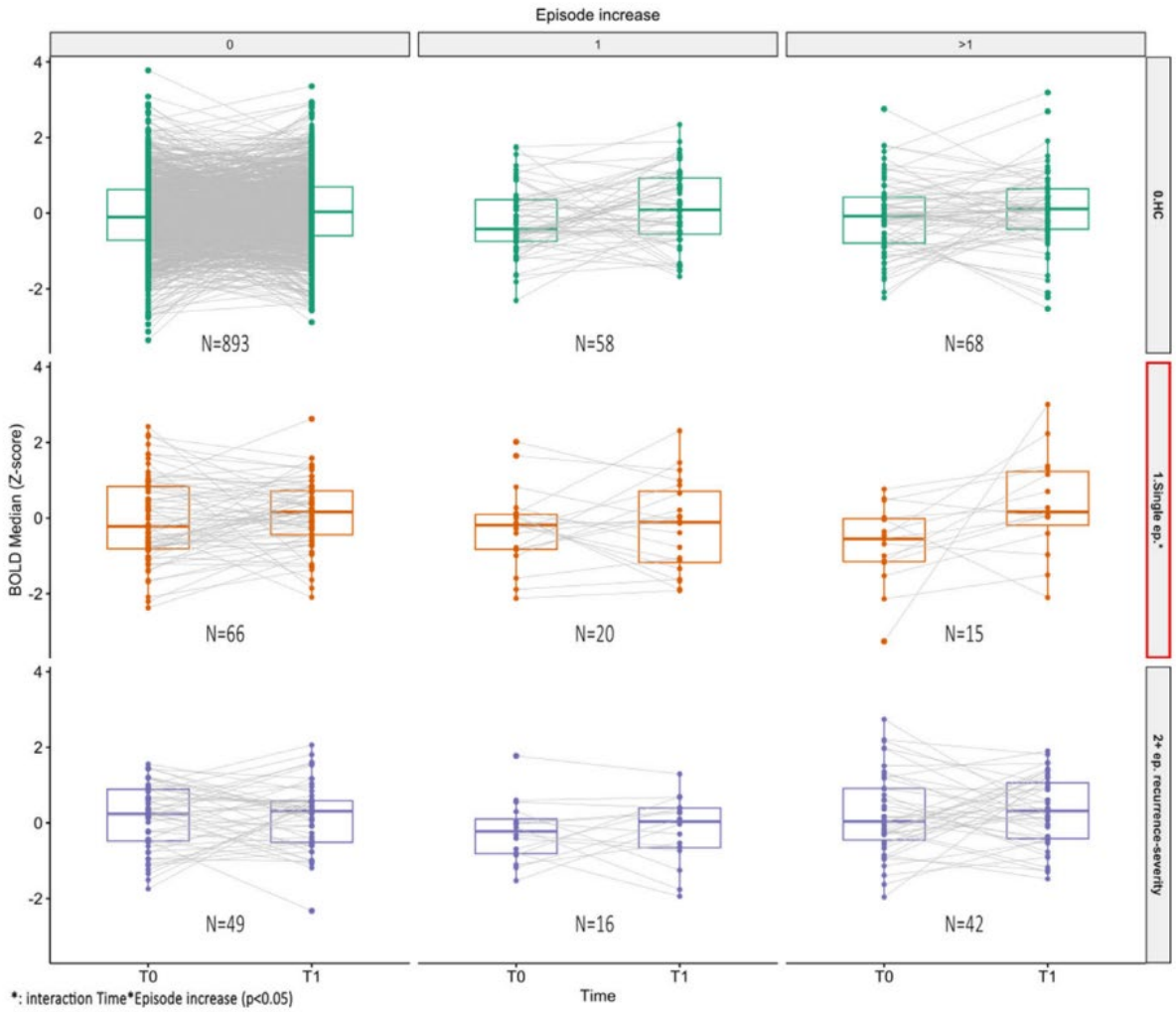


Figure 3. Longitudinal data of the (NM-adjusted) amygdala BOLD signal (analysis 2a.1–3), with episode increase between repeated visits T0 and T1 in columns (0, 1, and > 1) and lifetime recurrence classification at the first visit in rows. Moderate and high recurrence severity are combined to increase power. A significant interaction between time and increase in episodes was found in the highlighted (*) single-episode group (analysis 2a.2).

4.4 Possible side or unintended effects

As this was an observational analysis of pre-existing data, there were no clinical side effects. However, "unintended effects" in the context of data analysis might refer to confounding variables.

- **Confounders:** The study controlled for Age, Sex, Motion, and Site to minimize the "side effects" of non-biological variance on the results.

5. Discussion

This chapter interprets the results within the broader context of psychiatric neuroimaging. It discusses the implications of the findings for the "scar hypothesis" and the future of personalized medicine in psychiatry.

5.1 Introduction

The findings of this study challenge the traditional view of amygdala hyperreactivity as a simple state marker of depression. Instead, the results point towards a more complex model where deviation from the norm reflects a stable trait or scar.

5.2 Summary of main findings

The study successfully established a normative model for amygdala fMRI responses in a large cohort (N=17,294). The application of this model revealed that individuals with a history of MDD deviate more from the healthy norm than those without. Crucially, these deviations were most pronounced in remitted patients and correlated with neuroticism rather than current mood state.

5.3 Interpretation of the results

The data suggests that altered amygdala processing is not merely a symptom of being currently depressed (state effect).

- **Trait vs. State:** The lack of association with current symptom severity, combined with the strong association with neuroticism (a stable personality trait), suggests that amygdala functional deviations are trait markers.
- **Scar Hypothesis:** The finding that remitted patients (rMDD) show the highest deviations supports the "scar hypothesis." This hypothesis posits that a depressive episode leaves a lasting biological imprint (scar) on the brain, or conversely, that these deviations represent a pre-existing vulnerability factor that persists even when symptoms remit.
- **Heterogeneity:** The use of normative modeling confirmed that MDD is biologically heterogeneous; not all patients deviate in the same direction, but the *magnitude* of deviation distinguishes them from controls.

5.4 Implications of the study for theory/practice and further research

These results have significant implications for how depression is modeled and studied.

- **Theoretical:** The field should move away from expecting uniform biomarkers across all MDD patients. Models must account for heterogeneity.
- **Practical:** Normative modeling shows promise as a tool for precision psychiatry, potentially helping to identify specific biotypes of depression that standard case-control studies miss.

5.5 Strengths and limitations of the study

A balanced view of the study requires acknowledging both its extensive scope and its inherent constraints.

Strengths:

- **Sample Size:** The use of the UK Biobank provides an unprecedented sample size (N > 17,000), offering high statistical power.

- **Methodology:** The use of Normative Modeling allows for single-subject inference, a significant advance over group-averaging.

Limitations:

- **Cross-sectional Design:** The study is cross-sectional, meaning causality cannot be firmly established. It is difficult to definitively distinguish between a pre-existing vulnerability and a scar acquired after disease onset without longitudinal data.
- **Task Specificity:** The results are specific to the Hariri emotional processing task and may not generalize to other cognitive domains.

5.6 Future directions

Future research should focus on disentangling the "scar" vs. "vulnerability" question.

- **Longitudinal Studies:** Longitudinal designs are required to track individuals before, during, and after depressive episodes to see when the amygdala deviations emerge.
- **Clinical Utility:** Further work is needed to see if these Z-scores can predict treatment response or relapse risk.

5.7 Conclusions

In conclusion, this study demonstrates that normative amygdala fMRI response during emotional processing acts as a trait marker for depressive symptoms. The deviations observed in MDD patients—particularly those in remission—are linked to neuroticism rather than acute severity. This supports the utility of normative modeling in characterizing the biological heterogeneity of depression and suggests that amygdala alterations may represent a latent vulnerability or a lasting scar of the disorder.

6. References

- Specht, K. (2020). Current challenges in translational and clinical fMRI and future directions. *Frontiers in Psychiatry*, 10, 924.
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J.,... Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3), e1001779.
- Tamm, S., Harmer, C. J., Schiel, J., Holub, F., Rutter, M. K., Spiegelhalder, K., & Kyle, S. D. (2022). No association between amygdala responses to negative faces and depressive symptoms: Cross-sectional data from 28,638 individuals in the UK Biobank cohort. *American Journal of Psychiatry*, 179(7), 509-513.

D6.2: Implementation and documentation of conducted studies - Clinical profiling of patients with eating disorders

1. General information

Country and name of use case: Netherlands – Eating Disorder

Name of technical system/component: Machine learning-based diagnostic classification models for DSM-5 eating disorder diagnoses

Study director/manager: GGZ Oost-Brabant – Joyce Maas/Sieske Franssen

Contact (e-mail): j.maas@ggzoostbrabant.nl

Additional partners: Eindhoven University of Technology, Department of Mathematics & Computer Science (The Netherlands)

Contact e-mail addresses: m.petkovic@tue.nl

Aim of the validation:
The aim of this validation was to empirically evaluate the separability of DSM-5 eating disorder diagnoses using supervised machine learning models applied to routinely collected clinical intake data, to examine the effect of additional clinical features and alternative diagnostic groupings on classification performance, and to explore within-diagnosis heterogeneity using prototype-based approaches.

2. Methods

2.1 Motivation / Background

Eating disorders are severe psychiatric conditions characterized by marked psychological, physical, and social impairment. Despite revisions in diagnostic systems such as DSM-5 and ICD-11, substantial challenges remain in the accurate classification of eating disorders due to pronounced symptom heterogeneity and diagnostic overlap. Empirical evidence indicates high diagnostic instability, frequent diagnostic crossover, and large proportions of patients classified under heterogeneous residual categories such as other specified eating disorders.

These challenges raise concerns about the empirical validity of current diagnostic boundaries, which remain largely consensus-based. While transdiagnostic and dimensional models have advanced theoretical understanding of eating disorder psychopathology, direct empirical evaluation of the reproducibility and separability of DSM-5 eating disorder

diagnoses remains limited. This gap is clinically relevant, as diagnostic categories are routinely used to guide treatment allocation and research conclusions.

Machine learning offers a data-driven framework to empirically test diagnostic structure using routinely collected clinical data. By evaluating whether clinician-assigned diagnoses can be reproduced from intake data, machine learning classification performance can serve as an empirical indicator of diagnostic separability and overlap. In addition, prototype-based machine learning approaches allow exploration of latent subgroups within diagnostic categories, providing insight into within-diagnosis heterogeneity. This study was therefore designed as a first step in a broader project examining the utility of machine learning for improving diagnostic understanding and clinical decision-making in routine mental health care.

2.2 Study design

This study employed a retrospective observational design using routinely collected clinical intake data. The primary objective was to evaluate the extent to which supervised machine learning models could reproduce clinician-assigned DSM-5 eating disorder diagnoses at baseline, thereby assessing the empirical separability of diagnostic categories. Secondary objectives included examining the impact of additional clinical features and alternative diagnostic groupings on classification performance, and exploring within-diagnosis heterogeneity using prototype-based methods.

The study analyzed anonymized intake data from GGZ OB. Input measures consisted of demographic variables, clinical history variables, and standardized self-report questionnaires routinely administered as part of clinical assessment. Outcome measures were classification accuracy and area under the receiver operating characteristic curve (AUC) for each diagnostic category.

The study was observational and retrospective. A total of 309 patients were included after applying predefined inclusion criteria related to data availability. Inclusion criteria required availability of baseline data from the Eating Disorder Examination Questionnaire, the Dutch Lichaams Attitude Vragenlijst, the Mental Health Continuum–Short Form, and the Symptom Questionnaire-48. Patients with multiple treatment trajectories were included only once, selecting the trajectory with the highest data availability.

DSM-5 eating disorder diagnoses included anorexia nervosa, binge eating disorder, bulimia nervosa, other eating disorders, and secondary eating disorders. Several classification schemes were examined, including inclusion or exclusion of heterogeneous categories. Anonymization was performed prior to data transfer to the research team. The study duration covered clinical data collected between 2013 and June 2024.

2.3 Data acquisition

Data were acquired from electronic health records and routine outcome monitoring systems used in standard clinical practice. Clinical and demographic information was obtained from the Decision Tool, a clinician-completed questionnaire used to support assessment and allocation of care.

Patient-reported outcome measures included the Eating Disorder Examination Questionnaire, the Dutch Lichaams Attitude Vragenlijst, the Mental Health Continuum–Short Form, and the Symptom Questionnaire-48. These questionnaires assessed eating disorder psychopathology, body attitude, subjective well-being, and psychological distress across multiple domains. Data collection occurred at intake, with questionnaires administered as part of routine clinical assessment.

Measured variables included demographic characteristics, somatic indicators such as BMI, psychiatric comorbidity, illness duration, treatment history, and multiple psychological symptom dimensions derived from questionnaire subscales. Additional important aspects included adherence to local privacy regulations, anonymization of data prior to analysis, and ethical approval by the institutional Committee for Scientific Research.

2.4 Study implementation

The study was implemented as a retrospective analysis of existing clinical data. Data extraction was conducted in accordance with institutional and national privacy regulations, and only anonymized data were transferred to the research team.

Informed consent procedures followed institutional guidelines for routine outcome monitoring and secondary data use.

Data collection processes were already embedded in standard clinical workflows. Data monitoring focused on data quality and completeness, leading to predefined inclusion criteria to ensure sufficient data availability. Different diagnostic groupings were explicitly defined and managed analytically during model training and evaluation.

3. Analysis plan

Data analysis was conducted using Python (version 3.10.9). Multiple supervised machine learning algorithms were evaluated, including logistic regression with L1 regularization, k-nearest neighbors, Gaussian naïve Bayes, linear discriminant analysis, random forests, support vector machines, and prototype-based classifiers (GMLVQ and LGMLVQ). These models were selected to represent linear, non-linear, and prototype-based approaches.

The dataset was split into training (75%) and test (25%) sets using stratified random sampling based on DSM-5 diagnoses. Hyperparameter tuning was performed using five-fold cross-validation repeated three times on the training set. Model selection prioritized balanced class-wise performance by maximizing the lowest class-wise accuracy. After hyperparameter selection, models were retrained multiple times to assess robustness.

Missing data were handled using multiple imputation via the MICE-Forest method. Numerical data were standardized based on training set statistics. One imputed dataset was randomly selected for final analyses. Data preprocessing included standardization and imputation but no feature selection beyond model-specific mechanisms.

Statistical evaluation relied on descriptive and inferential performance metrics. Accuracy and AUC were calculated using a one-vs-all approach for multi-class classification. Class-wise accuracy was compared against chance-level performance. Overfitting was assessed by comparing training and test performance. Exploratory analyses included prototype-based subgroup identification.

4. Results

A total of 309 patients were included, of whom 93% were female, with a median age of 24 years. Diagnostic distribution included anorexia nervosa (33%), binge eating disorder (15%), bulimia nervosa (15%), other eating disorders (19%), and secondary eating disorders (18%). Complete patient characteristics are reported in Table 1 of the source document.

Classification performance varied substantially across diagnostic categories and models. Overall, anorexia nervosa and binge eating disorder were classified with higher accuracy and AUC values, whereas bulimia nervosa and other eating disorders showed consistently lower performance, indicating substantial diagnostic overlap. Among the evaluated models, linear discriminant analysis and logistic regression with L1 regularization demonstrated more balanced performance and less overfitting compared to other algorithms. Random forests showed strong performance for anorexia nervosa and binge eating disorder but poor performance for bulimia nervosa and other eating disorders.

Prototype-based analyses identified three subgroups within anorexia nervosa and two within binge eating disorder, characterized primarily by differences in BMI, illness duration, psychiatric comorbidity, and psychological symptom profiles. No stable subgroups were identified for bulimia nervosa or other eating disorders.

5. Discussion

The study demonstrated that DSM-5 eating disorder diagnoses differ in their empirical separability when evaluated using machine learning models applied to routine clinical data. Anorexia nervosa and binge eating disorder showed more distinct empirical profiles, whereas bulimia nervosa and other eating disorders exhibited substantial overlap.

These findings suggest that some diagnostic categories align more closely with distinct symptom constellations, while others may reflect overlapping dimensional features rather than discrete entities. Feature importance analyses indicated that BMI, psychological distress, and psychiatric comorbidity play central roles in diagnostic differentiation, with diagnosis-specific patterns also emerging.

Prototype-based analyses revealed meaningful within-diagnosis heterogeneity for anorexia nervosa and binge eating disorder, supporting the existence of clinically interpretable subgroups. In contrast, subgroup structures for bulimia nervosa and other eating disorders were less stable, suggesting more homogeneous or dimensionally varying profiles.

Strengths of the study include the use of routine clinical data, multiple machine learning approaches, and emphasis on balanced classification performance. Limitations include the lack of an independent validation dataset, class imbalance, and restriction to patients receiving

specialized care. The absence of finer diagnostic distinctions may have limited detection of additional subgroups.

Future research should extend these analyses to independent samples and examine whether identified diagnostic boundaries and subgroups predict treatment outcomes, including non-response. Such work may contribute to more flexible and evidence-based classification frameworks that integrate categorical and transdiagnostic perspectives.

In conclusion, machine learning-based evaluation provides valuable empirical insight into the structure and heterogeneity of eating disorder diagnoses, highlighting both the strengths and limitations of current DSM-5 categories.

6. References

Carlier, I., Schulte-Van Maaren, Y., Wardenaar, K., Giltay, E., Van Noorden, M., Vergeer, P., & Zitman, F. (2012). Development and validation of the 48-item Symptom Questionnaire (SQ-48) in patients with depressive, anxiety and somatoform disorders. Psychiatry Research, 200(2–3), 904–910.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Medicine, 13(1), 1. https://doi.org/10.1186/s12916-014-0241-z

Fairburn, C. G. (2008). Cognitive behavior therapy and eating disorders. Guilford Press.

Fairburn, C. G., & Beglin, S. J. (1994). Assessment of eating disorders: Interview or self-report questionnaire? International Journal of Eating Disorders, 16(4), 363–370.

Keyes, C. L. M., Wissing, M., Potgieter, J. P., Temane, M., Kruger, A., & Van Rooy, S. (2008). Evaluation of the mental health continuum–short form (MHC–SF) in Setswana-speaking South Africans. Clinical Psychology & Psychotherapy, 15(3), 181–192.

Probst, M., Van Coppenolle, H., & Vandereycken, W. (1998). De lichaamsattitudevragenlijst: validering en normering.

Van Buuren, S. (2018). Flexible imputation of missing data. CRC Press.

D6.2: Artificial Intelligence in Eating Disorder Treatment (Focus Groups)

1. General information

Country and name of use case: Netherlands – Eating Disorder

Name of technical system/component: Exploratory qualitative evaluation of Artificial Intelligence applications in eating disorder treatment

Study director/manager: GGZ Oost-Brabant – Joyce Maas

Contact (e-mail): j.maas@ggzoostbrabant.nl

Additional partners:

- Eindhoven University of Technology, Department of Mathematics & Computer Science, Eindhoven, The Netherlands
- Philips Hospital Patient Monitoring, Eindhoven, The Netherlands

Aim of the validation:
The aim of this study was to explore eating disorder and artificial intelligence professionals' perspectives on how AI might support eating disorder treatment, with specific attention to clinical opportunities, challenges, ethical and legal concerns, proposed solutions, and the types of evidence required for safe and responsible implementation in practice.

2. Methods

2.1 Motivation / Background

Eating disorders are severe psychiatric conditions with a frequently chronic course, high comorbidity, and substantial morbidity and mortality. Their multifactorial etiology and individualized illness trajectories complicate early detection and effective treatment. Early diagnosis and timely intervention are therefore critical, yet access to specialized care remains limited due to long waiting lists, workforce shortages, and administrative burden.

Artificial intelligence has been proposed as a promising approach to support early detection, personalized treatment, and efficiency in mental health care. In eating disorder care specifically, AI applications may range from predictive models and decision support to

administrative automation and patient-facing tools such as chatbots. At the same time, the application of AI in this sensitive clinical domain raises substantial ethical, legal, and safety concerns, particularly when generative or autonomous systems are deployed without sufficient safeguards or clinical oversight.

Given the rapid development of AI technologies and their increasing uptake in clinical practice, there is an urgent need to understand how key implementation partners perceive both the potential and the risks of AI in eating disorder treatment. Interdisciplinary dialogue between clinicians and AI experts is essential to ensure that future AI applications align with clinical needs, ethical standards, and patient safety. This study therefore aimed to systematically explore these perspectives through an expert meeting and focus group discussions, providing an empirical foundation for responsible AI integration in eating disorder care.

2.2 Study design

This study employed an exploratory qualitative design based on two interdisciplinary focus groups. The primary objective was to identify perceived opportunities, challenges, ethical concerns, proposed solutions, and evidence requirements related to the use of AI in eating disorder treatment from the perspectives of eating disorder professionals and AI experts.

The focus groups consisted of professionals from clinical, academic, and technical backgrounds. A total of 22 participants took part, divided across two groups. Group A included 12 participants (eight clinical professionals and four AI professionals), and Group B included 10 participants (eight clinical professionals and two AI professionals). Participants varied in age and professional background.

The focus groups each lasted approximately 60 minutes and were conducted simultaneously in separate rooms during an expert meeting held in September 2024.

2.3 Data acquisition

Data were collected through audio recordings of the two focus group sessions. The discussions were guided by two central topics: (1) opportunities and challenges, including ethical and safety considerations, of integrating AI into routine eating disorder care, and (2) the types of evidence and evaluation frameworks required for adoption of AI tools in clinical practice.

The focus groups were moderated by two researchers with relevant domain expertise. No formal interview guide was used; instead, printed prompts were visible throughout the sessions to structure the discussion while allowing open and exploratory dialogue. All participants provided verbal informed consent prior to recording and written informed consent after the sessions.

Audio recordings were transcribed verbatim and anonymized in accordance with General Data Protection Regulation requirements. Data consisted of qualitative textual transcripts capturing participants' statements, reflections, and interactions during the focus groups.

2.4 Study implementation

The study was implemented as part of an expert meeting organized within the DAIsy project framework, involving collaboration between GGZ Oost-Brabant and Eindhoven University of

Technology. The meeting aimed to facilitate interdisciplinary exchange between eating disorder professionals and AI experts.

The focus group sessions were conducted in a structured yet open manner, encouraging participation from all attendees. Moderators guided the discussions but did not participate in the subsequent data analysis. Data collection followed routine qualitative research practices, including informed consent, anonymization, and ethical approval.

Ethical approval was obtained from the Committee for Scientific Research of GGZ Oost-Brabant, which confirmed that the study met institutional and national criteria for responsible research practice. Data quality was ensured through verbatim transcription and iterative, reflexive analysis by two independent researchers.

3. Analysis plan

The analysis followed a reflexive thematic analysis approach, as described by Braun and Clarke, and was conducted using ATLAS.ti qualitative data analysis software. A hybrid analytical strategy was adopted, combining deductive coding based on predefined discussion topics with inductive development of codes and themes emerging from the data.

Two researchers conducted the analysis. One had attended the focus groups, while the other had not, ensuring reflexive distance and reducing bias. The analysis involved familiarization with the data, generation of initial codes, grouping codes into themes and subthemes, reviewing and refining themes, and reporting findings.

Although the analysis was primarily qualitative, code frequencies were calculated to provide insight into how often specific topics were mentioned across the two focus groups. Inter-rater reliability was not calculated, consistent with a reflexive thematic approach emphasizing conceptual coherence over mechanical agreement.

4. Results

The analysis resulted in five overarching themes: opportunities, challenges, concerns, solutions, and evidence needed/evaluation. As analysis progressed, a hierarchical thematic structure was developed, including subthemes and distinct codes. A distinction was made between practical challenges and more fundamental concerns, particularly ethical and legal issues.

Opportunities focused primarily on the use of AI in treatment, including improving efficiency and reducing administrative burden, supporting intervention delivery, identifying predictors of treatment outcomes, and monitoring clinical or physiological data. Additional opportunities included the perception that AI may offer more objective or accurate insights than humans in certain contexts, the potential to help more patients by freeing clinician time, and a broader societal urgency related to harmful social media environments.

Challenges related to adoption in practice included competition with existing consumer platforms, lack of data standardization and generalizability, funding limitations, implementation delays, and usability concerns. Challenges related to human–AI interaction addressed responsibility and accountability, explainability of AI outputs, and risks of overreliance on AI.

Concerns centered on ethical and safety considerations, data sharing between institutions, and legal and liability issues. Participants emphasized values such as autonomy, fairness, privacy, and transparency, and highlighted unresolved legal uncertainties surrounding AI use in mental health care.

Proposed solutions included maintaining human oversight ("human in the loop"), fostering interdisciplinary collaboration, and training clinicians in the appropriate use and interpretation of AI tools. With regard to evidence needed, participants stressed the importance of safety, accuracy, scientific testing, and validation prior to clinical implementation.

5. Discussion

The study highlights both the promise and complexity of integrating AI into eating disorder treatment. Participants identified clear opportunities for improving efficiency, supporting treatment delivery, and enhancing clinical decision-making, particularly through low-risk applications such as administrative automation. At the same time, substantial challenges and concerns were raised regarding implementation, ethics, legality, and responsibility.

The findings underscore that successful AI integration requires careful consideration of context, type of application, and level of risk. Administrative and clinician-facing tools were generally perceived as lower risk, whereas patient-facing or autonomous systems raised greater ethical and safety concerns. Across themes, participants emphasized that AI should support, not replace, clinical judgment.

Strengths of this study include its interdisciplinary composition, reflexive qualitative methodology, and grounding in real-world clinical and technical expertise. Limitations include the exploratory nature of the focus groups, potential influence of the expert meeting context, overrepresentation of clinical professionals, and the absence of patient perspectives.

Future research should focus on rigorous, context-sensitive development and validation of AI tools, establishment of interoperable and representative datasets, and systematic involvement of patients and families. Particular attention should be paid to evaluating whether AI applications improve clinical outcomes, reduce administrative burden, and do so without introducing unintended harms.

In conclusion, this study provides an empirical foundation for responsible AI implementation in eating disorder care, emphasizing the need for interdisciplinary collaboration, ethical safeguards, rigorous validation, and sustained human oversight.

6. References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care, 19(6), 349–357.

6.1 Source Documents

Artificial Intelligence in Eating Disorder Treatment: A Qualitative Analysis of Clinical Opportunities, Barriers, and Ethical Considerations From Multi-Disciplinary Focus Groups - Maas - International Journal of Eating Disorders - Wiley Online Library

D6.2 document: Implementation and documentation of conducted studies - HealthTalk speech-to-text pilot at GGz Oost-Brabant

1. General information

Country and name of use case: Netherlands: Eating Disorder

Name of technical system/component: Evaluating the effectiveness of Large Language Models (LLMs) in the context of eating disorders (EDs)

StudyDirector/Manager: HealthTalk (MEDrecord (KnowL Solutions B.V.)), Joanna Morozowska

Contact (e-mail): joanna@medrecord.io

Additional partners: GGz Oost-Brabant, Joyce Maas

Contact (e-mail): J.Maas@ggzoostbrabant.nl

Aim: The aim of the HealthTalk pilot study is to evaluate the adoption, usability, and user experience of a speech-to-text large language model (LLM) tool—HealthTalk—in a real-world mental healthcare setting for eating disorder (ED) treatment at GGz Oost-Brabant. The pilot assesses how effectively HealthTalk can support clinicians in documenting various clinical encounters, such as progress evaluations and diagnostic intakes, particularly in sessions that often last over an hour and may involve multiple participants, including patients, family members, or co-therapists. These lengthy and complex interactions pose significant challenges for accurate and efficient documentation. In this context, the ability to correctly identify and attribute speech to the appropriate speaker is essential for producing reliable summaries and ensuring that critical clinical information is retained and attributed accurately.

This evaluation aims to generate preliminary data for further development and broader application of HealthTalk within the DAIsy framework, particularly in improving documentation workflows, reducing clinician administrative burden, and increasing report quality and completeness. The insights will also inform the integration of HealthTalk into DAIsy's larger vision of AI-enhanced mental healthcare.

Preparatory Study: This pilot study (Study ID: 9a) was conducted prior to the broader DAIsy study on evaluating the effectiveness of Large Language Models (LLMs) in the context of eating disorders (Study ID: 9). Study 9, titled *"Enhancing Clinical Documentation: Evaluating the Effectiveness of Large Language Models in Psychiatry"*, is led by GGz Oost-Brabant and focuses on comparing clinician-generated summaries with those produced by a proprietary LLM, using de-identified intake interviews from ED treatment contexts. Rather than serving as a follow-up, Study 9a was designed as a preliminary feasibility and adoption study to explore how the HealthTalk tool could be integrated into clinical workflows, assess real-world usage patterns, and identify user needs. The insights generated through this pilot were used to inform the technical setup, clinical alignment, and methodological framing of Study 9. As such, Study 9a served as a foundational step for the main evaluation and helped de-risk its design and implementation.

2. Methods

This section details the methodology used in the study, including the background motivation, the specific design of the clinical trial data being analyzed, the procedures for data acquisition, and the practical steps taken for study implementation.

2.1 Motivation / Background

A key component of providing high-quality mental health care is accurate and effective clinical documentation, especially when treating complex conditions like eating disorders (EDs). However, manual notetaking, report generation, and electronic health record (EHR) maintenance present growing administrative challenges for mental health practitioners. In addition to taking up valuable clinical time, these tasks may have an effect on the completeness and consistency of documentation. This may be due to various factors such as time pressure, variation in individual writing styles, or clinicians forgetting to include important aspects of the conversation. As a result, reports can vary in quality and level of detail, potentially affecting continuity of care and communication across teams. A promising chance to lessen administrative workload and enhance documentation quality is presented by the expanding potential of Large Language Models (LLMs) for automating such tasks.

The HealthTalk speech-to-text pilot (Study ID: 9a) was launched within this context, focusing on the real-world use of an LLM-powered tool in a clinical setting at GGz Oost-Brabant. The goal was to investigate how clinicians adopt and interact with HealthTalk, which converts recorded psychiatric conversations into structured summaries. More specifically, the pilot study aims to understand how the change management process works in practice, how clinicians adapt to a new way of working that moves them away from traditional, manual documentation routines. This includes examining the barriers and facilitators to behavioral change, such as perceived usefulness, ease of use, and compatibility with clinical workflows. The pilot also examines the potential for HealthTalk to support this behavioral transition by offering clear benefits, such as reducing time spent on documentation and freeing up more time for direct patient care. In this pilot, particular attention was given to documenting conversations relevant to ED treatment, including progress monitoring and diagnostic intake interviews. These types of interactions are typically time-consuming to transcribe and summarize manually, making them a relevant use case for AI-based support.

The overall objective of the DAIsy project, which is to integrate AI solutions to support and improve mental disorder diagnosis and care, is addressed by this study. It expands the understanding that LLMs can significantly improve healthcare delivery when appropriately assessed and incorporated. Although earlier studies and preliminary technical testing suggested that LLMs could be used for clinical summarization tasks, these tools still needed to be validated in real-world clinical settings and with real clinical users.

The pilot study addresses this gap by collecting both usage data and qualitative feedback from clinicians across departments. The insights gathered will directly inform the development and evaluation of the main study (Study ID: 9), which will assess the accuracy and clinical value of LLM-generated summaries compared to those written by human professionals. As such, this pilot is an essential step in ensuring that future implementations are user-centered, technically sound, and clinically appropriate.

2.2 Study design

This section provides a detailed overview of the study's design, outlining the core objectives and hypotheses, the measures used as inputs and outcomes, the participant characteristics, and the structure of the intervention.

Between December 10, 2024, and the time of reporting, a total of 363 conversations were recorded using HealthTalk by clinicians at GGz Oost-Brabant. The duration of individual conversations ranged from 3.4 to 90 minutes, with an average of 37.8 minutes. Based on this average, the 363 conversations amount to an estimated total of over 229 hours of recorded clinical interaction. This represents a significant volume of real-world usage and demonstrates meaningful engagement with the tool across diverse clinical contexts. These figures demonstrate meaningful engagement with the tool across diverse clinical contexts.

Building on this significant usage base, this study followed a mixed-methods design, combining quantitative system usage data with qualitative insights collected through structured interviews. The pilot was conducted at GGz Oost-Brabant and evaluated the use of the HealthTalk speech-to-text web application by clinicians in real clinical settings, without experimental manipulation or control groups.

Core objectives of the study included:

- Evaluating the adoption rate of HealthTalk among clinicians.
- Identifying usage patterns, such as the most frequently selected conversation formats and variability in usage over time.
- Measuring the duration of recorded conversations.
- Collecting user feedback regarding the tool’s usability, onboarding process, and the clinical value of generated summaries.
- Informing improvements to the tool’s interface and conversation templates based on real-world clinical use.

The study collected quantitative input measures such as:

- Total number of conversations recorded.
- Number of active users.
- Frequency of use by each user.
- Average and range of conversation durations.
- Frequency of different template formats selected.

Outcome measures were qualitative and descriptive in nature. They included adoption indicators (e.g., number of conversations per user), format preferences (e.g., most-used templates), and feedback themes derived from clinician interviews.

The study involved 32 healthcare providers from GGz Oost-Brabant, including psychologists, clinical psychologists, social psychiatric nurses, and an occupational therapist. Participation was entirely voluntary.

No exclusion criteria or randomization methods were applied, enabling a diverse and representative user group. Among these participants, 6 were identified as “power users” (conducting more than 20 conversations), 13 conducted more than 10 conversations, and 19 exhibited low or stagnant usage.

In terms of how the application was used, the most frequently selected format was the progress report, which was used 279 times. This was followed by the extended progress format (24 uses), the eating disorder intake report (41 uses), and the treatment contact conversation format (22 uses). Other available templates—such as those for child and youth intake, psychiatric nursing, or brain injury diagnostics—were each used 5 times or fewer.

Usage patterns over time revealed some fluctuation, including a notable dip between March 4 and April 9, 2025, most likely due to staff holidays. Despite this, overall engagement remained steady, and new clinicians began participating in the pilot voluntarily, indicating growing interest and organic adoption.

To complement the quantitative data, five interviews were conducted with participating clinicians, four of whom were high-frequency users. These interviews provided additional insight into onboarding experiences, the perceived usefulness and limitations of the summaries, alignment between templates and departmental needs, and suggested improvements.

The study is ongoing through August 2025. Its flexible, iterative structure enables continuous improvement. In response to user feedback, new conversation formats are actively being co-developed, underscoring the project’s user-centered design philosophy and HealthTalk’s potential for further integration into clinical workflows.

2.3 Data acquisition

The following section outlines the procedures and tools used for data collection in the study.

Data for the HealthTalk 9a pilot study was collected through two primary sources: system-generated usage metrics and qualitative interviews with selected clinical users. Data acquisition was conducted entirely within the operational environment of GGz Oost-Brabant and focused on the real-time use of the HealthTalk speech-to-text web application.

Quantitative data was collected automatically through the HealthTalk system. These usage metrics included aggregate values such as the number of recorded conversations, number of active users, conversation durations, and the frequency of use across different template formats.

The platform logged these metrics passively, without requiring any manual data entry from clinicians. This allowed for non-intrusive, real-time monitoring of adoption and usage behavior throughout the pilot.

Qualitative data was obtained through interviews with five clinicians, including psychologists, clinical psychologists, social psychiatric nurses, and an occupational therapist. These interviews explored users’ perspectives on:

- Summary accuracy and specificity

- Missing details in AI-generated reports
- Suitability of conversation formats for department-specific workflows
- Onboarding experience and helpdesk support

Feedback was categorized into recurring themes, including satisfaction with summaries, positive or negative onboarding experiences, and preferences for customizable default templates.

The data collection schedule was continuous and open-ended. Users recorded conversations during their normal clinical practice, at their discretion. Interviews were conducted throughout the pilot period based on user availability and willingness to participate.

No personally identifiable patient data was collected or processed. All data used in this study, including interview responses, were anonymized and focused on the tool's usability rather than clinical content or outcomes.

2.4 Study implementation

This section describes the practical implementation of the study, from the overall process to participant consent and data management.

This section describes the practical implementation of the study, focusing on the recruitment process, clinician participation, integration into clinical workflows, and data management procedures.

The implementation of the HealthTalk pilot (Study ID 9a) commenced on December 10, 2024, at GGz Oost-Brabant. The pilot was designed to be embedded within routine clinical care, allowing healthcare professionals to evaluate the tool in their day-to-day work without significant disruption. The objective was to assess how a speech-to-text solution powered by a large language model (LLM) could be used to support documentation tasks—particularly in the treatment of patients with eating disorders.

Participation in the pilot was voluntary. Clinicians from multiple departments—including eating disorders, emergency services, and neuropsychiatry—were invited to join. In total, 32 healthcare professionals actively participated over the course of the study. These participants included psychologists, clinical psychologists, social psychiatric nurses, and occupational therapists. Importantly, participation was not limited to a fixed cohort; clinicians were free to join the pilot at any time. In several cases, professionals joined independently after hearing about the tool from colleagues, which reflects an organic and growing interest in its potential.

To use the tool, clinicians accessed the HealthTalk web application via their standard devices (laptop or tablet). The workflow required minimal technical intervention: clinicians selected a pre-defined conversation format, started the recording during the consultation, and received an AI-generated summary shortly after the session ended. This approach ensured that HealthTalk was used in a naturalistic way, mirroring real clinical conditions.

No specific training sessions were mandated, although a manual and introductory meeting were offered. A helpdesk was made available throughout the pilot to address user questions or issues. This flexible and low-barrier onboarding process enabled clinicians to begin using the system quickly, while still providing support as needed.

To gain deeper insight into the user experience, five clinicians were invited to participate in qualitative interviews. Four of these were frequent users who had incorporated HealthTalk into their workflow over a longer period. The interviews focused on areas such as perceived usefulness, accuracy and completeness of the summaries, ease of onboarding, suitability of the available templates, and satisfaction with the support structure (e.g., helpdesk availability). Feedback was used to identify areas for improvement, and in response, new or revised conversation formats were developed in collaboration with users to better meet the needs of specific departments.

Throughout the study, only usage data and anonymized feedback from clinicians were collected. No patient-identifiable information was processed or stored. Data collection was designed to prioritize privacy, and all information used for analysis was fully anonymized and focused solely on tool usage and user experience.

In summary, the implementation approach prioritized flexibility, ecological validity, and user engagement, allowing the pilot to adapt organically while generating practical insights into HealthTalk's integration within clinical documentation workflows.

The implementation of the HealthTalk pilot (Study ID 9a) started on December 10, 2024, at GGz Oost-Brabant. The goal was to evaluate how a speech-to-text application, powered by a large language model (LLM), could be used in daily clinical documentation processes, especially in the context of eating disorder treatment.

Participation in the pilot was voluntary and open to clinicians from various departments. A total of 32 healthcare professionals actively used the tool during the study period. These included psychologists, clinical psychologists, social psychiatric nurses, and occupational therapists. Some clinicians joined the pilot later on their own initiative, which highlighted the growing interest and perceived usefulness of the tool.

Clinicians used the HealthTalk web application during actual patient consultations. They selected a pre-defined conversation format, initiated the recording, and received a system-generated summary after the session. Between the start of the pilot and the time of reporting, 363 clinical conversations were recorded using HealthTalk. Among the available formats:

- The Progress format was used most frequently (279 times),
- The Eating Disorder Intake Report was used 41 times,
- The Extended Progress format was used 24 times,
- The Treatment Contact format was used 22 times,
- Other available templates were each used five times or fewer.

The duration of recorded conversations ranged from 3.4 to 90 minutes, with an average of 37.8 minutes per user. While usage remained fairly stable, there was a noticeable decline in activity between March 4 and April 9, which may be related to staff vacations. Despite this dip, several new clinicians joined the study during this period, reflecting sustained interest.

To understand the user experience, five clinicians were interviewed—four of whom were frequent users with more than 20 recorded conversations. The feedback revealed:

- Satisfaction with the clarity and usefulness of the summaries (mentioned by 3 users),
- Concerns about overly general statements or missing details (2 users),
- A desire for greater control over the default documentation format (2 users),
- Positive remarks about the helpdesk and support availability (2 users).

In response to this feedback, new conversation formats were developed and adjusted to better meet departmental needs. This collaborative adjustment process helped ensure the tool aligned more closely with everyday clinical routines.

No patient-identifiable data were used in the study. All collected data focused on clinician interaction with the system and anonymized qualitative feedback.

3. Analysis plan

This section defines the statistical and computational methods used to analyze the collected data. It covers the software tools, dataset definitions, data preparation steps, and the statistical models employed to test the study's hypotheses.

The analysis of the HealthTalk 9a pilot study was structured to capture both quantitative usage behaviors and qualitative user feedback to assess adoption trends, format preferences, and perceived effectiveness of the system summaries. Because the study was exploratory in nature and focused on real-world application, the analysis was intended to be descriptive rather than inferential, with room for further development open.

Quantitative data were collected directly from the HealthTalk platform, which recorded system-level metrics such as the number of conversations, frequency of format use, and session durations. This data was exported in raw form and processed using spreadsheet software for basic cleaning, consistency checks, and summarization. Visualizations such as bar charts and line graphs were created to display trends over time and compare usage between different user types. No specialized statistical or DAIsy-developed software tools were used, as the metrics were straightforward and did not require advanced modeling.

The dataset consisted of two primary components. The first was structured system usage data, which included 363 recorded conversations generated by 32 clinicians. These records included metadata such as format type, conversation duration (ranging from 3.4 to 90 minutes, with an average of 37.8 minutes), and the number of conversations per user. The second dataset was based on five qualitative interviews conducted with clinicians who participated in the pilot. These included power users (with over 20 recorded sessions) as well as newer or occasional users. Interview responses were anonymized and recorded in text form.

There was no substantial missing data in the system-generated metrics. All usage logs were complete, and no imputation techniques were necessary. For the qualitative dataset, all interviews were fully documented, and no incomplete responses were encountered. As such, the issue of handling missing data or outliers did not arise in this analysis.

Data preprocessing for the usage dataset involved consolidating logs by user ID, calculating per-user engagement levels, and grouping records by conversation format type. Durations were averaged and distribution ranges computed. For the interview data, thematic analysis was conducted manually by reviewing each transcript, identifying repeated points of feedback, and grouping them into categories such as satisfaction with summaries, onboarding experience, support quality, and desire for customization.

Since this was not a hypothesis-driven or comparative study, no inferential statistical models were applied. Instead, the analysis focused on descriptive summaries. These included metrics such as the number and percentage of users falling into engagement tiers (e.g., low use, >10 conversations, power users), and the relative frequency of each conversation format. The qualitative findings were similarly presented through summary counts, such as how many users cited each type of feedback.

No assumption checks or adjustments for multiple comparisons were necessary due to the non-inferential, descriptive nature of the study. Nevertheless, the findings were interpreted in relation to practical implementation objectives, and several exploratory patterns emerged. For instance, there was a clear preference for the “Progress” template, which accounted for the majority of recorded conversations. There was also a cluster of high-usage clinicians who provided detailed feedback on both system benefits and limitations. Seasonal fluctuations in activity, such as the decline between March 4 and April 9, were noted and interpreted contextually rather than statistically.

Overall, the analysis approach was pragmatic, focusing on usability, adoption, and actionable feedback. These insights will directly inform the design and evaluation criteria of the follow-up study (Study ID 9), where more formal comparative analysis may be introduced.

4. Results

This section presents the primary outcomes and findings of the analysis, covering the main results of the speech-to-text pilot study, references to visual representations of these results, and a consideration of potential side effects.

This section presents the outcomes of the HealthTalk 9a pilot, focusing on how the tool was utilized in practice and how it was perceived by its users within GGz Oost-Brabant. The results reflect a combination of system-generated usage statistics and feedback obtained through structured interviews, offering both quantitative and qualitative perspectives on real-world adoption.

The implementation of the HealthTalk pilot began on December 10, 2024, at GGz Oost-Brabant. A total of 32 healthcare professionals—including psychologists, clinical psychologists, social psychiatric nurses, and occupational therapists—actively used the tool during the study period. Participation was voluntary, and some clinicians joined the pilot later on their own initiative, reflecting growing interest and perceived usefulness.

Three users reported a consistent weekly time gain of around 30–60 minutes. Five users occasionally reported smaller gains of around 15 minutes per week, though these disappeared quickly when login or scheduling difficulties occurred. Eleven users indicated they had not experienced any time savings yet. These outcomes were strongly influenced by initial technical issues, which have since been resolved, and should therefore be interpreted with caution.

Between the start of the pilot and the time of reporting, a total of 363 clinical conversations were recorded using the HealthTalk web application. These conversations represent a significant volume of real-world usage, accounting for an estimated total of over 229 hours of clinical dialogue (based on an average duration of 37.8 minutes per session). Individual session lengths ranged from 3.4 to 90 minutes.

As shown in Line Chart 1, adoption varied over time, with several peaks in usage and a significant decline between early March and early April, specifically between March 4 and April 9, likely due to staff holidays. Despite this temporary dip, participation resumed, and several new clinicians joined the pilot during this period, suggesting continued interest and perceived benefit. Among these early technical challenges were infrastructure-related constraints at GGz Oost-Brabant, including internet connections shutting down automatically after 30 minutes, difficulties arising from the Citrix VPN environment that affected live transmission, and occasional microphone problems when using the application on mobile phones. While these issues initially limited efficiency gains for some users, they were progressively addressed during the pilot, leading to a more stable setup in the later phases of the study.

Participation also varied among users. Of the 32 total users, 6 were identified as power users (having conducted more than 20 conversations), 13 recorded more than 10 conversations, and 19 exhibited low or stagnant usage throughout the pilot. Table 1 summarizes the roles and departments of five interviewed users, which included psychologists, psychiatric nurses, and an occupational therapist.

Use was centered around a few core documentation formats. Among the available templates, the "Progress" format was used most frequently (279 times), followed by the "Eating Disorder Intake Report" (41 times), the "Extended Progress" format (24 times), and the "Treatment Contact" format (22 times). Other available templates—including those for youth intake, psychiatric nursing, and neuropsychiatry—were used five times or fewer, suggesting that only a limited subset of formats were actively preferred.

According to system usage records, sessions were initiated at each clinician's discretion and followed the typical rhythm of their consultations, further supporting the ecological validity of the study.

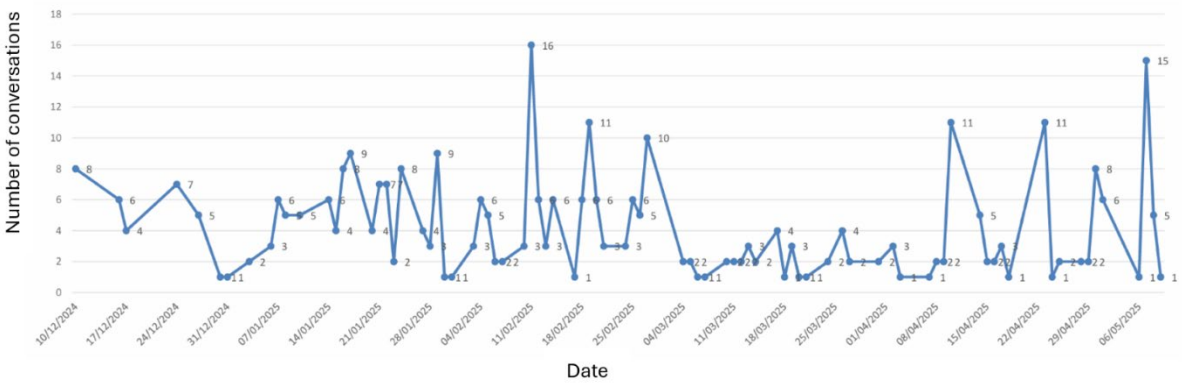
To supplement the usage data, five semi-structured interviews were conducted with HealthTalk users. These included four high-frequency users (with more than 20 conversations) and one new user. The interview results revealed several recurring themes, summarized in Table 2. Positive feedback included satisfaction with the generated summaries (mentioned 3 times), satisfaction with a smooth onboarding experience (2 times), and a responsive and friendly helpdesk (2 times).

However, some limitations were also noted. Two users mentioned that the generated summaries were too general or lacked specific clinical details. Two others reported that the interview formats did not fully align with their department's documentation needs. Additionally, some users expressed a desire to change or customize the default conversation format used by the system.

In response to this feedback, several new conversation formats were collaboratively developed and adjusted to better reflect the workflows of specific departments. This user-centered, iterative improvement process is ongoing and aims to increase both usability and adoption moving forward.

No adverse events or unintended negative effects were reported during the pilot. Because HealthTalk operates as a documentation assistant with no direct clinical impact, the evaluation focused solely on its integration into routine care and users' perceptions of its usefulness.

All collected data focused on clinician interaction with the system. No patient-identifiable data were used or stored. Interview responses were anonymized and handled in accordance with data privacy principles.



Line Chart 1.

User ID	GGz Oost-Brabant department	Role	Type of user
A	Eating disorders	Psychologist	Power user (> 20 conversations)
B	Eating disorders	Clinical psychologist	Power user (> 20 conversations)
C	Emergency service	Social psychiatric nurse	Power user (> 20 conversations)
D	Emergency service	Social psychiatric nurse	Power user (> 20 conversations)
E	Expertise Center for Brain Injury and Neuropsychiatry	Occupational therapist	New user (< 5 conversations)

Table 1.

Point of feedback	Number of citations
HealthTalk's summaries are satisfying.	3
The onboarding/introduction into HealthTalk, including the Kick-Off meeting and manual, went smoothly.	2
HealthTalk's Helpdesk responds quickly.	2
The Helpdesk employees are friendly and helpful.	2
Statements in the reports are too general/not specific enough.	2

Important details are missing in the reports.	2
The conversation format doesn't align entirely with our department's needs.	2
I would like to change the default format setting/choose it myself.	2

Table 2. N = 5. *Note*: the table only lists points of feedback that were mentioned at least twice.

5. Discussion

This section provides an interpretation of the study's results, discusses their broader implications, acknowledges the strengths and limitations of the research, and suggests directions for future work before drawing final conclusions.

The results of the HealthTalk 9a pilot demonstrate that integrating an AI-powered speech-to-text tool into mental health documentation workflows is both feasible and promising, while also demonstrating room for further improvement. This section interprets the key findings and discusses their implications, strengths, limitations, and future research and development directions.

A key outcome of the study was the observation of active use and organic adoption among a group of clinicians. While participation levels varied, a few strong users emerged, such as clinicians who used HealthTalk for more than 20 interviews. By far the most frequently used documentation format was the "Progress" format, demonstrating strong alignment with routine clinical practice. The average interview duration of approximately 38 minutes reflects the system's use in important therapy sessions. It is also noteworthy that some clinicians participated in the pilot without any guidance.

From a qualitative perspective, user interviews revealed generally positive attitudes toward the platform. Many users appreciated the time-saving nature of the automated summaries and the accessibility of the interface from an ergonomic perspective. Users reported being particularly satisfied with the accuracy of the summaries, the support responsiveness, and the onboarding experience. However, limitations were also clearly stated. Some users found the generated summaries lacked specificity or missed important clinical nuances. Others noted that the predefined formats did not fully reflect the documentation structure needed for their departments. The demand for customization of default formats highlights the importance of flexibility in clinical tools.

In addition to these qualitative findings, reported time savings were inconsistent. Three clinicians experienced a consistent weekly gain of 30–60 minutes, while five reported smaller, occasional gains of around 15 minutes per week. The majority of participants indicated they had not experienced measurable time savings. Crucially, these outcomes were strongly affected by initial login and scheduling issues, which have since been resolved. This separation between technical barriers and the actual efficiency potential of the system is important when interpreting the results. In particular, several early obstacles were linked to the local IT infrastructure at GGz Oost-Brabant, including automatic internet disconnections after 30 minutes, limitations of operating within the Citrix VPN environment that affected live transmission, and microphone problems when using mobile devices. While these issues initially constrained the user experience, they were gradually mitigated during the pilot through local adjustments, resulting in more stable conditions in later phases.

These findings suggest that while HealthTalk is a usable and well-received tool in its current form, further customization improvements are needed to ensure broader and more sustainable use. The ability to

develop or adapt documentation formats in collaboration with clinical users demonstrates the application's strengths. It also aligns with best practices for user-centered design in health technology.

Methodologically, integrating the pilot into routine care and collecting real-world usage and perception data avoided potential limitations for the project. It also allowed for observations of how clinicians interacted with the platform. On the other hand, combining usage metrics with qualitative feedback provided a comprehensive perspective on both behavioral and experiential dimensions.

However, some limitations must be acknowledged. The study focused on early-stage user adoption and did not assess the tool's impact on clinical accuracy, workflow efficiency, or patient outcomes. Furthermore, the paucity of interview data may have negatively impacted the positive evaluation results. For example, the fact that four of the five interviewed participants were power users may have contributed to the positive feedback during the evaluation phase. Furthermore, while the study tracked engagement over time, it could not analyze whether usage was related to clinician roles, patient types, or workload. Another important factor observed during the pilot was that not all clinicians were equally willing to adjust their established routines. This hesitancy to change working habits may have limited adoption for some users, underlining the importance of organizational readiness alongside technical reliability.

Future evaluations should therefore include a renewed assessment of time savings now that technical issues have been resolved, to more accurately determine HealthTalk’s contribution to efficiency. Combining this with pre-post comparisons and larger, more diverse samples will strengthen the evidence base.

Looking forward, future studies should include more diverse participant samples, implement pre-post comparisons to measure changes in documentation quality and time, and potentially examine patients' perspectives on the role of such tools. A follow-up study (Study ID 9) is currently planned to compare the quality of clinician-generated summaries with HealthTalk-generated summaries using de-identified transcripts. This will allow for a more rigorous evaluation of the tool's value in clinical settings.

In conclusion, the HealthTalk 9a pilot demonstrated that AI-based documentation support is acceptable, usable, and potentially effective in real-world mental health settings. With continuous improvements based on user feedback, HealthTalk can help reduce clinician workload and improve documentation quality—both factors that are increasingly critical in mental health care.

6. References

[1] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering 17, 734-749.

[2] DiClemente, C. C., & Velasquez, M. M. (2002). Motivational interviewing and the stages of change. Em W. R. Miller, & S. Rollnick, Motivational interviewing: Preparing people for change 2 (pp. 201-216). New York: The Guildford press.

[3] Gregertsen, E. C., Mandy, W., & Serpell, L. (2017, December 22). The Egosyntonic Nature of Anorexia: An Impediment to Recovery in Anorexia Nervosa Treatment. Frontiers Psychology 8, pp. 1-9.

[4] Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (May de 1995). Recommending and evaluating choices in a virtual community of use. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 194-201.

[5] Lalonde, M. P., O'Connor, K., Aardema, F., & Coelho, J. S. (1 de March de 2015). Food for Thought: Ego-dystonicity and Fear of Self in Eating Disorders. European Eating Disorders Review 23, pp. 179-184.

[4] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. eprint arXiv:2303.13375, pp. 1-35.

[5] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM computing surveys (CSUR) 52,1-38.

D6.2: Implementation and documentation of conducted studies - Bewell–Vestel Wearable Physiological Monitoring Device & Multimodal Video Protocol

1. General information

Country and name of use case: Türkiye: Major Depressive Disorders (melancholia, and catatonic, seasonal, and psychotic depression)

Name of technical system/component: Bewell–Vestel Wearable Physiological Monitoring Device & Multimodal Video Protocol

Study director/manager: ÖZEL NPİ NÖROPSİKİYATRİ İST. SAĞ. EĞİTİM DANŞ. YAY. İNŞ. SAN. VE TİC. AŞ. - Ahmet Furkan Tarhan

Contact (e-mail): bilgi@npistanbul.com

Additional partners:

- VESTEL SAĞLIK TEKNOLOJİLERİ A.Ş. (Öner Tekin)
- BEWELL TEKNOLOJİ SANAYİ VE TİC. A.Ş. (Taşkın Kızıl)
- ARD GRUP BİLİŞİM TEKNOLOJİLERİ ANONİM ŞİRKETİ (Berk Cengiz)
- ETİYA BİLGİ TEKNOLOJİLERİ YAZILIM SANAYİ VE TİCARET ANONİM ŞİRKETİ (Taner Orta)

Aim of the validation: The primary aim of this validation was to technically evaluate the functionality, usability, data quality, and system integration of the Bewell–Vestel wearable device and the associated video protocols within the DAIsy platform. Specifically, the study aimed to assess hardware ergonomics, battery optimization, and the reliability of the data transmission chain (BLE-to-Gateway) during a pre-clinical pilot to ensure readiness for deployment in Major Depressive Disorder clinical trials.

If follow-up, name of the previous validation: [Information not available in source document]

2. Methods

This chapter outlines the methodological framework established for the study. It details the underlying motivation for the technology's development, the specific design of the pre-clinical pilot, the instruments used for data acquisition, and the procedural implementation of the study.

2.1 Motivation / Background

The development of the DAIsy ecosystem is driven by the need to improve the diagnosis and care of mental diseases through Artificial Intelligence. A cornerstone of this initiative is the ability to continuously and objectively monitor patients' physiological parameters and emotional states. The Bewell–Vestel wearable device and the multimodal video protocol were designed to meet this need by providing a stream of physiological data (Heart Rate, SpO2, activity) and behavioral data (facial expressions, voice prosody).

However, before these technologies can be deployed in a sensitive clinical setting involving patients with Major Depressive Disorder, rigorous technical and usability validation is required. The motivation for this specific pre-clinical study arose from the necessity to verify significant hardware and software updates. These included a new ergonomic curved shell design for the wearable to improve patient comfort during long-term use, and a robust data transmission architecture capable of handling connectivity interruptions via offline buffering. Furthermore, the study sought to standardize the "Valence-Arousal" video protocol to ensure consistent data collection across different physical locations. The background of this study is therefore rooted in the transition from prototype

development to clinical readiness, ensuring that the system is safe, reliable, and capable of generating high-quality data for subsequent AI analysis.

2.2 Study design

The study was structured as a pre-clinical pilot and feasibility study. It was designed to stress-test the technical components and assess user experience in a real-world environment before the commencement of the main clinical trials.

- **Study objectives/hypotheses:**
    - To verify the ergonomic comfort of the new curved device design and the usability of the magnetic charging system.
    - To validate the reliability of the end-to-end data transmission chain, specifically measuring packet loss, latency, and the success rate of the offline buffering and reconnection mechanisms.
    - To assess the feasibility and standardization of the "Valence-Arousal" video protocol across different recording environments.
    - To evaluate the accuracy of the new current-based battery level estimation system.
- **Input measures:**
    - **Physiological Signals:** Photoplethysmography (PPG) signals, acceleration data.
    - **System Metrics:** Connection status (BLE/MQTT), battery voltage and current, timestamp accuracy.
    - **Video Inputs:** Facial video recordings, audio recordings of voice prosody.
    - **Subjective Inputs:** User responses to System Usability Scale (SUS) and NASA-TLX questionnaires.
- **Outcome measures:**
    - **Usability:** Scores derived from SUS and NASA-TLX surveys; qualitative feedback on comfort and ergonomics.
    - **Technical Performance:** Packet loss percentage, median data latency, reconnection success rate, battery duration (hours/days).
    - **Safety:** Incidence of skin irritation or physical discomfort.
    - **Protocol Adherence:** Valid session rates for video recording, synchronization error rates, and inter-rater reliability (Cohen’s kappa) for annotations.
- **Type of study:** Pre-clinical pilot / Feasibility study.
- **Number of participants and selection criteria:**
    - The study involved 25 voluntary participants.
    - The selection criteria focused on healthy controls, specifically university students, who were capable of using mobile applications and wearing the device continuously.
- **Power analysis:** [Information not available in source document]
- **Inclusion and exclusion criteria:**

  - **Inclusion:** Healthy volunteers able to provide informed consent and adhere to the monitoring protocol.
  - **Exclusion:** [Information not available in source document]
- **Randomization/blinding:** [Information not available in source document]
- **Type of anonymisation coding list/code word:** Data was managed using token-based authentication and role-based access control (RBAC) to ensure privacy.
- **Duration of the study:** The wearable device testing phase lasted for an average of 8 days per participant.
- **Description of the intervention:** Participants were provided with the Bewell–Vestel wearable device and the "Meditis" mobile application. They were instructed to wear the device during their daily routine for the duration of the study, ensuring continuous data collection. They also participated in video recording sessions using tablets stationed in three different locations, responding to a standardized set of questions designed to elicit specific emotional states.
- **Control groups:** No specific control group was utilized for this technical feasibility pilot.

2.3 Data acquisition

Data acquisition was multifaceted, utilizing a combination of hardware sensors, mobile software, and standardized recording protocols to capture a comprehensive dataset.

- **Tools and instruments used:**
  - **Wearable Device:** A Vestel-manufactured wristband equipped with optical sensors for Heart Rate and SpO2, an accelerometer for activity tracking, and a temperature sensor.
  - **Mobile Application:** The "Meditis" app served as the primary user interface and data gateway, facilitating device pairing via QR codes and visualizing health metrics.
  - **Video Recording Setup:** Tablets were deployed in three standardized locations to record user responses.
  - **Questionnaires:** Standardized scales (SUS, NASA-TLX) were used to assess usability and workload.
- **Data collection schedule:**
  - **Physiological Data:** Collected continuously (24/7) throughout the 8-day pilot period.
  - **Video Data:** Collected in specific sessions, with a target duration of approximately 20 minutes per session. Each question in the protocol aimed for a response time of 30-60 seconds, with an upper limit of 80 seconds.
- **Measured variables:**
  - **Physiological:** Heart Rate (BPM), Oxygen Saturation (SpO2 %), Step Count, Sleep Duration.
  - **Technical:** Battery level (%), BLE connection status, Data packet timestamps.
  - **Behavioral:** Facial micro-expressions, voice prosody features.
- **Other important facts missing in this list:**

- The system utilized Automatic Gain Control (AGC) and Adaptive Artifact Cancellation (AEC) to manage signal quality and reduce motion artifacts during acquisition.
- The video protocol specifically targeted the "Positive Valence – High Arousal" axis.

2.4 Study implementation

The implementation phase involved the coordination of technical partners and clinical staff to ensure smooth deployment and monitoring of the devices.

- **Description of the study process:** The study commenced with a briefing session where participants were issued the wearable devices. They were guided through the onboarding process, which involved pairing the device with the "Meditis" app using a QR code. Over the subsequent 8 days, participants went about their daily lives while the devices passively collected data. The video protocol was administered separately, with participants responding to the standardized question set.
- **If applicable, pilot measurements:** A preliminary technical pilot was conducted with university students for one week prior to the main pre-clinical verification to test initial stability.
- **If necessary, training of staff:** [Information not available in source document]
- **Participant recruitment:** Recruitment was targeted at university students and healthy volunteers associated with the partner institutions to facilitate easy technical support and device retrieval.
- **Informed consent process:** The study was conducted in accordance with ethical standards, having received approval from the Üsküdar University Non-Interventional Research Ethics Committee. All participants provided informed consent regarding the collection and R&D usage of their data.
- **Implementation of data collection process:**
  - **Wearable Data Flow:** Sensor data was buffered locally on the device and transmitted via Bluetooth Low Energy (BLE) to the gateway (mobile app). The gateway then timestamped the reports, checked for format integrity, and transmitted them to the cloud backend via MQTT.
  - **Video Data Flow:** Video sessions were recorded on tablets and processed through a pipeline that included face marker extraction and transcript generation.
- **Data monitoring:** The technical team actively monitored the incoming data stream using the "Meditis Pro" web dashboard. This allowed for the real-time detection of data gaps, verifying the successful operation of the offline buffering and reconnection logic.
- **Description of groups/control groups and control group management:** [Information not available in source document]

3. Analysis plan

The analysis plan focused on validating the technical integrity of the system and assessing the quality of the collected data for future AI modeling.

- **Software and tools supporting the data analysis of the study:**
  - **Backend Infrastructure:** A microservices architecture built with Spring Boot, Node.js, and FastAPI was used for data processing.
  - **Database:** InfluxDB was utilized for the storage of time-series physiological data.

- **Visualization:** Grafana and the custom "Meditis Pro" dashboard were used for data visualization and monitoring.
- **Video Processing:** Custom pipelines were employed for extracting facial markers and voice prosody features.

- **Description of datasets:**
  - **Dataset 1 (Physiological):** A time-series dataset comprising Heart Rate, SpO2, and activity metrics from 25 participants over an average of 8 days.
  - **Dataset 2 (Video):** A collection of video recordings annotated according to the Valence-Arousal model, including "Positive Valence – High Arousal" states.
- **Handling of missing data and outliers:**
  - **Missing Data:** The system's "offline buffering" capability was the primary mechanism for handling missing data due to connectivity loss. Data stored on the device was transmitted upon reconnection.
  - **Outliers:** The sensor hub provided confidence metrics and state bits which were used to filter out unreliable data points, particularly those caused by motion artifacts.
- **Data (pre)processing:**
  - **Physiological:** Pre-processing involved timestamp synchronization to ensure temporal alignment, unit standardization, and data integrity checks (checksums).
  - **Video:** The pipeline included transcript generation, prosody extraction, and the verification of annotation consistency.
- **Statistical methods and models:**
  - **Descriptive Statistics:** Mean, median, and standard deviation were calculated for technical metrics such as battery life and packet loss percentage.
  - **Reliability Metrics:** Inter-rater reliability for video annotations was assessed using Cohen’s kappa.
- **Inferential/descriptive statistics:** The analysis primarily relied on descriptive statistics to summarize technical performance and usability scores.
- **Assumption check:** [Information not available in source document]
- **Adjustment for multiple comparison:** [Information not available in source document]
- **Exploratory/unplanned analysis:** [Information not available in source document]

4. Results

The results of the pre-clinical pilot demonstrate the system's readiness for deployment, highlighting key successes in hardware ergonomics and data stability.

- **Main findings:**
  - **Ergonomics and Usability:** The new curved shell design of the wearable device was validated as comfortable for continuous wear, with no reported incidents of skin irritation. The magnetic charging system was confirmed to be user-friendly.
  - **Data Transmission and Integrity:** The "offline buffering" mechanism functioned correctly, successfully storing and re-transmitting data during connection drops. The

BLE connection demonstrated stability in hospital environments, with reliable automatic reconnection.

- **Battery Performance:** The upgraded current-based battery measurement system provided accurate discharge estimations, and the device successfully met the operational duration requirements for the study.
- **Video Protocol Feasibility:** The standardized "Valence-Arousal" video protocol was successfully implemented across three locations, achieving valid session rates and satisfactory lighting/audio quality.

- **Visual representations of findings:**
  - [Image: See Figure 1.a Schematic Circuit Design in source document DAIsy 2025-1. Dönem Teknik Rapor Taslağı.docx]
  - [Image: See Figure 1.b Wearable device PCB design in source document DAIsy 2025-1. Dönem Teknik Rapor Taslağı.docx]
  - [Image: See Figure 3 Layered Architecture and Data Flow in source document DAIsy 2025-1. Dönem Teknik Rapor Taslağı.docx]
- **Possible side or unintended effects:** No adverse physical side effects were observed during the study. Minor synchronization delays were noted in the mobile application but did not compromise the completeness of the data.

5. Discussion

This chapter interprets the findings of the pre-clinical pilot, placing them in the context of the project's broader goals and outlining the path forward for clinical trials.

- **Summary of main findings:** The study successfully validated the technical maturity of the Bewell–Vestel wearable system and the DAIsy multimodal video protocol. The hardware improvements proved effective in ensuring user comfort, while the software architecture demonstrated resilience in data collection and transmission.
- **Interpretation of the results:** The successful execution of the "offline buffering" and reconnection logic is critical, as it indicates the system can perform reliably in clinical settings where connectivity is not guaranteed. The positive usability feedback suggests that the device will be well-tolerated by patients with Major Depressive Disorder, which is essential for compliance in the main study.
- **Implications of the study for theory/practice and further research:** The validation of these technologies establishes a robust foundation for the main clinical trials. It confirms that the standard operating procedures (SOPs) for device management are effective. The data collected provides a verified baseline for training the AI models that will be used for disease state assessment.
- **Strengths and limitations of the study:**
  - **Strengths:** The study was conducted in a real-world setting with continuous 24/7 monitoring, rather than a limited laboratory test. It tested the complete end-to-end system, from sensor to cloud.
  - **Limitations:** The participant pool consisted of healthy volunteers (students), whose activity levels and compliance behavior may differ from the target clinical population. The sample size (25 participants) was sufficient for technical validation but limited for broader generalization.

- **Future directions:** The immediate next step is the commencement of the full clinical trial with Major Depressive Disorder patients at NP Istanbul Brain Hospital. Future work will focus on the fusion of the physiological and video data to train the AI models for emotion recognition and health scoring. The "Micro-habit insight service" will be further optimized based on the insights gained from this pilot.
- **Conclusions:** The DAIsy technology ecosystem has passed its pre-clinical verification. The wearable device, mobile application, and backend infrastructure are technically robust, secure, and ready for clinical deployment. The study confirms that the technical risks associated with data loss and device usability have been effectively mitigated.

Evaluation Report: Monitoring of Movement and Behavioral Patterns for the Recognition of the Severity of Clinical Depression

1. General Information

- **Country:** Germany
- **Name of Study:** Monitoring of Movement and Behavioral Patterns for the Recognition of the Severity of Clinical Depression
- **Study Director/Manager:** Dr. Patrick Elfert
- **Partner (Name):** Carl von Ossietzky University Oldenburg, Faculty II - Computer Science, Economics and Law; OFFIS - Institute for Information Technology, Health Division / Biomedical Devices and Systems
- **(Brief) Aim of the Study/Evaluation:** The overarching goal of this study was the development and evaluation of a mobile system in the form of an Android application ("DAIsy DC"). This system was designed to collect a wide range of data associated with depression and general mental well-being in a scientific context, using commercially available smartphones and optionally connected smartwatches. In the long term, such a system is intended to serve as a basis for better understanding and potentially automatically assessing the individual severity of clinical depression based on the collected behavioral and movement patterns.
- **Study Start and End Date:** The development and conception of the study culminated in a final report on August 14, 2023. The actual field study for data collection extended over a planned period of two weeks per participant, with staggered recruitment and individual start times influencing the exact survey period.
- **Place of Study Implementation:** The study was realized as a field test. This means that participants used the developed app and connected devices in their natural everyday environment, without specific laboratory conditions. The sample consisted primarily of participants without a medically diagnosed depressive disorder.

2. Methods

This chapter outlines the methodological approach employed in this study. It details the study design, the procedures for data acquisition, the specifics of the study implementation, and the plan for data analysis.

2.1 Study Design

The research design of this work aimed to investigate the feasibility and potential benefits of the developed mobile application "DAIsy DC" for collecting depression-relevant data. The **primary goal** was to evaluate whether the app is capable of reliably recording the factors identified as relevant in the literature and through expert interviews in the context

of depression and mental well-being. A **secondary goal** was the identification of initial, preliminary indications of correlations between these recorded factors and the mental well-being self-reported by the participants. Furthermore, possible limitations and challenges of the chosen approach were to be uncovered in order to support the **preparation for future, more extensive studies**, especially with clinically depressed individuals.

Based on a comprehensive literature review and expert discussions with psychological specialists from the University Hospital Bonn (UKB), specific hypotheses were formulated concerning the relationship between certain behaviors or environmental factors and mental well-being:
It was assumed that high physical activity has a positive effect on mental well-being, while low entropy of visited places (an indicator of low spatial variability) is associated with a negative mental state. Furthermore, it was postulated that high smartphone and especially social media use, as well as the presence of appetite or sleep disorders (identifiable, e.g., by calorie intake or sleep duration), correlate negatively with well-being. The weather (temperature, humidity), time spent at home, and ambient brightness (light intensity) were also attributed an influence on mental state. Finally, it was expected that high scores on established depression questionnaires (PHQ-9 and BDI) would be associated with lower self-reported mental well-being among participants.
The central **outcome measures** were the data on movement and behavioral patterns collected via the app and connected systems (detailed in Section 2.2), the mental well-being self-assessed by participants several times a day on a 5-point Likert scale, and the weekly collected scores from the standardized depression questionnaires PHQ-9 and BDI. The analysis focused on descriptive correlations between these outcome measures.

The investigation was conducted as a **field study**. This approach was chosen to enable data collection under real-life everyday conditions, thereby achieving higher ecological validity. The **number of participants** was 18 individuals (13 male, 5 female) aged **between 18 and 56 years**. The essential **inclusion criterion** was the use of an Android-based mobile phone (Android Version 8 or higher), as the developed app "DAIsy DC" was designed for this operating system. The study was primarily aimed at mentally healthy individuals to initially test the basic functionality and acceptance of the app. Explicit **exclusion criteria** were not formulated, but the selection was implicitly based on the assumption of general mental health. One person with a medically diagnosed depression and corresponding medication participated in the study, but their data could not be included in the final analysis due to an insufficient amount of data.

The **duration of data collection** was set for a period of two weeks for each individual participant.

The central **intervention** consisted of the use of the Android application "DAIsy DC" (Data

Collector), newly developed for this research. This app was developed with the aim of automatically and "ambiently" (i.e., as unobtrusively as possible in the background) recording a variety of parameters. For this, the app uses both the sensors integrated into the smartphone, such as the light sensor, GPS for location tracking, accelerometers, and gyroscope for activity recognition. On the other hand, a connection to Google Health Connect (GHC) was implemented. This Android API allows data from various third-party apps (e.g., Google Fit for movement data, MyFitnessPal for nutritional data, Spotify for music preferences) and connected wearables (e.g., smartwatches for pulse and sleep data) to be aggregated in a standardized way and made available to the DAIsy DC app. The recorded parameters included, among others, call behavior, ambient light intensity, details of smartphone use (total duration, applications used), various location information (for calculating entropy and time spent at home), daily step count, nutritional data (calories, macronutrients), weather data of the current location, listened to music tracks and artists, sleep duration and phases, as well as heart rate and heart rate variability. In addition to these passively collected data, demographic information was collected once, and the established depression questionnaires PHQ-9 and BDI were administered weekly. A core element was also the daily, multiple queries of current mental well-being via an easy-to-use 5-point Likert scale. A schematic representation of the system architecture and data flows is provided (see Figure 6).

As this was an exploratory field test to evaluate app functionality and for initial data acquisition, there was no **control group** in the classic sense of a randomized controlled trial.

2.2 Data Acquisition

Data acquisition in this study relied on a multimodal system, the core of which was the self-developed Android application, the **DAIsy DC App**. This app served as the primary tool for collecting both passively recorded sensor data and information actively entered by the participants. To ensure a broad data base and to enable the integration of existing health and fitness applications, **Google Health Connect (GHC)** was used as a central Android API. GHC acted as an interface to aggregate data from various **third-party apps** and connected wearables. The third-party apps relevant to the study included **Google Fit** (especially for activity and movement data, if no smartwatch with direct GHC synchronization was used), **MyFitnessPal** or, alternatively, the **Digital Nutrition Diary (DND)** developed by OFFIS (for manual recording of nutritional data), and **Spotify** (for automated recording of listened-to music via the Spotify SDK).

Another important component of data acquisition was standardized **questionnaires** integrated directly into the DAIsy DC App. This included an initial **questionnaire on demographic information** (age, gender, education level, etc.), which was completed once upon registration. Repeatedly, namely weekly, the **Patient Health Questionnaire-**

9 (PHQ-9) and the **Beck Depression Inventory (BDI)** were used to collect established measures of depressive symptomatology. Current **mental well-being** was queried several times a day using a visual **5-point Likert scale** with smileys, accessible both via the app's home screen and a widget on the smartphone's home screen (a representation of the scale is shown in Figure 14).

The DAIsy DC App accessed various **smartphone sensors** to collect behavioral and environmental data. The **light sensor** recorded ambient brightness in lux. **GPS** was used to determine latitude and longitude, as well as the derived address (encrypted), for location determination. The **accelerometer**, **gyroscope**, and **magnetometer** provided raw data for activity recognition and the analysis of usage patterns. Additionally, external **APIs** were connected: The **OpenWeather API** provided location-based weather and bioweather data, while the **Spotify SDK** enabled the recording of music listening habits.

The data collection schedule provided for continuous or interval-based recording of most passive data, as well as regular active input by the participants:
Mental well-being (Likert scale) was to be entered at least three times a day; push notifications also served as reminders for this. The PHQ-9 and BDI questionnaires were queried weekly, also with a reminder function. Demographic information was collected once after the first login.
Passive data recording was as follows: Light intensity was recorded minutely when the screen was on. Location was determined every 15 minutes (latitude/longitude and address were SHA-256 encrypted and stored). Smartphone usage (apps used, duration) was retrieved daily for the preceding day. Smartphone usage while sitting, standing, or lying down at the defined home location was tracked continuously when inactivity and a switched-on screen were detected at the respective location. Call behavior (start time, duration) was recorded for every call made via the mobile network. Data from Google Health Connect (such as steps, movement minutes, nutritional data, sleep, heart rate, heart rate variability, and specific sports activities) were read out daily for the previous day; an automatic check and synchronization also occurred every eight hours if the app was not opened, to minimize data loss. Weather and bioweather data were retrieved once a day based on the current location. Music listened to via Spotify (title, artist) was recorded with each newly played song.
In total, the following measured variables were considered for the analysis:
Demographic data included age, gender, relationship status, highest level of education, smoking habits, presence of children, and information on medical history (current medical or psychiatric treatment, medication intake). Mental well-being was recorded as a self-reported value on the 5-point Likert scale (coded from -2 "very poor" to +2 "very good"). Depressive symptomatology was operationalized via the sum scores of the PHQ-9 and BDI questionnaires. Physical activity included the daily step count, total duration of movement minutes, and type, duration, and start/end times of specific sports activities.

Sleep duration was recorded in milliseconds. Nutritional data included energy intake in kilocalories, and amounts of protein, carbohydrates, and fat in grams. Smartphone usage was recorded in detail by duration of use of individual apps, total smartphone usage time, and duration of use while sitting, standing, or lying down at the home location. Social interaction was approximated by call behavior (duration of calls). Environmental factors included light intensity (in lux), various weather data (temperature, perceived temperature, humidity, air pressure), and bioweather information. Mobility and routine were mapped using location information (encrypted latitude/longitude and addresses), the entropy of visited places calculated from this, and time spent at home (duration). Music preference was recorded by the titles and artists listened to on Spotify. Physiological parameters, if available via GHC and connected devices, included the average, minimum, and maximum heart rate per day, and the RMSSD value as a measure of heart rate variability.

2.3 Study Implementation

The **implementation of the study** began with instructing participants to install the developed DAIsy DC app and, depending on their individual device setup and preferences, necessary third-party apps such as Google Fit, MyFitnessPal (or DND), and Spotify on their Android smartphones. The installation of Google Health Connect was also required to enable data integration. After successful installation, an initial registration in the DAIsy DC app was carried out, using a username assigned by the study management and a password chosen by the participants themselves. Immediately thereafter, participants completed a one-time questionnaire on their demographic data. Upon completion of these setup steps, the two-week phase of active and passive data collection began. The DAIsy DC app continuously collected passive sensor data and information from connected services in the background. In parallel, users were actively prompted to enter their current mental well-being multiple times a day via the Likert scale and to complete the PHQ-9 and BDI depression questionnaires weekly. Push notifications served as reminders for these active inputs. The study was designed as a field test, meaning participants continued their normal daily routines without specific laboratory instructions to obtain as natural a data basis as possible. Detailed instructions for installing and using the app were provided to the participants.

The **recruitment of participants** is not described in detail, but a sample of 18 individuals was obtained for participation in the study.

The **informed consent process** was a central component of the study preparation. Participation was purely voluntary. All potential participants received comprehensive written information about the study's objectives, the type of data to be collected, the methods of data collection and processing, data protection measures (including pseudonymization using a coding list), and their right to withdraw from participation at any

time without giving reasons and without any disadvantages. This information was summarized in a dedicated participant information document. Consent to participate was documented.

The technical **implementation of the data collection process** was based on the DAIsy DC app. This app utilized various Android system functions such as UsageStatsManager for recording app usage, SensorManager for accessing hardware sensors (light, acceleration, etc.), LocationClient for location data, and TelephonyManager for call information. A key component was the integration of the Google Health Connect API, which enabled standardized access to data from other health and fitness apps. The collected data were cached locally on the smartphone in an SQLite database and also sent to a central SQL server. An implemented synchronization mechanism ensured that data collected offline (without an active internet connection) were transmitted to the server at the next opportunity to minimize data loss (illustrated in the activity diagram, Figure 7). Participants also had the option to granularly adjust permissions for data collection by the DAIsy DC app and for Google Health Connect, thereby maintaining control over their data (see Figure 17, which details app permissions). The entity-relationship model of the database is depicted in Figure 16.

Regarding **data monitoring for quality and quantity**, several challenges were identified during the analysis phase: *Erroneous data* occurred, for example, in smartphone usage, where unrealistic values of over 24 hours of use per day were recorded in isolated cases. Call behavior was also incompletely recorded (only calls via the mobile network were tracked, not those via messenger services), and there was a high number of duplicates in the Spotify music data. Furthermore, there were *missing data*, especially for physiological measures such as heart rate and heart rate variability. This was attributed to the low prevalence or use of smartwatches or other compatible wearables among the participants. Sleep data, which had to be entered manually in the absence of a smartwatch, were also often incomplete. Finally, the relatively short study duration of two weeks per person, combined with the sometimes late registration of some participants, led to a *limited amount of data* for statistical analysis. For each factor under investigation, data from only a small subgroup of individuals who provided usable data for at least eight days could often be considered. This significantly limited the conclusiveness of the statistical analyses.

3. Analysis Plan

A range of software and tools were used for the **data analysis**. The primary **database infrastructure** consisted of a local SQLite database on the participants' Android smartphones for temporary storage and a central SQL server for consolidated data storage. The actual **data analysis software** was the Anaconda Distribution (Version 4.10.3), which provides a comprehensive environment for scientific computing with

Python (Version 3.10.11). Within this environment, specific **Python libraries** were used: Jupyter Notebook (Version 8.1.0) served as an interactive development environment for the analysis scripts. NumPy (Version 1.24.3) was used for numerical calculations and array handling. The pandas library (Version 1.5.3) was central to data manipulation and analysis, particularly through the use of DataFrame structures. Matplotlib (Version 3.7.1) was used for creating visualizations, supplemented by Seaborn (Version 0.12.2), which provides an interface for more attractive statistical graphics.

The **description of the datasets** for the analysis included the variables detailed in Section 2.2. For most factors, the raw data were aggregated into daily values to examine daily patterns and relationships. Due to the varying data quality and quantity per participant and factor, a minimum data density was established for the descriptive analysis: Only participants who had provided usable data for the respective factor for at least eight days were included in the specific factor analysis.

The **handling of missing data and outliers** was an important consideration. *Missing data*, as mentioned in the section on study implementation, was a present issue. Factors for which very few data points were available (e.g., heart rate variability due to lack of smartwatch use) could not be included in the detailed correlation analysis. The general limitation due to the small amount of data per person and factor was highlighted as a significant constraint for in-depth statistical analyses. *Outliers and obviously erroneous data* (e.g., smartphone usage values exceeding 24 hours per day) were identified during data exploration. The discussion of the results mentioned these data quality issues, and they led to some factors (such as call behavior and music usage) not being included in the results presentation, or only to a very limited extent. A systematic statistical outlier treatment was not described in detail due to the primarily descriptive nature of the analysis; however, extreme values (e.g., very high daily step counts) were mentioned in the discussion as potentially important aspects to be considered for future, more detailed analyses.

Data (pre)processing involved several steps to prepare the raw data for analysis. Self-reported mental well-being, recorded on a 5-point Likert scale, was converted into numerical values for quantitative analysis: "very poor" received a value of -2, "poor" -1, "ok" 0, "good" 1, and "very good" 2. Location data (latitude and longitude) underwent a cleaning process where multiple entries within a 15-minute window were grouped into a single place of stay. This was done to determine the number of unique places visited for at least 15 minutes per day, which then served as the basis for calculating location entropy. For the comparative presentation of results from the depression questionnaires (PHQ-9 and BDI) in relation to mental well-being, the respective sum scores were scaled to a uniform range between 0 and 1 using min-max normalization to ensure better comparability.

As **statistical methods and models**, primarily **descriptive statistics** were used due to the aforementioned limited amount of data and the exploratory nature of the study. To obtain initial indications of possible relationships between the recorded factors and mental well-being, a **descriptive correlation analysis** was performed for individual participants and factors. For this, scatter plots were created, and Pearson correlation coefficients and their Fisher's Z-transformation were calculated as quantitative measures. The associated p-values were considered as an indication of the possible randomness of the observed correlations, but without claiming statistical significance for the entire study population. Data **visualization** played a central role in presenting the results. Box plots of the Fisher-transformed correlation coefficients for the evaluated factors across all considered participants were created (see Figure 33) to provide an overview of the distribution and direction of individual relationships. Individual relationships were also illustrated by scatter plots with regression lines. Normalized questionnaire scores were presented in bar charts.

4. Results

The **main findings** of the study indicated that the developed DAIsy DC app was, in principle, capable of collecting a wide range of data potentially relevant to the study of depression and mental well-being. However, the descriptive analysis of the collected data primarily revealed individual differences in the relationships between the investigated factors and the self-reported mental well-being of the participants.

Regarding **physical activity**, measured by steps and minutes of movement, some participants showed positive associations with mental well-being. For example, user ID 20 showed a tendency towards better well-being with a higher step count (see Figure 20), and user ID 14 tended to report better well-being on days with more minutes of movement. However, there were also contrary observations, such as with user ID 23, whose well-being tended to decrease with very high step counts; this specific negative correlation even showed statistical significance at the level of this individual case with a p-value of 0.05 (see Figure 21).

The **entropy of location variance**, a measure of the diversity of places visited, also showed mixed results. For some participants, such as user ID 22, higher entropy (i.e., visiting more different places) was associated with better mental well-being, while for others, such as user ID 20, lower entropy correlated with better well-being (see Figures 23 & 24).

Contrary to the original hypothesis, the (limited) data on **smartphone usage** from three individuals tended to show a positive association with mental well-being (see Figure 25 for user ID 23). Similarly inconsistent were the results for **social media usage**: Of the two individuals whose data could be analyzed for this, one showed a positive and the other a

negative tendency in relation to well-being (see Figures 26 & 27).

Data on **nutrition**, particularly calorie intake, were successfully collected and could be related to well-being. For user ID 14, for example, there was a tendency towards better well-being with higher calorie intake (see Figure 28).

The influence of **weather** (temperature and humidity) on mental well-being was also highly individual. For some participants, such as user ID 20, well-being decreased with increasing humidity (see Figure 29).

Possible **sleep disorders**, operationalized by sleep duration, could potentially be identified from the data. For two of the three individuals with analyzable sleep data, well-being decreased with very long sleep duration (over seven hours), while one person reported an increase in well-being with increasing sleep duration (Figure 30, pertaining to user ID 20, also shows values with very short sleep duration).

Time spent at home also showed mixed correlations: For some participants, such as user ID 20, more time at home correlated with better well-being, while for others it was the opposite (see Figure 31).

Surprisingly, for the majority of the analyzed individuals (6 out of 8), a higher average daily **light intensity** correlated with *poorer* mental well-being, which contradicted the original hypothesis.

Finally, the analysis of the **questionnaires (PHQ-9 and BDI)** in relation to self-reported mental well-being showed an expected trend: Individuals with higher (normalized) sum scores on the depression questionnaires tended to report lower values for (normalized) mental well-being, and vice versa. This is visualized in Figure 32, which compares the normalized values.

The **visual representations of the results** include numerous scatter plots for individual factors and participants (Figures 20-31), illustrating the individual trends. Figure 33 summarizes the distribution of correlation coefficients for various factors in a box plot and provides a general overview of the direction of effect (positive, neutral, negative) of the factors on mental well-being, based on the individual correlations.

As **possible side or unintended effects** of the study, the identified data quality problems were primarily noted. These included erroneous or incomplete data recordings for smartphone usage, call behavior, music recognition, as well as heart rate and sleep data (especially with manual entry or missing wearables). The study thus also highlighted the limitations of a purely smartphone-based recording for certain parameters. Another point was that the recording was limited to the smartphone, and the use of other electronic devices (tablets, PCs) for potentially relevant activities such as social media consumption

or work could not be taken into account.

5. Discussion (corresponds to point 6 in the template)

The **summary of the main findings** clarifies that the study was able to demonstrate the fundamental feasibility of collecting a wide range of potentially depression-relevant factors using the developed DAIsy DC app. Initial, albeit highly individually variable, trends were observed in the relationships between the recorded behavioral and movement patterns and the participants' self-reported mental well-being. The results of the standardized depression questionnaires (PHQ-9 and BDI) showed a tendency to correlate with subjective well-being as expected.

The **interpretation of the results** must be made with caution. Due to the short study duration of two weeks per participant and the resulting, often small number of fully usable datasets per factor and person, the findings are primarily to be regarded as descriptive and exploratory. They do not allow for generalizations or statistically secured conclusions for the overall group of potentially affected individuals. The observed correlations showed considerable inter-individual variability, indicating the complexity of the relationships between lifestyle factors and mental state. The need to improve data quality and quantity in follow-up studies became clear, for example, through greater integration and use of wearables (especially for physiological data such as heart rate and detailed sleep analyses) and by choosing longer observation periods. Some results, such as the tendency for a negative correlation between ambient brightness and well-being or the partially positive relationship between smartphone use and well-being, contradicted the initial hypotheses and the results of some other studies. These discrepant findings underscore the need for further, more detailed research, which may also need to consider contextual factors more strongly.

The **implications of the study for theory, practice, and further research** are manifold. The work lays an important technical and conceptual foundation for more extensive follow-up studies. Of particular interest here would be investigations with clinical populations, i.e., individuals with a medically diagnosed depression, in order to examine the identified factors and their correlations with the actual severity of the illness in more detail. The collected multimodal and longitudinal data could in the future provide a valuable basis for the development of machine learning models (Artificial Intelligence). Such models could potentially be used to create personalized risk assessments for depressive episodes or to derive individualized intervention recommendations. The study has also highlighted the need to better understand the context of data collection – for example, the reasons for low physical activity or altered sleep behavior – in order to interpret the data more meaningfully.

Among the strengths of the study are the development of a comprehensive system for

collecting a wide range of factors potentially related to depression and mental well-being. The use of Google Health Connect as an integration platform for data from various sources is also to be highlighted positively. Conducting it as a field test in the participants' natural environment increases the ecological validity of the results. The combination of passive sensor data collection and active self-reporting enabled a multi-faceted view. Last but not least, the detailed participant information and the focus on data protection and user control over data sharing are to be valued as strengths.
However, the limitations of the study are also significant. The aforementioned short study duration and the small number of participants with complete datasets over the entire period severely limit the generalizability and statistical robustness of the results. Data quality problems with some specific factors (smartphone use, call behavior, music data) as well as data gaps due to the low use of smartwatches by participants are further important limitations. The need for manual data entry (e.g., for nutrition, partly for sleep) carries the risk of incompleteness and inaccuracy. The restriction of recording to the smartphone means that the use of other electronic devices (tablets, PCs) for potentially relevant activities was not taken into account. Finally, possible reactivity effects – i.e., a change in participants' behavior solely due to the knowledge of the continuous recording of their behavior – cannot be ruled out and could have influenced the results.
Clear recommendations for **future directions** emerge from these findings. An extension of the study duration and the recruitment of a larger number of participants are essential in order to be able to make statistically valid statements and to better model inter-individual differences. Conducting studies with individuals actually suffering from depression is necessary to investigate the clinical relevance of the recorded factors and possible digital biomarkers. Technical improvements to the app are also indicated; these include the implementation of a data overview and a correction option for users, as well as the optimization of recording modules for those factors that proved problematic (e.g., recording communication via messenger services instead of pure telephony). Stronger promotion of the use or even the provision of wearables (smartwatches, fitness trackers) for study participants could significantly improve the quality and quantity of physiological data (especially sleep and heart rate data). The collected, rich, and multidimensional datasets offer great potential for the development and application of machine learning models to identify complex patterns, non-linear relationships, and cross-correlations that might not be detected with conventional statistical methods. Finally, future research should also increasingly aim to investigate causalities – for example, whether low physical activity is a cause or a consequence of poor mental well-being.

In **conclusion**, it can be stated that the DAIsy DC app developed for this research is a promising tool for collecting data that may be relevant for a deeper understanding and potentially for the early detection or therapy-accompanying support of depression. The pilot study provided valuable initial insights into the feasibility of such an approach and at the same time uncovered important challenges and limitations. Future research should

consistently build on these results, address the identified methodological limitations, and further explore the considerable potential of digital technologies for clinical applications in the field of mental health.

6. References (corresponds to point 7 in the template)

A comprehensive list of the literature that informed this research is available. This body of work includes, among others:

- Sources on the definition and diagnostics of depression (e.g., WHO, National Treatment Guideline Unipolar Depression, ICD-10, DSM-IV).
- Validation studies on questionnaires (PHQ-9, BDI).
- Research papers on the relationship between mental well-being and various factors such as physical activity (Harris & Ashley, 2018), smartphone use (Daniyal et al., 2022; Masud et al., 2020; Cao et al., 2020), social media (Karim et al., 2020), demographic factors, sleep, nutrition, weather (Ding et al., 2016; Taniguchi et al., 2022), light intensity (Brown & Jacobs, 2011).
- Works on technical aspects such as smartphone sensor technology, activity recognition, and existing apps/platforms (e.g., RADAR-base).

(A detailed list of all specific references would exceed the scope of this document; key citations can be provided upon request, and an exhaustive list informed the original research.)

D6.2: Review Machine Learning in Eating Disorders

1. General information

Country and name of use case: Netherlands – Eating Disorder

Name of technical system/component: Narrative review and comparative evaluation of machine learning and artificial intelligence solutions for eating disorders

Study director/manager: TUe, Milan Petkovic

Contact (e-mail): m.petkovic@tue.nl

Additional partners: GGZ Oost Brabant, Centre for Eating Disorders

Aim of the validation:
The aim of this study was to review, categorize, and critically evaluate existing machine learning and artificial intelligence applications in eating disorders, with a specific focus on clinical management, to identify current state-of-the-art approaches, limitations of existing methods and datasets, and requirements for achieving actionable, trustworthy, and ethically responsible AI-supported eating disorder care.

2. Methods

2.1 Motivation / Background

Eating disorders are among the most complex psychiatric disorders, associated with severe impairment in psychological, physical, and social functioning, and substantial disease burden. Despite advances in research, early detection rates remain low, recovery rates are limited, and relapse rates are high. The multifactorial etiology of eating disorders, involving genetic, psychological, and socio-environmental factors, complicates diagnosis, prognosis, and treatment personalization.

Machine learning and artificial intelligence offer promising tools to model complex, high-dimensional data and have demonstrated value in other medical domains, including oncology and cardiology. In eating disorders, AI has the potential to support early detection, risk assessment, diagnosis, monitoring, treatment response prediction, and development of digital interventions. However, increasing model complexity raises concerns regarding explainability, fairness, trust, and clinical applicability.

Given the rapid expansion of AI research in eating disorders, a structured synthesis is needed to bridge the knowledge gap between eating disorder researchers and AI practitioners, to map current applications, and to identify methodological and practical barriers that hinder translation to actionable healthcare. This review was conducted to address this need.

2.2 Study design

This study employed a narrative review design. The primary objective was to identify and synthesize existing machine learning and artificial intelligence applications in eating disorders, with a focus on clinical management use cases. Secondary objectives were to compare AI techniques across use cases, evaluate their strengths and limitations, and identify constraints related to data quality, explainability, and clinical integration.

The review systematically categorized published studies into key use cases, including social media and internet analysis, early detection and risk factor identification, diagnosis, monitoring and prognosis, and treatment development.

2.3 Data acquisition

Data were acquired through a structured literature search, described in detail in the source document's appendix. Peer-reviewed publications reporting the application of machine learning, artificial intelligence, or causal models in eating disorders were identified.

Extracted information included eating disorder subtype, study aim, sample characteristics, predictor variables, data time points, algorithms used, and reported performance metrics. The review encompassed studies using a wide range of data types, including questionnaires, clinical records, neuroimaging, genetic data, physiological signals, wearable sensors, and social media content.

Measured variables and outcomes varied across studies and included diagnostic accuracy, risk prediction, symptom severity, treatment response, and feasibility of digital interventions. Additional relevant information included model explainability, generalizability, and healthcare constraints.

2.4 Study implementation

The review was implemented through iterative reading, categorization, and synthesis of the identified literature. Studies were grouped into clinically relevant use cases and further analyzed according to the machine learning techniques employed.

Data quality considerations focused on transparency of reporting, dataset size and representativeness, and methodological rigor of the reviewed studies.

3. Analysis plan

The analysis consisted of qualitative and descriptive comparative synthesis. Studies were first categorized by eating disorder use case and then by the type of AI or machine learning technique applied. Techniques were compared along dimensions including complexity, flexibility, explainability, generalizability, performance, and adaptability to healthcare constraints.

The review distinguished between traditional statistical learning, shallow machine learning, and deep learning approaches. Strengths and limitations of each class of techniques were analyzed, with particular attention to assumptions, data requirements, robustness to missing data, and interpretability.

4. Results

Machine learning and AI applications in eating disorders were identified across five main domains: social media and internet analysis, early detection and risk factor identification, diagnosis, monitoring and prognosis, and treatment development. A wide range of algorithms were applied, including logistic regression, decision trees, random forests, support vector machines, clustering methods, neural networks, reinforcement learning, and causal models.

Results showed that ML models can achieve moderate to high performance in specific tasks such as distinguishing eating disorder patients from healthy controls, predicting binge eating disorder, and identifying risk factors. However, performance varied widely across use cases and datasets. Many models showed limited generalizability and marginal improvement over traditional statistical approaches.

Monitoring and prognosis studies demonstrated potential for predicting treatment response and symptom trajectories, but effect sizes were often modest. Treatment-oriented applications, including chatbots and mobile interventions, showed promise but raised concerns regarding safety, personalization, and unintended harm.

5. Discussion

The review highlights substantial growth in AI research related to eating disorders, particularly in detection and monitoring. However, significant methodological and practical limitations remain. Data scarcity, heterogeneity, and lack of representative datasets constrain model robustness. Increasing model complexity often comes at the expense of explainability, which is critical for clinical trust and ethical deployment.

The findings underscore the need for careful AI model selection tailored to specific clinical questions, rather than one-size-fits-all solutions. Joint efforts between eating disorder researchers, clinicians, patients, and AI practitioners are essential to develop high-quality datasets and secure, transparent AI frameworks.

Strengths of this study include its comprehensive scope and focus on clinical management. Limitations include reliance on published literature and heterogeneity in study designs and reporting standards.

Future research should prioritize explainable, causal, and fairness-aware models, rigorous validation in clinical settings, and alignment with ethical and legal requirements.

In conclusion, while AI holds promise for advancing eating disorder care, its translation into actionable healthcare requires methodological rigor, interdisciplinary collaboration, and sustained attention to trustworthiness and patient safety.

6. References

Ghosh, S., Burger, P., Simeunovic-Ostojic, M., Maas, J., & Petković, M. (2024). Review of machine learning solutions for eating disorders. International Journal of Medical Informatics, 189, 105526. https://doi.org/10.1016/j.ijmedinf.2024.105526

Fairburn, C. G. (2008). Cognitive behavior therapy and eating disorders. Guilford Press.

Van Buuren, S. (2018). Flexible imputation of missing data. CRC Press.

World Health Organization. (2019). International classification of diseases for mortality and morbidity statistics (11th Revision).

6.1 Source Documents

Review of machine learning solutions for eating disorders - ScienceDirect

D6.2: Detection of Refeeding Syndrome using Unsupervised Anomaly Detection

1. General information

Country and name of use case: Netherlands – Eating Disorder

Name of technical system/component: Unsupervised anomaly detection models for early identification of Refeeding Syndrome risk

Study director/manager: TUe, Milan Petkovic

Contact (e-mail): m.petkovic@tue.nl

Additional partners: GGZ Oost Brabant, Centre for Eating Disorders

Aim of the validation:
The aim of this study was to evaluate whether unsupervised anomaly detection models can identify abnormal physiological patterns associated with the onset of Refeeding Syndrome (RFS) in patients with anorexia nervosa using routinely collected inpatient clinical data.

2. Methods

2.1 Motivation / Background

Refeeding Syndrome is a potentially life-threatening complication that can occur during nutritional rehabilitation of severely malnourished patients, particularly individuals with anorexia nervosa. It is characterised by metabolic and electrolyte disturbances, most notably a rapid decline in serum phosphate levels following the reintroduction of nutrition. Early detection is clinically challenging, as symptoms may be subtle, variable over time, and confounded by individual patient characteristics.

In clinical practice, RFS is diagnosed using criteria defined by the American Society for Parenteral and Enteral Nutrition (ASPEN), which rely primarily on phosphate decline within a specified time window. However, these criteria may fail to capture all clinically relevant cases, especially when refeeding is non-linear or interrupted. Moreover, labelled RFS data are scarce and often incomplete, limiting the applicability of supervised machine learning approaches.

Unsupervised anomaly detection offers an alternative by learning patterns of “normal” physiological behaviour and identifying deviations that may indicate emerging clinical risk. This study explores whether such models, applied to routinely collected clinical data, can support earlier or more sensitive identification of RFS risk in an inpatient eating disorder setting.

2.2 Study design

This study employed a retrospective observational design using routinely collected inpatient clinical data. The primary objective was to compare the effectiveness of three unsupervised anomaly detection algorithms—Isolation Forest, One-Class Support Vector Machine, and Local Outlier Factor—in identifying anomalies corresponding to confirmed cases of Refeeding Syndrome.

The study addressed three sub-questions: which anomaly detection model performs best in detecting RFS-related anomalies, which physiological features contribute most strongly to anomaly detection, and whether incorporating rate-of-change features improves model performance.

Input measures consisted of demographic variables, vital signs, laboratory biomarkers, and derived temporal features. Outcome measures included binary anomaly classifications and performance metrics computed against clinically validated RFS labels derived from ASPEN criteria.

A total of 111 unique patients diagnosed with anorexia nervosa and admitted for inpatient treatment were included. Patients with fewer than three laboratory measurements were excluded. Data were pseudonymised through shifted dates and anonymised patient and intake identifiers.

The duration of the study covered the full inpatient admission periods available in the dataset.

2.3 Data acquisition

Data were obtained from GGZ Oost-Brabant and collected as part of routine clinical care. Three datasets were used: laboratory data, vital signs data, and demographic and diagnostic information.

Laboratory data included biochemical markers such as phosphate, potassium, magnesium, glucose, ALT, AST, and leucocytes. Vital signs included heart rate, systolic and diastolic blood pressure, body temperature, weight, height, and body mass index. Demographic data included age, sex, and eating disorder diagnosis.

Measurements were recorded repeatedly during inpatient admission. Dates were converted to UNIX timestamps for time-series modelling. Missing values were handled using Multiple Imputation by Chained Equations. Derived variables included BMI and rate-of-change features (absolute and percentage changes) for dynamic clinical variables.

RFS labels were assigned at the measurement level using ASPEN criteria, based primarily on a ≥10% drop in serum phosphate within five days of refeeding initiation, though labels were

applied across the entire admission period to reflect clinical practice. Severity levels were not modelled, and labels were binary.

2.4 Study implementation

The dataset comprised 111 patients, split into a training set (94 patients) and an independent test set (17 patients). The test set included 11 confirmed RFS patients and 6 control patients. Strict separation between training and test patients was maintained to prevent data leakage.

Three unsupervised models were implemented: Isolation Forest, One-Class Support Vector Machine, and Local Outlier Factor. Each model was trained on the training set and evaluated on the test set under three feature configurations: raw features, delta (absolute change) features, and percentage change features.

Hyperparameters were tuned using grid search. Model outputs were anomaly scores converted to binary anomaly labels using predefined contamination thresholds. Model interpretability was explored using SHapley Additive exPlanations (SHAP).

3. Analysis plan

Factor analysis was first conducted on a subset of biochemical markers to identify dominant latent dimensions relevant to RFS risk. Principal axis factoring was applied, with the number of components determined using a scree plot and the elbow method. Varimax and Promax rotations were used to enhance interpretability.

Model performance was evaluated by comparing predicted anomalies to clinically validated RFS labels. The primary evaluation metric was the F2-score, prioritising recall due to the clinical importance of minimising false negatives. Precision, recall, and accuracy were also computed.

Performance was compared across models and feature configurations to assess the contribution of temporal features. False positives and false negatives were examined to explore clinically meaningful deviations beyond labelled RFS cases.

4. Results

Factor analysis identified two consistent latent dimensions explaining 55.2% of the variance. The first factor was strongly associated with liver enzymes ALT and AST, while the second was characterised by electrolyte regulation, primarily phosphate and potassium. These findings were robust across unrotated, Varimax, and Promax solutions.

In model evaluation, Isolation Forest achieved its best performance under the raw feature configuration, with an F2-score of 0.5123 and recall of 0.8929. One-Class Support Vector Machine performed best under the delta feature configuration, achieving the highest overall F2-score of 0.5760 and recall of 0.8929. Local Outlier Factor showed the weakest performance across configurations.

Rate-of-change features improved performance for OCSVM but not consistently for Isolation Forest or LOF. Several false negatives occurred near the ASPEN phosphate threshold, highlighting the difficulty of detecting borderline cases.

5. Discussion

The results demonstrate that unsupervised anomaly detection models can identify physiological patterns associated with Refeeding Syndrome using routinely collected clinical data. OCSVM with delta features showed the most favourable balance between sensitivity and precision, suggesting that temporal dynamics are particularly informative for detecting early deterioration.

Factor analysis supported the clinical relevance of electrolyte and liver enzyme markers, reinforcing their importance in RFS risk assessment. The study also illustrates limitations of relying solely on static diagnostic thresholds, as some clinically relevant anomalies were detected outside ASPEN-defined criteria.

Strengths of this study include use of real-world clinical data, strict train-test separation, and emphasis on recall-oriented evaluation. Limitations include small test set size, reliance on binary labels, and restriction to a single clinical centre.

Overall, the findings suggest that unsupervised anomaly detection may complement existing clinical criteria and support earlier identification of RFS risk in inpatient eating disorder care.

6. References

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93–104.

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245–276.

Chapple, L. S., et al. (2019). ASPEN consensus recommendations for refeeding syndrome. Journal of Parenteral and Enteral Nutrition.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. British Journal of Statistical Psychology, 17(1), 65–70.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6, 27.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23(3), 187–200.

Ktena, S., et al. (2024). Machine learning challenges in clinical anomaly detection.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. John Wiley & Sons.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. Proceedings of the 2008 IEEE International Conference on Data Mining, 413–422.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 4765–4774.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 18(1), 50–60.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67.

Wei, W., et al. (2024). Challenges of anomaly detection in imbalanced clinical datasets.

D6.2: Implementation and documentation of conducted studies – Virtual Therapy Assistant

In the generation of this report, AI was used for linguistic purposes. All AI-generated content was reviewed again by a human editor.

1. General information

The overarching project consists of four separate mini-Randomized Controlled Trials (mini-RCTs) designed to assess the feasibility and clinical utility of four distinct therapeutic modules for patients with Major Depressive Disorder (MDD): Activity Planning and Behavioral Activation (Modul 1), Stress Regulation Techniques (Modul 2), Stabilization of Self-Esteem (Modul 3), and Rumination Control Techniques (Modul 4). This report is specifically limited to the findings of **Modul 2: Stress Regulation Techniques**, as its data collection and analysis have already been completed and submitted for publication.

Country and name of use case: Germany: **Virtual Therapy Assistant**

Name of study: Evaluation of therapeutic content for self-directed use in a modular app for patients with unipolar depression

Study director/manager: UKB – Dr. Niclas Braun

Contact (e-mail): Niclas.braun@ukbonn.de

Aim of the validation: Feasibility and clinical utility investigation of four separate therapy modules for MDD, conveyed via a self-constructed digital intervention.

2. Methods

This chapter describes the comprehensive methodological approach employed for the evaluation of the digital intervention module, covering the design, participant characteristics, intervention details, and data collection procedures.

2.1 Motivation / Background

Depression is one of the most prevalent mental disorders worldwide, affecting around 5% of the global population and nearly 9.5 million people in Germany in 2022. Individuals with depression often experience persistent low mood, loss of interest, cognitive impairments, and functional limitations, imposing a heavy personal, societal, and economic burden. In Germany, depression accounted for approximately €9.5 billion in direct healthcare costs in 2020 and led to more than 53 million sick leave days in 2022.

Standard treatments—pharmacotherapy and psychotherapy—have limitations. Medication may bring side effects and shows limited clinical benefit in mild to moderate cases, while psychotherapy is difficult to access due to long waiting times (averaging 142 days). Digital health interventions, especially smartphone-based programs, offer an accessible, scalable, and cost-efficient alternative. Evidence shows small to moderate effects of digital interventions on depressive symptoms; however, their effectiveness varies, and it remains unclear which specific therapeutic components contribute most to outcomes.

Stress plays a central role in the onset and maintenance of depression. Non-digital stress management programs show moderate effects, and some digital stress regulation interventions (DSRIs) have also shown promise. However, most previous trials did not include formally diagnosed patients, and many interventions were longer and therapist-supported. This study aimed to close this gap by testing a four-week, fully self-guided DSRI as an adjunct to TAU in clinically diagnosed adults.

2.2 Study design

The research employed a randomized controlled trial with three assessment points: a baseline diagnostic and self-report assessment (T0), a post-intervention evaluation after four weeks (T1), and a follow-up after twelve weeks (T2). The primary hypothesis was that adding the DSRI to guideline-based TAU would result in a greater reduction of depressive symptom severity (PHQ-9) than TAU alone. Secondary hypotheses were that the DSRI would improve anxiety (GAD-7), perceived stress (PSS-10), and quality of life (WHOQOL-BREF). Additionally, the study assessed usability and user satisfaction.

Participants were recruited via university channels, local and online media, and an online project website. A total of 114 individuals were screened; 62 adults meeting criteria were randomized (31 DSRI + TAU, 31 TAU). Inclusion criteria were: formal ICD-10 or DSM-5 diagnosis of depression, a minimum GRID-HAMD-21 score of 9, ongoing guideline-based depression treatment (at least one past and one future appointment with a psychiatrist/psychotherapist within four months), German language proficiency, and age 18–70. Exclusion criteria included increased suicide risk, psychosis, bipolar disorder, schizophrenia spectrum disorders, or current substance dependence.

Participants were randomly assigned to DSRI + TAU or TAU alone.

The DSRI lasted four weeks and comprised four self-guided online modules:

1. Psychoeducation on stress and the “stress traffic light” model.
2. Regenerative stress management through Progressive Muscle Relaxation (PMR).
3. Cognitive restructuring of stress-amplifying thoughts.
4. Problem-solving training for stressors.

Each module took ~90 minutes and included exercises, reflection tasks, and homework. The DSRI was accessible via browser (desktop or mobile) and as an optional Android app. Participants received a brief introduction but no further guidance.

The control group continued with **treatment as usual (TAU)** per German depression guidelines, provided by psychiatrists, general practitioners, psychotherapists or neurologists in outpatient or inpatient care.

Participant codes were used for data storage and intervention access.

2.3 Data acquisition

Data collection combined remote clinical assessment and self-report measures. Baseline eligibility was confirmed through a structured ninety-minute video call using the secure platform RedMedical, which included a diagnostic interview with the GRID-HAMD-21 to assess depression severity and the Mini-DIPS for common mental disorders. Self-report questionnaires were completed online.

The primary clinical outcome was the severity of depressive symptoms measured by the PHQ-9. Secondary outcomes included general anxiety assessed with the Generalized Anxiety Disorder-7 (GAD-7), perceived stress measured by the Perceived Stress Scale-10 (PSS-10), and quality of life evaluated using the WHOQOL-BREF. Usability and user satisfaction were assessed post-intervention using the System Usability Scale (SUS), the Client Satisfaction Questionnaire – Internet (CSQ-I), task completion rates, and a self-developed qualitative user survey covering user experience, engagement and perceived strengths and weaknesses of the DSRI. Assessments were conducted at baseline (T0), after the four-week intervention (T1), and at twelve-week follow-up (T2).

2.4 Study implementation

Recruitment was conducted via internal university channels, local/online media, and an online project page. Interested participants completed an online contact form and were contacted by the study team for screening. Baseline assessments confirmed eligibility, gathered informed consent, and collected clinical and sociodemographic data. Participants were then randomized to the two study arms.

The DSRI was provided as a web-based platform (desktop and mobile) and optional Android app. Participants completed the modules independently. Adherence was monitored through task completion logs. Dropout rates and user evaluations were tracked to assess feasibility.

3. Analysis plan

Data preprocessing and analysis were conducted in R version 4.5.0. Linear mixed-effects models (LMM) were used to examine changes in all outcome variables across time and between groups. Each model included fixed effects for time (baseline, post-intervention, follow-up), group (DSRI plus TAU versus TAU), and their interaction, as well as a random intercept for participants. Age, gender, and education were added as covariates. Bonferroni-corrected post-hoc comparisons of estimated marginal means were performed where relevant. Statistical significance was set at $\alpha = .05$. Exploratory analyses summarized adherence (task completion) and user feedback descriptively.

4. Results

Sixty-two participants with depression (female: n = 44, male: n = 18; mean age 46.4 ± 14.1 years) were randomized. Baseline depressive symptom severity was moderate (PHQ-9 M = 13.87 ± 4.32) and nearly half of the sample reported depressive symptoms lasting two years or more, indicating chronicity. Groups were comparable at baseline except that TAU participants had higher baseline social and environmental quality of life.

Attrition was substantial: 23% dropped out by T1 (DSRI 35% vs. TAU 10%, p = .034) and 29% by T2 (DSRI 39% vs. TAU 19%, ns).

For the primary outcome, both groups showed a reduction in depressive symptoms over time, but there was no statistically significant interaction between time and group ($F(2,97.47)=0.77$, $p=.464$). The DSRI group did not achieve greater reductions in PHQ-9 scores than TAU alone.

Secondary outcomes followed a similar pattern. Anxiety levels slightly increased post-intervention and then decreased again at follow-up, with no significant group differences $F(2,96.90)=0.98$, $p=.380$). Perceived stress was initially higher in the TAU group at baseline but did not differ significantly between groups at later time points. Quality of life improved modestly in both groups, with no significant differences between them ($F(2,93.14)=0.08$, $p=.923$).

The digital intervention was well accepted. Usability, measured by the SUS, averaged 75.5, indicating good user-friendliness. User satisfaction, assessed by the CSQ-I, reached a medium to high level with an average score of 24.37. Objective adherence data showed that among participants who completed the post-intervention assessment, the median completion of tasks was 70%, though rates ranged widely from 3 to 100%. Participants reported spending about an hour per session on average, with substantial variability. Qualitative feedback revealed that many participants appreciated the intervention's structure, clarity, and opportunities for reflection and progress tracking. However, a slight majority described their diligence in completing the modules as low and criticized the text-heavy format, navigation difficulties, and lack of personalization or multimedia elements.

5. Discussion

This randomized controlled trial evaluated a four-week self-guided DSRI as an adjunct to treatment as usual in adults with clinically diagnosed depression. The intervention was rated as feasible, user-friendly, and moderately satisfying but did not produce superior clinical outcomes compared with standard care alone. Depressive symptoms, anxiety, perceived stress, and quality of life improved slightly in both groups without significant differences.

Several factors may explain these null findings. The intervention period was relatively short, which may have been insufficient for meaningful clinical change, particularly given that prior meta-analyses have shown dose–response effects with longer and more intensive digital stress management programs. The content of the DSRI, which emphasized relaxation and mental stress management, may also have limited its effectiveness. Evidence from component network meta-analyses suggests that some digital cognitive behavioral therapy elements,

such as problem solving, may be beneficial, while relaxation-based components can sometimes reduce overall effectiveness. Furthermore, the sample was highly clinical and chronic, differing from many previous studies that recruited participants with less severe or undiagnosed depressive symptoms, and such populations may respond less robustly to brief self-guided digital interventions.

Usability and user experience findings were encouraging but also highlighted clear improvement needs. Participants generally found the platform easy to use and appreciated its clear structure and reflective exercises. Nevertheless, many reported low motivation to complete the modules and suggested that the intervention was overly text-heavy and insufficiently tailored to individual needs. The lack of reminders or motivational prompts likely contributed to the high dropout rate observed in the DSRI group.

Strengths of the study include its randomized design, the formal diagnostic confirmation of depression, and the comprehensive evaluation of clinical and usability outcomes. Limitations are the small sample size, which reduced statistical power and precluded subgroup analyses, the high attrition rate, the short intervention duration, and the absence of therapist guidance, personalized content, or automated notifications to sustain engagement.

Future research should investigate longer and more intensive DSRI formats, integrate therapist support or automated motivational elements such as reminders, and explore personalization strategies, for instance by tailoring content to user characteristics or incorporating conversational agents. Blended-care approaches combining digital modules with clinician input could enhance adherence and effectiveness, although such models require additional resources. Examining differential effects by symptom severity and comorbidity may also provide insights into which patient groups benefit most from digital stress interventions.

Overall, the present study shows that while the tested DSRI was technically feasible and well accepted, it did not provide measurable clinical benefit beyond usual care for adults with diagnosed depression.Overall, the present study shows that while the tested DSRI was technically feasible and well accepted, it did not provide measurable clinical benefit beyond usual care for adults with diagnosed depression.

6. References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.).

Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., & Hedman, E. (2019). Guided internet-based vs. face-to-face cognitive behavior therapy: A systematic review and meta-analysis. *World Psychiatry, 18*(1), 20–28.

Baumeister, H., Reichler, L., Munzinger, M., & Lin, J.-B. (2014). The impact of guidance on Internet-based mental health interventions—A systematic review. *Internet Interventions, 1*(4), 205–215.

Harrer, M., Nixon, P., Sprenger, A. A., Heber, E., Boß, L., Heckendorf, H., … Lehr, D. (2024). Are web-based stress management interventions effective as an indirect treatment for depression? *BMJ Mental Health, 27*(1), e300846.

Kaluza, G. (2023). *Stressbewältigung: Das Manual zur psychologischen Gesundheitsförderung*. Springer.

Karyotaki, E., et al. (2021). Dismantling, optimising, and personalising internet CBT for depression: A component network meta-analysis. *The Lancet Psychiatry, 8*(6), 500–511.

Torous, J., et al. (2020). Dropout rates in digital mental health interventions. *World Psychiatry*.

D6.2: Implementation and documentation of conducted studies – Multimodal Neurofeedback

In the generation of this report, AI was used for linguistic purposes. All AI-generated content was reviewed again by a human editor.

1. General information

Country and name of use case: Germany: **Multimodal Neurofeedback**

Name of study: Efficacy study of ILF neurofeedback training for depression

Study director/manager: UKB – Dr. Niclas Braun

Contact (e-mail): niclas.braun@ukbonn.de

Additional partners: Beemedic, Offis

Aim of the validation: To evaluate the potential effectiveness of infra-low frequency neurofeedback (ILF-NF) as an adjunctive treatment to standard inpatient or day-patient care for individuals with major depressive disorder (MDD). The study aimed to determine whether ILF-NF produces superior improvements in depressive symptoms and related outcomes compared with a sham neurofeedback control condition.

2. Methods

2.1 Motivation / Background

Major depressive disorder (MDD) is characterized by the occurrence of at least five depressive symptoms over two weeks or longer, including depressed mood or loss of pleasure, changes in weight or sleep, psychomotor disturbances, fatigue, feelings of worthlessness, impaired concentration, and suicidal thoughts (American Psychiatric Association, 2013). Depression remains one of the leading causes of global disease burden (GBD 2019 Mental Disorders Collaborators, 2022), and its prevalence continues to rise worldwide (Moreno-Agostino et al., 2021).

Recommended first-line treatments include cognitive-behavioral therapy (CBT), antidepressant medication, or their combination (NICE, 2022). However, antidepressants often lead to side effects causing discontinuation rates of up to 57% (Garcia-Marin et al., 2023), and discontinuation increases relapse risk (Kato et al., 2021). Although CBT may provide better long-term outcomes (Voderholzer et al., 2024), limited therapist availability results in long waiting times. These challenges motivate the search for complementary interventions.

Neurofeedback (NF) aims to train self-regulation of brain activity by providing real-time feedback (Marzbani et al., 2016; Sitaram et al., 2017). Traditional NF protocols targeting EEG frequency bands (e.g., frontal alpha asymmetry, high-beta power reduction) have shown some promise but remain limited by small samples and few controlled trials (Fernández-Alvarez et al., 2022; Patil et al., 2023). Infra-low frequency neurofeedback (ILF-NF) focuses on infra-slow cortical oscillations (<0.1 Hz) and has been suggested to modulate resting-state networks (e.g., default mode network, saliency network) and autonomic regulation (Othmer & Othmer, 2016; Hiltunen et al., 2014; Bazzana et al., 2022).

Preliminary work suggests ILF-NF might improve mood and arousal regulation (Grin-Yatsenko et al., 2018; Dobrushina et al., 2020; Bazzana et al., 2022). However, evidence in MDD is scarce, with only a small case series to date. This study therefore aimed to test ILF-NF against a sham-controlled condition in a randomized, double-blind pilot trial.

2.2 Study design

This investigation was a double-blind, sham-controlled randomized controlled pilot trial (RCT) conducted at the Department of Psychiatry and Psychotherapy, University Hospital Bonn. The study was approved by the local ethics committee (protocol 326/23-EP) and pre-registered (DRKS00032973).

The primary objective was to test whether ILF-NF leads to greater reduction in depressive symptoms (experimenter-rated GRID-HAMD and self-rated PHQ-9) compared to sham NF when added to treatment-as-usual. Secondary objectives were to examine potential effects on anxiety, stress, quality of life, cognitive performance (Continuous Performance Test; CPT), and the P300 event-related potential.

Twenty-four adults (18–70 years) with a current unipolar depressive episode and GRID-HAMD-17 score ≥ 9 were enrolled. Exclusion criteria were high suicidality risk (GRID-HAMD suicidality item >2), schizophrenia spectrum or psychotic symptoms, bipolar or manic episodes, substance-related addiction, epilepsy or severe neurological disease. Participants were recruited from inpatients or day-patients and gave informed consent.

Participants were randomly assigned (1:1) to real ILF-NF or sham NF. Allocation was concealed from participants and experimenters, and the neurofeedback software provided identical interfaces for both conditions to maintain blinding.

All participants received 10 sessions (30 min each) of either real ILF-NF or sham NF using the Cygnet® system and NeuroAmp® II amplifier by BEE Medic GmbH. In real ILF-NF, infra-slow EEG signals (<0.1 Hz) were used to modulate visual and auditory feedback while participants watched TV episodes. Frequencies were individually adjusted (starting at 0.1 Hz, possibly reduced to 0.0001 Hz) to optimize calmness and alertness.

The sham NF condition matched session structure and feedback presentation but did not reflect actual brain activity but pre-recorded signals.

The training comprised 10 sessions. Assessments were conducted at baseline, mid-treatment (after the fifth session), post-treatment (after the tenth session), and, for self-reported outcomes, at a three-month follow-up. After completing the intervention, both participants and experimenters were asked to guess group allocation. Accuracy rates were below chance level, supporting the success of blinding.

2.3 Data acquisition

Data collection combined structured clinical interviews, self-report questionnaires, neuropsychological testing, and neurophysiological recordings. Psychiatric eligibility was established through the Mini-DIPS interview and the GRID-HAMD-17 depression rating. Participants completed a verbal intelligence test (MWTB) and a set of validated self-report measures: the PHQ-9 for depression severity, the DASS for anxiety and stress, and the WHO-QOL BREF for quality of life.

Neuropsychological assessment focused on the Continuous Performance Test (CPT), a sustained attention task requiring responses to a specific letter sequence (A followed by K). Reaction times and reaction time variability were extracted. EEG was simultaneously recorded using a 24-channel cap connected to a Smarting® wireless amplifier at 500 Hz sampling. The P300 event-related potential was derived from correct target trials during the CPT. Functional near-infrared spectroscopy (fNIRS) was also collected but is not analyzed in this report. Data streams were synchronized via Lab Streaming Layer.

2.4 Study implementation

Potential participants were screened at baseline through structured interviews and rating scales. Written informed consent was obtained before study inclusion. Experimenters underwent specialized training in ILF-NF techniques provided by BEE Medic GmbH. Neurofeedback sessions were conducted under standardized conditions, with electrodes placed according to the international 10–20 system and impedances maintained below 10 Ω. Real-time data collection was continuously monitored for quality and completeness.

Two participants discontinued early due to scheduling conflicts or clinical transfer. These drop-outs were replaced to maintain the planned sample size and balanced group allocation.

3. Analysis plan

EEG preprocessing was carried out in MATLAB R2022b using the EEGLAB toolbox. EEG data were bandpass-filtered between 1 and 40 Hz, detrended, cleaned for artifacts by combining statistical rejection and independent component analysis, and then re-referenced and low-pass filtered for event-related potential analysis. P300 amplitude and latency were extracted at channel Pz for each participant.

All statistical analyses were performed using R version 4.4.1. The primary and secondary outcome datasets comprised clinical ratings (GRID-HAMD, PHQ-9, DASS, WHO-QOL), behavioral metrics from the CPT (reaction time and variability), and EEG-derived P300 parameters.

Group and time effects were modeled using linear mixed-effects models with participant ID as a random factor. Pre-treatment scores were included as covariates. Residual diagnostics were conducted to verify model assumptions; robust models were applied when heteroscedasticity was detected but results were similar to standard models. Post-hoc comparisons were adjusted using the Bonferroni–Holm method, and significance was set at 0.05.

4. Results

Depressive symptoms, both observer-rated with the GRID-HAMD and self-rated with the PHQ-9, significantly decreased over time in both groups. However, no interaction between group and time was observed, indicating that ILF-NF did not provide additional symptom reduction compared to sham neurofeedback. In the GRID-HAMD analysis, there was a strong main effect of time ($p = .001$) but no group effect ($p = .443$). The PHQ-9 analysis also showed a significant time effect ($p = .011$) and additionally revealed that participants in the ILF-NF group reported higher depression scores overall compared to those in the sham group ($p = .035$). No time-by-group interaction was detected for the PHQ-9.

Self-reported anxiety and stress symptoms measured by the DASS did not show significant group differences. A small overall time effect was detected for the DASS depression subscale ($p = .049$), but post-hoc tests only indicated a trend-level improvement between mid- and post-treatment. Psychological quality of life improved significantly over time ($p = .041$), with increases most evident between mid-treatment and follow-up, but this change did not differ between the ILF-NF and sham groups.

Cognitive performance measured by the CPT suggested a potential but non-significant advantage of ILF-NF. Mean reaction times showed a trend toward a differential change: participants in the ILF-NF group tended to respond faster post-treatment, while those in the sham group showed slight slowing ($p = .054$). Reaction time variability did not change significantly in either group. Electrophysiological outcomes, specifically P300 amplitude and latency, did not show any significant effects of time, group, or their interaction. No adverse effects related to neurofeedback training were reported.

5. Discussion

This randomized, double-blind pilot study examined whether ILF-NF could provide benefits beyond sham neurofeedback when added to usual inpatient or day-patient treatment for major depressive disorder. The main finding was that ILF-NF did not produce superior improvements in depressive symptoms. Both observer- and self-rated depression decreased over time in both groups, most likely reflecting the effects of treatment-as-usual and nonspecific therapeutic factors such as clinician attention, participant expectations, and placebo effects, as previously discussed in the neurofeedback literature (Thibault & Raz, 2017). Interestingly, self-rated depression was consistently slightly higher in the ILF-NF group compared to the sham group, though this difference did not interact with time and may reflect chance or baseline variability.

Secondary outcomes mirrored the primary findings. Anxiety and stress levels did not significantly change over time in either group, and improvements in psychological quality of life were observed but were similar in both conditions. The only signal of a potentially ILF-NF–specific effect appeared in the domain of cognitive performance, where reaction times showed a non-significant trend toward improvement in the ILF-NF group compared to slight worsening in the sham group. Although preliminary, this pattern suggests that ILF-NF may influence processing speed, an effect that should be investigated in larger and more statistically powerful studies. Electrophysiological measures, namely P300 amplitude and latency, did not change following the intervention, suggesting either that ILF-NF does not influence these cognitive markers or that the current study was underpowered to detect small effects.

The study has several strengths, including a rigorous double-blind, sham-controlled design, careful EEG preprocessing, and the integration of both clinical and neurophysiological outcome measures. Nonetheless, there are notable limitations. The sample size was small, with only twelve participants per group, limiting the power to detect subtle effects. The number of neurofeedback sessions was restricted to ten for feasibility reasons, while previous reports and clinical recommendations often suggest twenty or more sessions for optimal efficacy (Bazzana et al., 2022; Othmer & Othmer, 2016). Treatment-as-usual was not standardized and included varying therapeutic interventions and frequent medication changes, which could have influenced outcomes. Moreover, most participants had comorbid psychiatric disorders, and the sample may have included heterogeneous subtypes of depression that respond differently to ILF-NF (Price et al., 2017; Yang et al., 2021).

Future research should include larger, adequately powered randomized trials with extended ILF-NF training schedules. It would be beneficial to control or stratify for concomitant treatments and to examine whether specific depression subtypes, potentially defined by neuroimaging markers, respond differently to ILF-NF. Further exploration of the potential cognitive benefits indicated by reaction time findings is also warranted.

In conclusion, this pilot RCT found no evidence that ILF-NF confers additional benefit over sham neurofeedback in alleviating depressive symptoms when combined with standard inpatient or day-patient care. While depression and quality of life improved over time across all participants, these improvements were not specific to ILF-NF. A possible effect on processing speed remains a promising but unconfirmed observation that merits further investigation.

6. References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5®).

Angermeyer, M. C., Kilian, R., & Matschinger, H. (2000). WHOQOL-BREF.

Balt, K., Du Toit, P. J., Smith, M., & Janse van Rensburg, C. (2020). The effect of infraslow frequency neurofeedback on autonomic nervous system function in adults with anxiety and related diseases. NeuroRegulation, 7(2), 64–74.

Bazzana, F., Finzi, S., Di Fini, G., & Veglia, F. (2022). Infra-Low Frequency Neurofeedback: A Systematic Mixed Studies Review. Frontiers in Human Neuroscience, 16.

Dobrushina, O. R., Vlasova, R. M., Rumshiskaya, A. D., et al. (2020). Modulation of intrinsic brain connectivity by implicit EEG neurofeedback. Frontiers in Human Neuroscience, 14.

Fernández-Alvarez, J., Grassi, M., Colombo, D., et al. (2022). Efficacy of bio- and neurofeedback for depression: A meta-analysis. Psychological Medicine, 52(2).

Garcia-Marin, L. M., Mulcahy, A., Byrne, E. M., et al. (2023). Discontinuation of antidepressant treatment: A retrospective cohort study. Annals of General Psychiatry, 22(1), 49.

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of mental disorders. The Lancet Psychiatry, 9(2), 137–150.

Grin-Yatsenko, V. A., Othmer, S., Ponomarev, V. A., et al. (2018). Infra-Low Frequency Neurofeedback in Depression: Three case studies. NeuroRegulation, 5(1).

Hiltunen, T., Kantola, J., Abou Elseoud, A., et al. (2014). Infra-slow EEG fluctuations and resting-state network dynamics. Journal of Neuroscience, 34(2).

Kato, M., Hori, H., Inoue, T., et al. (2021). Discontinuation of antidepressants after remission. Molecular Psychiatry, 26(1).

Marzbani, H., Marateb, H. R., & Mansourian, M. (2016). Neurofeedback: A comprehensive review. Basic and Clinical Neuroscience, 7(2).

Moreno-Agostino, D., Wu, Y.-T., Daskalopoulou, C., et al. (2021). Global trends in depression. Journal of Affective Disorders, 281.

National Institute for Health and Care Excellence [NICE]. (2022). Depression in adults: Treatment and management (Guideline 222).

Othmer, S., & Othmer, S. F. (2016). Infra-low-frequency neurofeedback for optimum performance. Biofeedback, 44(2).

Patil, A. U., Lin, C., Lee, S.-H., et al. (2023). Review of EEG-based neurofeedback as a therapeutic intervention. Psychiatry Research: Neuroimaging, 329.

Price, R. B., Gates, K., Kraynak, T. E., et al. (2017). Data-driven subgroups in depression. Neuropsychopharmacology, 42(13).

Santopetro, N. J., Thompson, B., Garron, A., et al. (2025). Meta-analysis: Depression and P300. Neuroscience & Biobehavioral Reviews, 175.

Sitaram, R., Ros, T., Stoeckel, L., et al. (2017). Closed-loop brain training: The science of neurofeedback. Nature Reviews Neuroscience, 18(2).

Thibault, R. T., & Raz, A. (2017). The psychology of neurofeedback. American Psychologist, 72(7).

Voderholzer, U., Barton, B. B., Favreau, M., et al. (2024). Enduring effects of psychotherapy and antidepressants. Frontiers in Psychiatry, 15.

Yang, X., Kumar, P., Nickerson, L. D., et al. (2021). Subgroups of MDD and brain networks. Biological Psychiatry Global Open Science, 1(2).

Zhang, Z., Zhang, Y., Wang, H., et al. (2025). Resting-state network alterations in depression: A meta-analysis. Psychological Medicine, 55.

D6.2: Implementation and documentation of conducted studies – Insight Generator

1. General information

Netherlands: **Eating Disorder**

Name of technical system/component: Insight Generator

Study director/manager: 5M Software – Marko Petkovic

Contact (e-mail): marko.petkovic@5msoftware.nl

Aim of the validation: Validation of insight generator performance and adaptability

2. Methods

2.1 Motivation / Background

For people with eating disorders, tracking meals and activities is an important part of treatment and recovery. These logs can reveal valuable information, but on their own they are often difficult to interpret and may not provide enough guidance for patients to make meaningful changes.

The developed component addresses this challenge by analyzing the logged data and generating personalized, actionable insights. It can highlight behavioral patterns—for example, connections between skipped meals, activity levels, or certain times of day—and turn them into practical suggestions for improvement. This makes the data more understandable and directly useful to patients.

Over time, the system learns from individual user profiles, refining the insights it generates so they remain relevant and tailored. This not only supports patients in building healthier habits but also provides clinicians with structured information that strengthens treatment. In this way, the component contributes to the broader aim of digital health tools: turning raw data into meaningful feedback that drives positive change.

2.2 Study design

The aim of this study is to evaluate the performance of the insight generator (based on [1]), specifically its ability to provide relevant and actionable feedback while adapting to changing user interests over time. The central hypothesis is that the system can maintain high accuracy and usefulness even when user preferences evolve, reflecting realistic patterns of behavior in patients with eating disorders.

To test this, we create a simple probabilistic model that generates synthetic user behavior over a 120-day period. The model samples daily activities and nutrition events, such as having breakfast of *x* calories at 10:00, taking a walk of *y* minutes at a given speed and distance at 11:00, or logging other meals and physical activities. These events are drawn from probability

distributions based on internal collected data, ensuring that the simulated behaviors resemble realistic patterns without relying on sensitive patient information. Each synthetic user is simulated multiple times, providing a robust dataset for evaluation.

Every month, the type of insights that a user is “interested in” is changed. For example, during the first month, the insight generator may focus on regularity of meal times; in the second month, it may shift to balance between nutrition and activity; and later to caloric intake or other dimensions. This variation tests whether the system can adapt to shifting priorities while still producing accurate and relevant insights. The outcome measures therefore focus on two main aspects: (1) the relevance of the generated insights to the user’s current stated interest, and (2) the accuracy of these insights in reflecting the underlying simulated behavior.

Since this is a simulation study, concepts such as participant recruitment, randomization, or anonymization are not applicable. All “participants” are synthetic, and all input is generated through the probabilistic model. The intervention is defined as the daily operation of the insight generator, while the evaluation consists of systematically assessing its outputs against the simulated data and the user’s changing preferences. The duration is fixed at 90 simulated days per user, with preference changes occurring on a monthly basis.

In summary, this simple study design allows us to isolate and rigorously test the adaptability and accuracy of the insight generator. By combining a probabilistic model of user behavior with controlled changes in interest over time, we can evaluate whether the system is capable of providing meaningful, personalized insights under dynamic conditions.

2.3 Data acquisition

The data for this study comes from an internal pilot project in which lifestyle information was collected through daily logging in the app. Participants recorded their nutritional intake and activities directly in the app, which served as the main tool for data collection. This approach ensured consistency and provided time-stamped entries that captured behaviors as they occurred, rather than relying on retrospective self-reporting. The schedule of collection was daily, and the measured variables included caloric content and timing of meals, as well as details of physical activity such as duration, distance, and speed.

These data were not analyzed directly for clinical outcomes but were instead used to fit a probabilistic model of user behavior. By capturing realistic distributions of eating and activity patterns, the model could generate synthetic daily routines that reflect plausible behaviors, such as a breakfast of a certain caloric value at a typical time or a walk of a given length and intensity later in the day. This probabilistic approach provided the foundation for simulating users in the study and evaluating the performance of the insight generator under realistic but controlled conditions.

2.4 Study implementation

The study was implemented entirely as a simulation, with synthetic users generated through the probabilistic model described earlier. Each simulated day consisted of a series of nutrition and activity events sampled from realistic distributions. As the day progressed, lifestyle data was logged continuously, mirroring the way real participants would record information through an app. At several fixed hours during the day, the system evaluated potential insights by predicting an "interestingness score," which reflected how relevant or useful each insight might be when comparing the user’s current and past behavior over different time windows (days, weeks, or months).

To balance exploration and exploitation, insights were sampled based on their predicted scores: some high-scoring insights were selected to ensure relevance, while others were included to allow the system to test new possibilities. Each selected insight was then labeled as "interesting" or "not interesting" according to the simulated user’s current monthly preference profile, with a degree of random noise added to reflect the variability of real user

responses. This labeling provided the feedback signal needed to evaluate and refine the model's predictions.

At the end of each simulated day, the data collection process was concluded by updating the training and test sets. The model's performance was assessed against the test set, and the dataset was adjusted to maintain a balanced representation of labels. Older insights, defined as those more than two weeks old, were removed to ensure that the training set reflected more recent patterns and to prevent the model from overfitting to outdated behaviors. Following this, the model was fine-tuned on the updated dataset, completing the cycle of daily simulation, prediction, labeling, and retraining.

Data monitoring in this study focused on tracking quality of generated insights over time. Performance metrics were recorded systematically to assess whether the model improved in its ability to predict user interest as preferences shifted from month to month. Since all participants were simulated, there were no control groups or participant management procedures in the traditional sense. Instead, variability was introduced through controlled changes in preference profiles, serving as a form of internal test condition to evaluate adaptability under different scenarios.

3. Analysis plan

The analysis focuses on evaluating the performance of the insight generator over time as user preferences change. The main outcome measure is accuracy, defined as the proportion of correctly predicted "interesting" or "not interesting" labels for generated insights, tracked across the 120 simulated days. In addition to this quantitative measure, selected insights are qualitatively inspected to confirm that the outputs are meaningful and relevant.

All data used in the analysis comes from the synthetic datasets generated by the probabilistic model, consisting of daily logs of nutrition and activities paired with system-generated insights and feedback labels. Since the data is simulated, missing values and outliers are not present, though some label noise is intentionally added to mimic real-world variability. Preprocessing involves organizing the simulation output into training and test sets, with insights older than two weeks dropped before retraining the model.

4. Results

The main finding of the study is that the insight generator adapts effectively to changing user preferences. When the type of insight a user is interested in changes at the start of a new month, the model initially experiences a slight drop in performance but recovers within a few days, demonstrating its ability to learn and realign with the updated priorities. Across multiple simulation runs, the average accuracy of the model is approximately 90%, showing consistently strong performance.

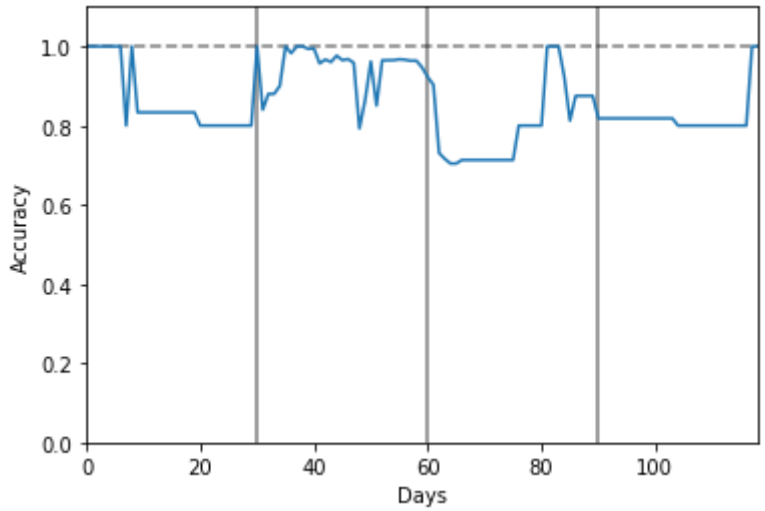


Figure 1 Accuracy of insights over time

An example run is illustrated in Figure 1, which shows the temporary dips in accuracy following preference changes, while performance never drops below 80%. This pattern is consistent across different simulated users and repeated runs, highlighting the robustness and reliability of the model under dynamic conditions.

No significant side or unintended effects were observed in the simulation. Minor fluctuations in accuracy occur due to the intentionally added label noise, but these are small and do not indicate systematic bias or model failure. Qualitative inspection of selected insights further confirms that the outputs remain meaningful and aligned with user interests throughout the simulation period.

5. Discussion

The results of this study demonstrate that the insight generator is capable of adapting to changing user preferences and providing relevant feedback in a simulated environment. Across multiple runs, the model achieved an average accuracy of approximately 90%, and even during preference shifts, performance did not drop below 80%. These findings indicate that the system is robust and able to learn user priorities over time, maintaining meaningful and actionable output.

The interpretation of these results suggests that the approach of combining a probabilistic model of user behavior with adaptive insight generation is effective for capturing and responding to dynamic preferences. While the current study is limited to simulated data with simplified representations of nutrition and activity patterns, the consistent performance indicates that the underlying methodology is sound and capable of supporting personalization in real-world settings.

In practical terms, these findings have implications for the development of digital health tools for patients with eating disorders. The ability to dynamically tailor insights to individual behaviors could enhance engagement, support self-awareness, and facilitate positive behavior change. From a research perspective, the study provides a foundation for further exploration into adaptive recommendation systems in health contexts, particularly those that need to respond to evolving user priorities.

The main strength of this study is its controlled evaluation of the insight generator under well-defined conditions, allowing clear assessment of its adaptation capabilities. However, the study is also limited by its reliance on simulated users, which cannot fully capture the complexity,

variability, and unpredictability of real human behavior. Additionally, the input data used for the probabilistic model is limited in scope, containing only basic nutrition and activity information.

Future directions include extending the model to incorporate richer user profiles and additional contextual information, as well as developing agents capable of more personalized analysis and communication of insights. This could include adaptive explanations, prioritization of actionable feedback, or integration with real-time monitoring data. Ultimately, the goal is to transition from simulation-based evaluation to real-world testing with actual users, which would allow validation of both the accuracy and the practical impact of the insights.

In conclusion, the study provides promising evidence that the insight generator can adapt to changing preferences and produce relevant feedback in a controlled simulation. While further work is needed to validate these findings in real-world settings, the results highlight the potential of adaptive, personalized insight systems in supporting behavior change and improving digital health interventions.

6. References

[1] Susaiyah, Allmin, et al. "Neural scoring of logical inferences from data using feedback." *International Journal of Interactive Multimedia and Artificial Intelligence* 6.5 (2021): 90-99.

D6.2 document: Implementation and documentation of conducted studies - Treatment Response Prediction in Major Depressive Disorder

1. General information

Country and name of use case: Netherlands: Major Depressive Disorder

Name of technical system/component: Multimodal Machine Learning Model for Treatment Response Prediction Study director/manager: AMC, University of Amsterdam - Dr. Liesbeth Reneman

Contact (e-mail): l.reneman@amsterdamumc.nl

Additional partners:

- Radboudumc, Nijmegen, the Netherlands (Henricus G. Ruhe)
- Western Norway University of Applied Sciences, Bergen, Norway (Ivan I. Maximov, Atle Bjørnerud)
- Vestfold Hospital Trust, Tønsberg, Norway (Inge R. Groote)
- Oslo University Hospital, Oslo, Norway (Inge R. Groote, Atle Bjørnerud)

Link to publication: https://pubmed.ncbi.nlm.nih.gov/38321916/

Aim: The aim was to determine whether a multimodal machine learning approach could predict early treatment response to the antidepressant sertraline in patients with major depressive disorder (MDD). The study also sought to assess the specific predictive contributions of MR neuroimaging and clinical assessments collected at baseline and after one week of treatment.

Follow-up: This study is a secondary analysis of data from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) study.

2. Methods

This section details the methodology used in the study, including the background motivation, the specific design of the clinical trial data being analyzed, the procedures for data acquisition, and the practical steps taken for study implementation.

2.1 Motivation / Background

Major Depressive Disorder (MDD) represents a significant global health challenge, being the second largest contributor to disability worldwide. A major difficulty in its treatment is the high variability in patient response to antidepressants. Currently, there are no objective predictors to guide the selection of treatment for an individual. As a result, initial antidepressant therapy leads to remission in only about a third of patients. This often forces patients to cycle through multiple sequential treatments and combinations, a process that prolongs the disease burden, exposes them to potential side effects, and incurs high societal costs. To shorten this trial-and-error process and expedite remission, there is a critical need for clinically valuable biomarkers that can indicate an individual's likely treatment response either before or very early after starting a new medication.

Neuroimaging with MRI has emerged as a promising avenue for discovering such predictive biomarkers. Over the years, several potential neuroimaging markers have been proposed, but their application in standard clinical practice remains distant. This is largely because existing studies are often heterogeneous, report small effect sizes, and tend to rely on single (unimodal) neuroimaging measures. To address these shortcomings, it has been suggested that combining individual neuroimaging predictors with other types of information, such as clinical data, using machine learning techniques could create a larger predictive effect and increase the robustness of the findings. Despite this recommendation, a review found that very few studies have adopted such a multimodal approach. This study builds on recent work that has shown the benefit of combining clinical and imaging data , taking it a step further by incorporating all clinically applicable MRI modalities with known predictive value to develop a comprehensive multimodal prediction model. The use of data from a large randomized clinical trial with a placebo arm allows for robust validation and testing of the model's specificity to sertraline versus a placebo effect.

2.2 Study design

This section provides a detailed overview of the study's design, outlining the core objectives and hypotheses, the measures used as inputs and outcomes, the participant characteristics, and the structure of the intervention.

The study was a preregistered secondary analysis of data from the Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care (EMBARC) study. The original EMBARC trial was a multisite, double-blind, placebo-controlled randomized clinical trial designed to identify predictors of antidepressant response.

The study had several clear objectives and hypotheses. The primary hypothesis was that a multimodal machine learning approach could predict sertraline treatment response significantly better than chance (i.e., the a priori response rates). A second hypothesis was that the predictive model would be specific to sertraline treatment; therefore, it was expected that the model's performance would be reduced when tested on placebo-treated patients but would remain high for placebo nonresponders who were subsequently treated with sertraline. The final hypothesis was that a multimodal approach integrating different data types would outperform unimodal models that rely on a single type of data.

The model's input measures consisted of data collected before treatment (pretreatment) and one week after the start of treatment (early treatment). These inputs included a wide range of clinical assessments (sociodemographic, behavioral, and neuropsychological) and multimodal MRI neuroimaging data. The specific MRI modalities used were T1-weighted structural MRI, diffusion-weighted imaging (DWI), resting-state functional MRI (rs-fMRI), and perfusion MRI using arterial spin labeling (ASL). DWI was only acquired at baseline. The primary outcome measures were treatment response and remission after eight weeks of treatment. Response was defined as a reduction of 50% or more in the 17-item Hamilton Depression Rating Scale (HAM-D) score, while remission was defined as a final HAM-D score of 7 or less.

The study analyzed data from an initial pool of 296 adult outpatients aged 18-65 with unmedicated recurrent or chronic MDD. After applying exclusion criteria, a total of 229 patients were included in the final analyses. The exclusion criteria included having missing data for the primary outcome at weeks 7 and 8, missing data for more than one MRI sequence, or not adhering to the pharmacological treatment

for at least two weeks. The study defined three key population subgroups for its analyses, as detailed in the CONSORT flow diagram (see Figure 1).

The original study was double-blind and involved randomization to either sertraline or a placebo control group. The trial's duration consisted of two 8-week phases. The intervention was treatment with either sertraline or placebo. In the second phase, nonresponders to placebo were switched to sertraline. The placebo group served as the primary control group for testing the model's specificity.

2.3 Data acquisition

The following section outlines the procedures and tools used for data collection in the study. Data was gathered at two main time points: before the start of treatment (baseline) and again one week after treatment began ("early treatment").

The tools and instruments used for data collection were comprehensive. They included a battery of clinical assessments covering sociodemographic, behavioral, and neuropsychological domains. Key questionnaires mentioned were the 17-item Hamilton Depression Rating Scale (HAM-D) for measuring symptom severity, the Mood and Anxiety Symptom Questionnaire (MASQ), and the Snaith-Hamilton Pleasure Scale (SHAPS) for assessing anhedonia. The neuroimaging component utilized several MRI methods. The four main sequences acquired were:

- T1-weighted (T1w) structural MRI
- Diffusion-weighted imaging (DWI)
- Resting-state functional MRI (rs-fMRI)
- Perfusion MRI via arterial spin labeling (ASL)

The data collection schedule involved acquiring the clinical and MRI data at baseline and at the one-week early treatment time point. The primary treatment outcomes, which included depression scores (HAM-D), quality of life metrics, and binary classifications of response and remission, were collected after 8 weeks of treatment for phase 1 and after 16 weeks for phase 2. A notable aspect of the data acquisition was that while task-based fMRI was also collected during the EMBARC trial, it was deliberately excluded from this analysis due to the practical challenges it poses for scalable clinical application.

2.4 Study implementation

This section describes the practical implementation of the study, from the overall process to participant consent and data management.

The study was implemented as a secondary analysis of the EMBARC trial data. In the original trial, participants were recruited and provided written informed consent after receiving a complete description of the study procedures. The study process involved randomizing these participants to receive either sertraline or a placebo for an 8-week period (phase 1). Following this, placebo nonresponders were enrolled in a second 8-week phase where they were treated with sertraline.

Participant recruitment resulted in the inclusion of 296 outpatients with MDD. The implementation of the data collection process was highly standardized, especially for the MRI data. Extensive preprocessing

pipelines were used for each MRI modality to ensure data quality and consistency across sites. This included using FreeSurfer and FastSurfer for T1-weighted scans, tract-based spatial statistics for DWI, fMRIPrep for rs-fMRI, and ExploreASL for perfusion data.

Data monitoring was a key part of the implementation. Systematic quality control (QC) was performed on the MRI data using automated tools like MRIQC and EDDY QC, supplemented by visual inspection of all segmentations. The generated predictors also underwent QC to check for missing values and outliers. The study carefully managed its experimental and control groups, which were defined as three distinct subgroups for analysis: Subgroup A (sertraline), Subgroup B (placebo control), and Subgroup C (sertraline-treated placebo nonresponders). The flow of participants through these groups is visually detailed in the study's CONSORT diagram (see Figure 1).

3. Analysis plan

This section defines the statistical and computational methods used to analyze the collected data. It covers the software tools, dataset definitions, data preparation steps, and the statistical models employed to test the study's hypotheses.

The analysis relied on a number of software packages and tools. The machine learning models were implemented in Python using the XGBoost library. Feature selection was performed using tools from Scikit-learn. The extensive MRI data preprocessing pipeline utilized specialized software including FreeSurfer, FastSurfer, PyRadiomics, MRIQC, EDDY QC, fMRIPrep, FSL, and ExploreASL.

The datasets for analysis were derived from the main EMBARC study cohort and organized into three subgroups: Subgroup A (sertraline-treated, n=109), Subgroup B (placebo-treated, n=120), and Subgroup C (sertraline-treated placebo nonresponders, n=58). The predictors from these datasets were further organized into three tiers based on their level of prior scientific evidence. Tier 1 included a small set of six MRI predictors with strong evidence from meta-analyses, plus key clinical variables. Tier 2 added 54 predictors with weaker evidence. Tier 3 included all 240 predictors without any pre-selection.

The analysis plan included specific procedures for handling missing data and outliers. Missing sequence-specific MRI predictors were imputed using K-nearest neighbor imputation, which was performed carefully to prevent information leakage across different MRI sequences. Missing HAM-D outcome scores were imputed using an iterative multivariate conditional modeling approach. Data preprocessing was a critical step. To mitigate confounding effects from the different acquisition sites, an open-source harmonization method called ComBat was used, with adjustments for age, sex, brain volume, and site. The data was also scaled, and to handle imbalanced outcomes, the minority class was randomly oversampled.

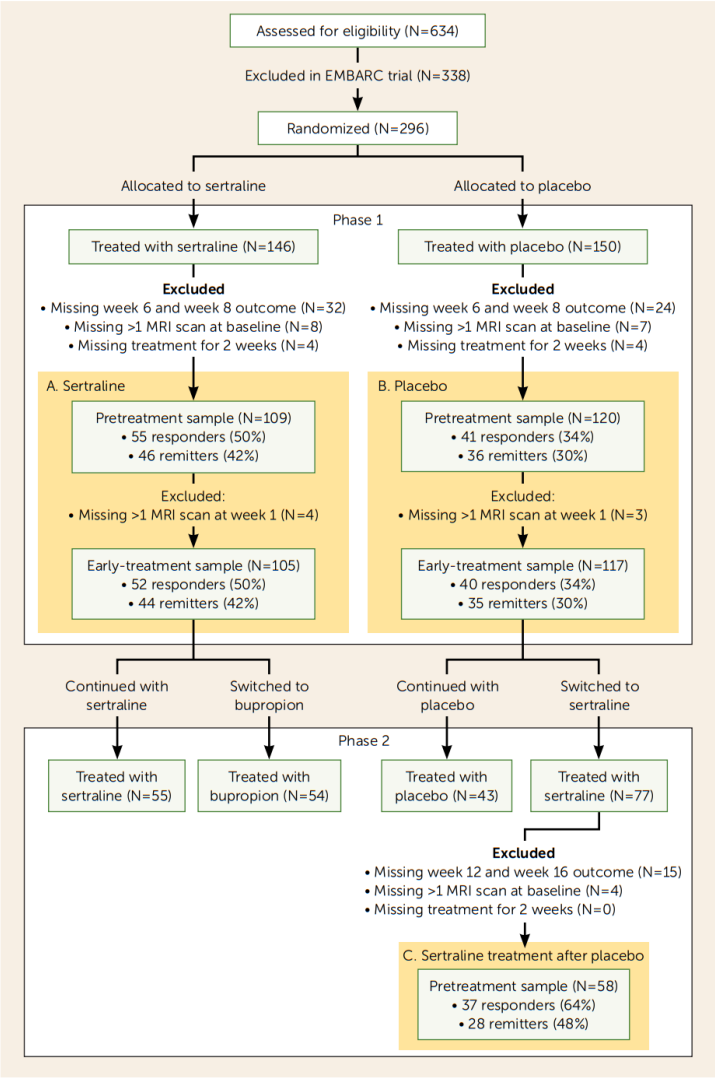
The core of the analysis involved statistical methods and models built within a nested cross-validation framework to reduce the risk of bias and overfitting. The primary machine learning model was an extreme gradient boosting classifier (XGBoost). Bayesian hyperparameter optimization was used to tune the model, and chi-square-based feature selection was embedded in this process. For inferential and descriptive statistics, the primary performance metrics were balanced accuracy (bAcc) and the area under the receiver operating characteristic curve (AUROC). Sensitivity and specificity were also reported. To test the study's hypotheses, one-tailed binomial tests were used to compare model performance against chance , one-sided dependent t-tests were used to assess treatment specificity , and one-tailed

sign tests were used to compare multimodal versus unimodal model performance. An assumption check was performed to ensure that the exclusion of patients did not significantly alter the population characteristics, using Student's t-tests and chi-square tests. The plan also included several exploratory (post hoc) analyses, such as comparing the XGBoost classifier with other models like support vector machines, and attempting to predict continuous outcome scores directly instead of binary classes.

4. Results

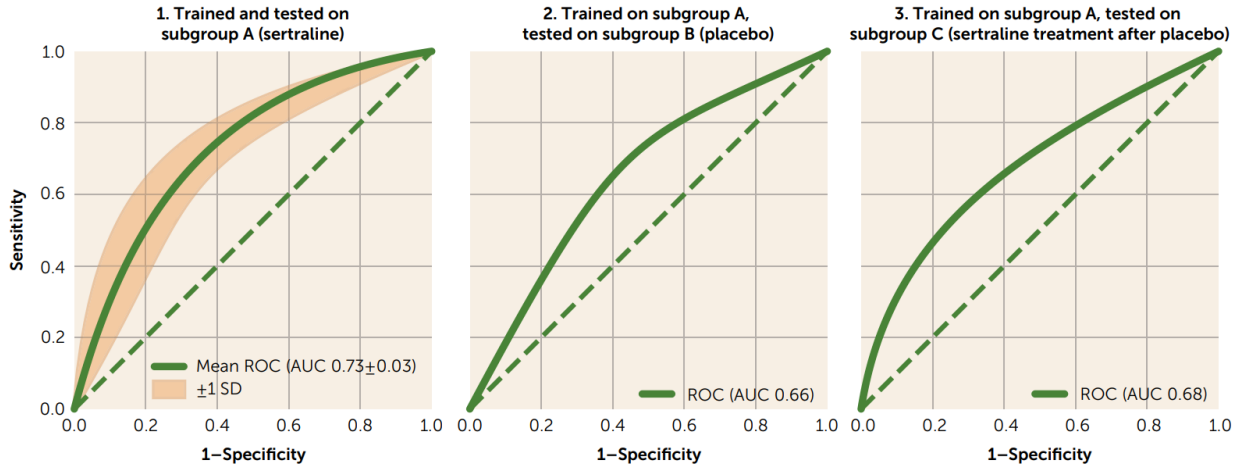
This section presents the primary outcomes and findings of the analysis, covering the main results of the predictive modeling, references to visual representations of these results, and a consideration of potential side effects.

FIGURE 1. CONSORT flow diagram for a study of treatment response prediction in major depressive disorder using multimodal MRI and clinical data[a]



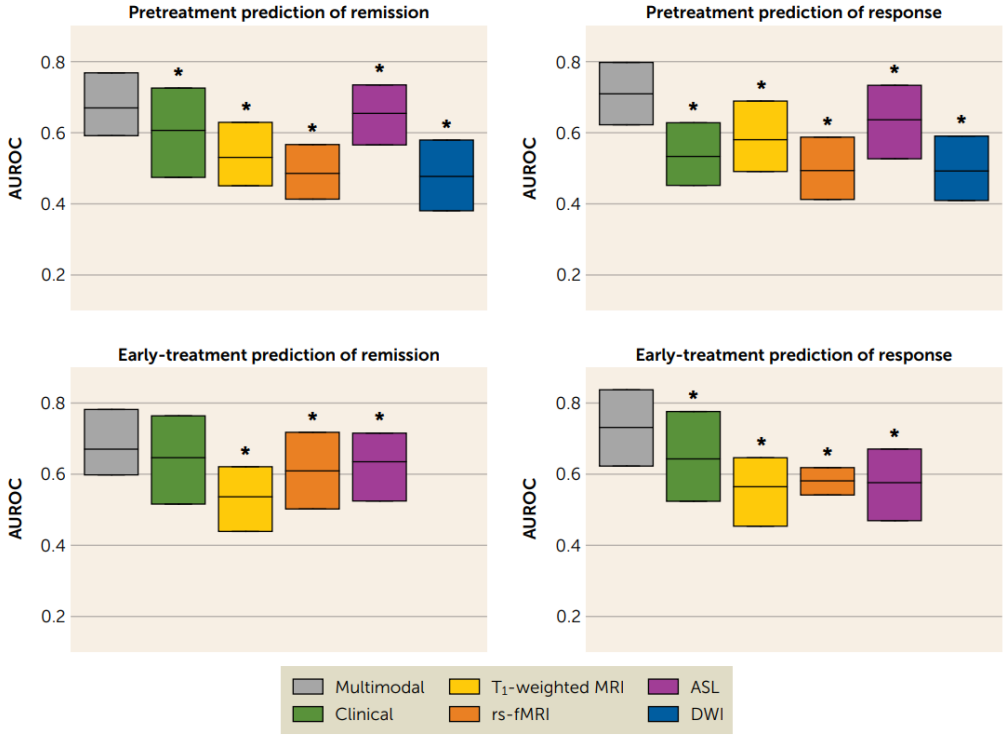
[a] The EMBARC clinical trial consists of two study arms and two 8-week phases. One arm was randomized to sertraline treatment and the other to placebo. Highlighted are three population subgroups used in our analyses. The first two subgroups, in phase 1 of the study, are those treated with sertraline (A) or placebo (B). The third subgroup consists of patients who did not respond to placebo treatment in phase 1 and were switched to treatment with sertraline in phase 2 of the study (C).

FIGURE 2. Receiver operating characteristic curves of the best-performing configuration of each analysis[a]



[a] Panel A shows the mean receiver operating characteristic curve (ROC) over 10 folds from subgroup A for early-treatment response prediction using tier 1 predictors; panel B shows early-treatment prediction of remission on subgroup B using tier 3 predictors; and panel C shows pretreatment prediction of remission on subgroup C using tier 1 predictors. AUC=area under the curve.

FIGURE 3. Box plots of best cross-validation performance of multimodal versus unimodal models[a]



[a] Boxes show two interquartile ranges with a median centerline. An asterisk marks unimodal cross-validation results that are significantly ($p<0.05$) worse than their multimodal counterpart. See Table S12 in the online supplement for statistical outcomes. ASL=arterial spin labeling; AUROC=area under the receiver operating characteristic curve; DWI=diffusion-weighted imaging; rs-fMRI=resting-state functional MRI.

The final analysis included 229 patients, with a mean age of 38.1 years and 65.9% being female. The exclusion of participants with missing data did not significantly affect the overall characteristics of the analyzed population.

The main findings of the study support the primary hypotheses. Firstly, the multimodal machine learning models were able to predict treatment response to sertraline with performance significantly better than chance. The best performance was achieved for the early-treatment prediction of response, which yielded a balanced accuracy (bAcc) of 68% and an AUROC of 0.73. The results from this analysis are visualized with receiver operating characteristic (ROC) curves (see Figure 2). Secondly, the models demonstrated specificity for sertraline treatment. When the model trained on sertraline patients was tested on placebo-treated patients (Subgroup B), its performance was significantly reduced. However, when tested on placebo nonresponders who were then switched to sertraline (Subgroup C), the predictive performance was not significantly lower than the original internal validation.

Thirdly, the results confirmed that multimodal models consistently outperformed unimodal models. A visual comparison of the performance of multimodal versus various unimodal models is provided via box plots (see Figure 3). Of the unimodal models, those using only arterial spin labeling (ASL) perfusion data for pretreatment prediction and those using only clinical assessment data for early-treatment prediction were the only ones that performed significantly better than chance. In terms of predictor importance, ASL perfusion metrics in the anterior cingulate cortex (ACC) were the most important predictors at pretreatment. For early-treatment prediction, the most important predictor shifted to the relative reduction in HAM-D score after one week. The flow of participants through the trial, which forms the basis of these results, is detailed in a CONSORT diagram (see Figure 1). Additional detailed results are available in supplementary materials, including post hoc analyses of regression models and feature importance (See Figures S1 and S2 in the online supplement).

As this study was a secondary analysis focused on developing and validating a predictive model from existing data, it did not involve administering interventions directly. Therefore, possible side or unintended effects related to the modeling approach itself are not applicable. The original clinical trial involved antidepressant medication, which is associated with side effects, but these were not a subject of this analytical study.

5. Discussion

This section provides an interpretation of the study's results, discusses their broader implications, acknowledges the strengths and limitations of the research, and suggests directions for future work before drawing final conclusions.

A summary of the main findings confirms that it is feasible to predict sertraline treatment response in patients with MDD using a combination of brain MRI and clinical data, with models performing significantly better than chance. The results also strongly suggest that these predictive models are specific to sertraline's pharmacological effect, as opposed to a general placebo effect. Furthermore, the study demonstrated that integrating data from multiple sources (multimodal) yields superior performance compared to using a single data type (unimodal), with arterial spin labeling (ASL) perfusion imaging standing out as the most powerful unimodal predictor from pretreatment data.

The interpretation of these results is that they build upon and improve previous work in the field. Compared to a recent similar study by Sajjadian et al., this work showed better predictive accuracy, which may be attributable to the inclusion of ASL data, a modality not available in the other study. The model's performance also exceeded the average benchmarks reported in recent meta-analyses of treatment prediction. The finding that perfusion in the anterior cingulate cortex (ACC) is a key predictor

aligns with existing theories about MDD, as the ACC is a critical hub in brain circuits that regulate emotion and are modulated by SSRIs like sertraline. The importance of early symptom reduction after one week is also consistent with prior literature, confirming it as a known and powerful predictor.

The implications of the study for both theory and practice are significant. In clinical practice, a validated tool based on this approach could help clinicians individualize treatment planning, moving away from the current trial-and-error standard. Such a tool could predict treatment efficacy early on, potentially justifying the costs of MRI scanning by preventing prolonged periods of ineffective treatment. For future research, the clear implication is the need for further validation. The model must be tested on larger, external datasets, for different antidepressants, and in more clinically diverse patient populations to ensure generalizability.

This study has notable strengths, including its use of the EMBARC dataset, which is the largest multimodal neuroimaging dataset for MDD available, reducing the risk of performance bias. The data was acquired across multiple sites, and advanced harmonization techniques were used, which increases the potential for the results to generalize to new populations. Another key strength is the preregistration of the study's hypotheses and methods, which minimizes the risk of overfitting and enhances the validity of the findings. However, several limitations must also be acknowledged. The results are currently limited to sertraline treatment and the specific data acquired in the EMBARC study. The analysis may have a bias towards treatment-adherent participants, as those who dropped out of the original trial were excluded. A potential selection bias also exists in the placebo nonresponder group (Subgroup C), although analysis suggests this did not significantly affect the conclusions. Finally, potentially useful data from task-based fMRI was excluded due to practical scalability issues, so its predictive value could not be assessed.

Future directions for this research include the crucial step of external validation, as mentioned. Future work could also focus on creating a leaner, more cost-effective model by combining the most powerful predictors, such as absolute ASL predictors with early clinical data. Other avenues for improvement include integrating predictors from other sources, like genetic data or information about previous treatment responses, and utilizing novel analytic methods such as normative modeling.

In conclusion, the study successfully demonstrates that a machine-learning-based method using multimodal MRI and clinical data can feasibly predict sertraline treatment response in MDD patients. The approach significantly outperforms chance and most unimodal models. The findings also provide evidence for the model's specificity to sertraline compared to placebo. With further external validation, this work could contribute to the development of predictive tools that help individualize the clinical treatment of MDD, ultimately improving patient care.

6. References

{1} Bromet E, Andrade LH, Hwang I, et al: Cross-national epidemiology of DSM-IV major depressive episode. BMC Med 2011; 9:90

{2} Kraemer HC: Messages for clinicians: moderators and mediators of treatment outcome in randomized clinical trials. Am J Psychiatry 2016; 173:672-679

{3} Trivedi MH, Rush AJ, Wisniewski SR, et al: Evaluation of outcomes with citalopram for depression using measurement-based care in STAR D: implications for clinical practice. Am J Psychiatry 2006; 163:28-40

{4} Al-Harbi KS: Treatment-resistant depression: therapeutic trends, challenges, and future directions. Patient Prefer Adherence 2012; 6:369-388

{5} Greenberg PE, Fournier AA, Sisitsky T, et al: The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry 2015; 76:155-162

{6} Phillips ML, Chase HW, Sheline YI, et al: Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: neuroimaging approaches. Am J Psychiatry 2015; 172:124-138

{7} Lee Y, Ragguett RM, Mansur RB, et al: Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 2018; 241:519-532

{8} Schrantee A, Ruhe HG, Reneman L: Psychoradiological biomarkers for psychopharmaceutical effects. Neuroimaging Clin N Am 2020; 30:53-63

{9} Schmaal L: The search for clinically useful neuroimaging markers of depression: a worthwhile pursuit or a futile quest? JAMA Psychiatry 2022; 79:845-846

{10} Cohen SE, Zantvoord JB, Wezenberg BN, et al: Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: a systematic review and meta-analysis. Transl Psychiatry 2021; 11:168

{11} Sajjadian M, Uher R, Ho K, et al: Prediction of depression treatment outcome from multimodal data: a CAN-BIND-1 report. Psychol Med 2022; 53:5374-5384

{12} Poirot MG, Ruhe HG, Maximov II, et al: Pretreatment and early-treatment prediction of SSRI treatment response in major depressive disorder: a multimodal MRI approach on the EMBARC dataset. Open Science Framework Registry, 2022. (https://doi.org/10.17605/OSF.IO/79CWY)

{13} Trivedi MH, McGrath PJ, Fava M, et al: Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC): rationale and design. J Psychiatr Res 2016; 78:11-23

{27} Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007; 8:118-127

{28} Chen T, Guestrin C: XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp 785-794

{31} Pedregosa F, Varoquaux G, Gramfort A, et al: Scikit-learn: machine learning in Python. J Mach Learn Res 2011; 12:2825-2830

{32} Combrisson E, Jerbi K: Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods 2015; 250:126–136

{33} Dixon WJ, Mood AM: The statistical sign test. J Am Stat Assoc 1946; 41:557–566

{36} Sajjadian M, Lam RW, Milev R, et al: Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychol Med 2021; 51:2742–2752

{38} Burke MJ, Romanella SM, Mencarelli L, et al: Placebo effects and neuromodulation for depression: a meta-analysis and evaluation of shared mechanisms. Mol Psychiatry 2022; 27:1658–1666

{42} Tejavibulya L, Rolison M, Gao S, et al: Predicting the future of neuroimaging predictive models in mental health. Mol Psychiatry 2022; 27:3129–3137

{43} Pizzagalli DA: Frontocingulate dysfunction in depression: toward biomarkers of treatment response. Neuropsychopharmacology 2011; 36:183–206

{44} Szegedi A, Jansen WT, van Willigenburg APP, et al: Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta analysis including 6562 patients. J Clin Psychiatry 2009; 70:344–353

{46} Orlhac F, Eertink JJ, Cottereau AS, et al: A guide to ComBat harmonization of imaging biomarkers in multicenter studies. J Nucl Med 2022; 63:172–179

{47} Paul R, Andlauer TFM, Czamara D, et al: Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Transl Psychiatry 2019; 9:187

{52} Rutherford S, Barkema P, Tso IF, et al: Evidence for embracing normative modeling. eLife 2023; 12:e85082

5.1 Source Documents

https://pubmed.ncbi.nlm.nih.gov/38321916/