Mass General Brigham

# Institutional Review Board
# Intervention/Interaction Detailed Protocol

| | |
|---|---|
| Principal Investigator: | Paola Pedrelli, PhD |
| Project Title: Depressive Disorder | Longitudinal assessment of biomarkers among patients with Major |
| Version Date: | 10.29.26 |
| Version Name/Number: | 1.2 |

## 1. Background and Significance

### Historical Background

Mental health disorders represent humanity's most critical health challenge, consuming 4% of global GDP (approximately US $5 trillion annually) and bringing immense suffering to millions of patients and their caregivers worldwide. Major Depressive Disorder (MDD) affects over 280 million people worldwide and is a leading cause of disability. Traditional face-to-face assessments rely on unstructured meetings, repetitive subjective self-report and potential clinician administered scales. These assessments are unevenly adopted due to the inherent time-consuming and burdensome efforts for patients and clinicians and remain prone to recall bias. Moreover, patient self-reporting tools face significant limitations, with real-world studies showing that only 19% of patients continue using mental health apps beyond six weeks, largely due to motivational deficits inherent to many mental health conditions. Anosognosia, the clinical inability of patients to recognize their own illness, further compromises accurate reporting. Yet, when clinicians use systematically standardized assessments, outcomes improve dramatically, with patients suffering from severe depression achieving remission three times more frequently and recovering twice as fast compared to standard care (Guo et al., 2015; Trivedi et al., 2006).

The solution to these assessment challenges is to unlock *reliable measurement-based care at scale*. A potential avenue and solution to this problem is “digital phenotyping”: digital phenotyping refers to the use of personal digital devices to collect objective behavioral data. This enables the continuous monitoring of symptoms in contrast with episodic assessment (Insel, 2017; Torous et al., 2016). Speech patterns (voice and language) are particularly promising markers since they can be collected through brief tasks with almost all devices, with little burden on patients, and especially reflect the underlying neurobiological changes associated with depression. Indeed, speech is the principal medium of psychiatric evaluation, and face-to-face conversation remains the core of psychiatric evaluation. Indeed, clinicians naturally analyze both what patients say and how they speak, looking for characteristics of speech disorganization or psychomotor changes (markers that represent even diagnostic criteria in schizophrenia and bipolarity). Besides, it has been found that depression diagnosis is associated with reduced pitch variability, slower speech rate, increased pause duration, and changes in voice quality (Cummins et al., 2015; Low et al., 2020; Mundt et al., 2012). These vocal features correlate with depression severity and change with treatment, suggesting sensitivity to clinical improvement (Mundt et al., 2012; Williamson et al., 2013). Machine learning approaches have successfully distinguished depressed individuals from non-depressed individuals with accuracy rates exceeding 70% in multiple studies (Alghowinem et al., 2012;

Taguchi et al., 2018; Riad et al., 2024). Depression rarely presents in isolation; it is frequently accompanied by residual symptoms such as insomnia and anxiety, even among patients who initially respond to treatment. These residual symptoms are not only burdensome in their own right, but they also serve as strong predictors of relapse and treatment failure (Carney et al., 2007; Fava et al., 2003). For instance, persistent residual insomnia increases the risk of depressive relapse by up to two-fold, while residual anxiety has been shown to delay full remission and worsen long-term outcomes. Thus, timely detection and monitoring of these symptoms is clinically essential.

Speech biomarkers, given their sensitivity to both cognitive-emotional and psychomotor processes, may offer a non-invasive window into the severity of residual insomnia or anxiety symptoms. Prior studies have shown that speech patterns can reflect anxiety states through markers of tension or hesitancy, while sleep disruption correlates with changes in vocal fatigue and speech fluency (Low et al., 2020; Cohn et al., 2015). Incorporating speech-based monitoring for insomnia and anxiety alongside core depressive symptoms has the potential to not only identify responders to treatment but also flag individuals at high risk of early relapse.

Ketamine and esketamine represent breakthrough treatments for treatment-resistant depression, a condition affecting approximately one-third of patients with major depression who do not achieve remission after multiple treatment trials (Gaynes et al., 2009). Unlike traditional antidepressants that require weeks to months for therapeutic effect, ketamine and esketamine demonstrate rapid onset of action, with significant symptom reduction occurring within hours to days (Zarate et al., 2006). Multiple randomized controlled trials have demonstrated significant antidepressant effects, including rapid reduction in suicidal ideation (Wilkinson et al., 2018). Importantly, residual insomnia and anxiety symptoms following initial treatment response are strong predictors of relapse in depression, making their early detection through speech biomarkers clinically valuable for relapse prevention.

The rapid onset of ketamine action is an ideal context for validating speech-based models, given the rapid evolution of symptom severity. The compressed timeframe of therapeutic response allows researchers to test whether speech features are sensitive to clinical change, to understand early predictors of treatment response, and to develop real-time monitoring tools. If speech biomarkers can detect the rapid changes associated with ketamine treatment, this would provide strong evidence for their utility in monitoring slower-acting treatments and tracking the natural course of depression. Similarly, a tumor necrosis factor inhibitor also exerts rapid-acting anti-inflammatory effects and has been shown to reduce depressive symptoms severity and anhedonia when delivered to depressed individuals with heightened inflammation Treadway et al., 2025; Mac Giollabhui et al., 2024).

**Experience in the field of Sponsor**

Callyope is a digital health company developing smartphone-based speech analysis technology for psychiatric assessment. The company is currently conducting 7 clinical trials in France and the UK to validate speech biomarkers across different psychiatric populations and treatment contexts (Riad et al. 2024, Ando et al. 2025, Roquefort et al. 2025), and plans for a Class II-A CE Marking in 2026. Callyope's approach addresses the measurement gap in mental health care by providing an objective, minimally burdensome assessment that can be completed remotely through patients' own smartphones, supporting the implementation of measurement-based care at scale.

The current study at Massachusetts General Hospital extends this validation to a clinical population receiving ketamine or esketamine treatment, as well as a TNF inhibitor for depression in a United States healthcare setting.

# 2. Specific Aims and Objectives

**Objectives**. The primary objective is to evaluate whether the Callyope model can (1) classify lower vs. moderate-to-severe depressive symptom severity and (2) estimate clinician-rated MADRS (Montgomery-Åsberg Depression Rating Scale) severity as a continuous score over 4 weeks. Secondary objectives include evaluating three-category classification, individual MDD symptoms, change in depressive symptom severity, treatment response, adherence to remote speech tasks, and the exploratory contribution of passive behavioral data. All clinical care proceeds according to standard practice and is not altered by this protocol.

**Endpoint 1 (co-primary):** Can the model discriminate lower depressive symptom severity (MADRS <20) from moderate-to-severe symptoms (MADRS ≥20) using paired speech and clinician-administered MADRS assessments at weeks 0 and 4? Three-class severity classification (MADRS <20, 20-34, and ≥35) and severe versus non-severe classification will be evaluated as secondary analyses.

**Endpoint 2 (co-primary):** How strongly does the model-predicted MADRS continuous score correlate with the clinician-administered MADRS total score using paired assessments at weeks 0 and 4

**Endpoint 3 (secondary):** Can the model predict categorical severity levels of residual symptoms such as insomnia, anxiety, anhedonia, sadness, psychomotor changes, fatigue, attention and suicidal thoughts based on AIS scores, GAD-7 scores, and MADRS and QIDS sub-items collected at week 0, 2 and 4?

**Endpoint 4 (secondary):** Can the model assess treatment response, defined as a ≥50% reduction in MADRS total score at week 4 and a ≥50% reduction in QIDS total score at week 2?

**Endpoint 5 (secondary):** Can the model discriminate mild depression (MADRS <20) from moderate severe symptoms (MADRS between 20 and 34) from severe (MADRS ≥35) using paired speech and clinician-administered MADRS assessments at weeks 0 and 4?

**Endpoint 6** (exploratory): Can the model predict the level of global depressive symptoms' severity (mild vs moderate vs. severe) based on the self-reported QIDS scores collected at week 0, 2 and 4?

**Endpoint 7** (exploratory): How strongly does the model-predicted QIDS continuous score correlate with the patient-reported QIDS total score using paired assessments at weeks 0, 2 and 4

**Endpoint 8** (exploratory): Can the combination of speech and passive biomarker analysis related to sleep and activity levels improve the performance of aims 1 to 5?

**Hypothesis 1:** The locked class probabilities derived from speech will discriminate lower depressive symptom severity (MADRS <20) from moderate-to-severe symptoms (MADRS ≥20) at weeks 0 and 4.

**Hypothesis 2:** The model-predicted MADRS score will be positively associated with the clinician-administered MADRS total score at weeks 0 and 4.

**Hypothesis 3:** Speech-derived outputs will be associated with individual symptom severity and treatment response.

**Hypothesis 4**: Combining speech features with passive behavioral data will improve predictive performance compared to speech features alone.

# 3. General Description of Study Design

**General Description of Study Design**.

The study will consist of a pragmatic, pre–post observational study to evaluate whether a brief smartphone-based speech task administered through the Callyope app is sensitive to symptom change among adults with MDD receiving ketamine or esketamine treatment at Massachusetts General Hospital (MGH) or participating in studies at the Depression Clinical and Research Program at MGH. In addition to the app-based self-report and speech collection, a clinician-administered MADRS interview will be conducted at baseline (week 0) and at week 4 to provide the reference standard for depressive symptom severity against which the speech-derived measures are evaluated.

# 4. Subject Selection

**Participant Selection**

Participants will be recruited from two primary sources: 1) the MGH Ketamine Clinic or 2) Dr. Naoise Mac Giollabhui studies examining the effects of TNF inhibitor (infliximab) on depression being conducted at the MGH-DCRP (2023P002988 via 2025P003067). The Clinical Research Coordinator (CRC) will identify potential participants at the ketamine clinic by using information collected during the standard intake consultation with a psychiatrist, which includes a comprehensive psychiatric and medical history. These intake records will serve as the primary source of information to determine initial eligibility. Potential patients will include both individuals initiating ketamine-based treatment and those receiving maintenance therapy. Recruitment at site 2 will be conducted by examining phone screens of participants interested in participating in I-CODE (Inflammation in Cognitive dysfunction in acute Depression 2023P002988).

Major Depressive Disorder is highly prevalent, and approximately one-third of patients experience treatment-resistant symptoms requiring interventions such as ketamine or esketamine. Participants will be recruited from the MGH Ketamine Clinic, which serves individuals actively receiving these treatments, and from Dr. Naoise Mac Giollabhui's study 2023P002988 via 2025P003067 within the Depression Clinical and Research Program. This recruitment approach targets a clinically relevant population consistent with the study objectives, and the planned sample size is sufficient to evaluate associations between speech features and symptom severity within this group.

Participants will be recruited from multiple sources, including both a clinical treatment setting and ongoing research studies within the MGH Depression Clinical and Research Program, which increases the likelihood of reaching individuals with varied backgrounds and treatment pathways. Recruitment will occur without restriction based on race, ethnicity, or socioeconomic status. While English proficiency is required due to the nature of speech-based assessments, all eligible individuals within the recruitment pool will be approached to support inclusive enrollment.

**Inclusion and exclusion criteria**

For inclusion in the clinical study, participants will be aged 18 years or above, meet criteria for MDD confirmed via chart review, and be able to provide informed consent and complete study procedures. Participants meeting criteria for any psychotic disorder (e.g., schizophrenia, delusional disorder) or substance use disorder mild, moderate, or severe will not be enrolled in the study. Due to the low risk of the studies (noninvasive remote monitoring and tablet-based questionnaires), participants will be allowed

to enroll concurrently in other research studies. There will be no prohibited medications; all concomitant medications will be documented. Participants with both iPhones and Androids will be able to participate.

Table 1. Inclusion and Exclusion Criteria

| |
|---|
| **Inclusion criteria:**<br>• Clinical diagnosis of MDD confirmed via chart review<br>• Aged 18 years or older<br>• Fluent in reading and speaking English<br>• Capable of providing informed consent and complying with the protocol in the estimation of the investigator.<br>• Capable of operating the Callyope app<br>• Capable of completing speech tasks<br>• Owning a smartphone |
| **Exclusion criteria:**<br>• Meeting DSM-5 criteria for any psychotic disorder (e.g., schizophrenia or delusional disorder) or a substance use disorder (excluding mild alcohol and cannabis use disorder, which will be permitted). |

# 5. Subject Enrollment

Participants will be recruited from two primary sources: 1) the MGH Ketamine Clinic or 2) Dr. Naoise Mac Giollabhui study (2023P002988) examining the effects of TNF inhibitor (infliximab) on depression being conducted at the MGH-DCRP. The Clinical Research Coordinator (CRC) will identify potential participants at the ketamine clinic by using information collected during the standard intake consultation with a psychiatrist, which includes a comprehensive psychiatric and medical history. These intake records will serve as the primary source of information to determine initial eligibility. Potential patients will include both individuals initiating ketamine-based treatment and those receiving maintenance therapy. If the participant is interested, the study staff will schedule a consent visit with Dr. Pedrelli. Recruitment at site 2 will be conducted by examining phone screens of participants interested in participating in I-CODE (Inflammation in Cognitive dysfunction in acute Depression 2023P002988). To ensure proper consent is obtained prior to contacting participants in ICODE, ICODE participants are asked to enroll in a future contact protocol (2025P003067). Participants from ICODE who have agreed to the future contact consent will be contacted. Study staff will call and/or email eligible patients identified at both sites to describe the study in detail using an IRB approved telephone script and/or email template and assess potential interest. Patients who express interest will be scheduled for a remote visit with Dr. Pedrelli, who will address any questions and complete the informed consent process via REDCap.

Participation in this study is entirely voluntary. Individuals will be informed that their decision to participate or not participate will not affect their clinical care or relationship with their providers. Recruitment and consent will be conducted by study staff, and care will be taken to separate clinical care from research participation. The recruitment sites for this study have limited overlap with the Principal

Investigator's direct clinical clientele, further minimizing the potential for perceived pressure to participate.

# 6. STUDY PROCEDURES

**Methods and Procedures**.

Once consent is obtained, the CRC will assist participants in downloading the *Callyope* app onto their mobile phone. Through this app, participants will complete the Quick Inventory of Depressive Symptomatology (QIDS), the Generalized Anxiety Disorder Scale (GAD-7), the Athens Insomnia Scale (AIS), and a standardized speech analysis. The speech task consists of questions about how the patient is feeling. Participants will respond by pressing a button on the app to record their answer. The voice recording is then saved on an encrypted cloud. There will also be the possibility to have an AI agent, specific to the Callyope app, to collect the speech tasks.

These self-report questionnaires will be administered at baseline and repeated two and four weeks after baseline (see Table 2 for study flow). Participants will undergo the informed consent process prior to any study activity.

The Montgomery–Åsberg Depression Rating Scale (MADRS) will additionally be administered at baseline (week 0) and at week 4 by Dr. Pedrelli or by a trained study clinician under her supervision.

Several studies on the use of smartphone data have shown that the digital phenotype can be used for clinical predictions, and it seems relevant to collect information about the patient's lifestyle habits. Therefore, if the patient agrees and his/her smartphone is compatible, passive data will be collected daily. The metrics include step count, distance, calories burned, and sleep duration.

At baseline, demographic information, treatment type (for example, intravenous ketamine versus intranasal esketamine), psychiatric diagnoses, and current medication regimen will also be collected through chart review.

**Assessments**

All assessments will occur remotely using MGB-approved platforms or at the MGH DCRP.

**Data Collection**

Data collected by the MGH team will be stored on REDCap, accessible only to MGH staff working on the study. Data will be sent to Callyope staff using MGB-approved methodology. Speech task data and self-report questionnaires collected using the Callyope app will be pseudonymized and stored in two HDS-certified (Health Data Hosting) cloud platforms– AWS (Amazon Web Services) and GCP (Google Cloud Platform) – with all servers located in France. For AWS, the certificate number is FR071635 - Version 3 and the file number is 12164167. For GCP, the certificate number is 2022-013, issued by EY CertifyPoint. To ensure the highest level of security, data is encrypted using AES (Advanced Encryption Standard) technology with 256-bit keys (AES-256). The pseudonymized database will be shared once a month with Dr. Pedrelli's team.

**Return of Research Results**

Participants will not receive individual research results from speech analyses, as these measures are investigational and not validated for clinical use. Clinically relevant findings, such as endorsement of suicidal ideation on study assessments, will be addressed according to established safety protocols, including referral for appropriate clinical evaluation. Aggregate study results may be shared through publications or presentations.

**Compensation**

Participants will be compensated for their involvement in the study. They will receive $10 for completing the consent visit and technology orientation, $10 for each of the three speech assessments, $10 for completing the questionnaires at baseline and at each follow-up, and $10 for each administration of the MADRS. The maximum compensation per participant will therefore be $90, although the payment amounts may be adjusted during contracting without altering the scientific aims of the study. Participants will be paid a lump sum upon completion of the study or at the end of their participation, should they choose to withdraw or be terminated. The full breakdown is listed below.

Consent: $5

Technology Orientation: $5

Speech Assessment: $10 each, $30 total

Self-Report Measures (QIDS, GAD7, AIS): $10 each, $30 total

MADRS: $10 each, $20 total

Week 0: $40

Week 2: $20

Week 4: $30

**Termination Criteria**

Participants may be withdrawn from the study under the following conditions: (1) withdrawal of consent at any time; (2) inability to complete key study procedures, defined as failure to complete baseline assessments or repeated non-completion of follow-up assessments (e.g., missing scheduled assessments after two consecutive reminders); (3) development of clinically significant worsening symptoms, including active suicidal ideation requiring escalation to a higher level of care; or (4) occurrence of adverse events that, in the judgment of the study clinician, warrant discontinuation of participation. The study may also be terminated early by the investigators if unexpected safety concerns arise.

**Standard of Care Comparison**

This study is observational and does not alter standard clinical care. All participants will continue to receive treatment as determined by their clinical providers. The only research-specific procedures include completion of questionnaires, speech tasks, and optional passive data collection via a smartphone application. The alternative to participation is to receive standard clinical care without involvement in research procedures.

Table 2. Study Procedure

| | **Week 0** | **Week 2** | **Week 4** | **Daily** |
|---|---|---|---|---|
| Consent | X | | | |

| | | | | |
|---|---|---|---|---|
| Technology orientation | X | | | |
| Demographic information form | X | | | |
| MADRS | X | | X | |
| QIDS | X | X | X | |
| GAD-7 | X | X | X | |
| AIS | X | X | X | |
| Speech Task | X | X | X | |
| Passive Data | | | | X |

# 7. Risks and Discomforts

**Potential Risks**

This study involves minimal risk. Potential risks include minor inconvenience or frustration related to use of the smartphone application, emotional discomfort when completing questionnaires related to mood or anxiety, and potential risks to confidentiality associated with electronic data collection. There are no physical risks associated with study participation.

**Risk Reduction Procedures**
To minimize risks, participants may skip any questions they do not wish to answer and may withdraw at any time. Data collected will be limited to the minimum necessary to achieve the study objectives and will be stored on secure, password-protected, and encrypted systems in accordance with Mass General Brigham policies. Identifiable information will be de-identified prior to analysis whenever feasible. Participants endorsing concerning symptoms, including suicidal ideation, will be promptly referred for appropriate clinical evaluation.

The study is designed as a non-invasive, observational study using procedures that are commonly used in clinical and research settings. No experimental interventions are introduced, and study procedures do not interfere with standard clinical care. As such, risks are minimized and reasonable in relation to the anticipated knowledge gained.

Several study assessments, including MADRS, QIDS, GAD-7, and AIS, are commonly used clinical tools for evaluating depressive symptoms, anxiety, and insomnia. Their use in this study reflects standard assessment practices, with the primary difference being collection through a mobile application and at specified research time points.

# 8. Benefits

There is no direct clinical benefit to participants. However, participants may derive personal satisfaction from contributing to research that aims to improve understanding and monitoring of depression. Participants will also receive compensation as described in the protocol. More broadly, this study may contribute to improved understanding of how speech and behavioral data can be used to monitor

depression and predict treatment response. This may support the development of more objective and accessible tools for mental health assessment in the future.

# 9. Statistical Analysis

This study is a prospective external evaluation of a fully locked Callyope model. The model version, input-processing pipeline, speech quality-control rules, predicted MADRS score, class probabilities, and classification thresholds will be fixed before study outcomes are examined.

No model training, feature selection, threshold optimization, or recalibration will be performed using study data. Accordingly, the full evaluable dataset will be used for external validation without a training-test split.

The primary analysis population will include all participants with at least one valid speech assessment paired with a clinician-administered MADRS assessment at the same study time point. The primary analyses will include paired assessments collected at weeks 0 and 4.

Assessments for which the MADRS score or model output is unavailable will not be imputed for the primary analyses. The number and reasons for missing, invalid, or excluded assessments will be summarized. All confidence intervals will be two-sided 95% confidence intervals unless otherwise specified.

Because the primary analyses pool repeated assessments from the same participants, confidence intervals for the co-primary endpoints will account for within-participant correlation using a participant-level cluster bootstrap. Complete analytic details will be prespecified in a Statistical Analysis Plan finalized before study outcome data are examined.

### 9.1 Co-primary endpoint 1: categorical severity classification

Observed depressive symptom severity will be dichotomized as lower severity (MADRS <20) versus moderate-to-severe symptoms (MADRS ≥20).

Discrimination will be quantified using AUROC, sensitivity, specificity, positive predictive value, negative predictive value, balanced accuracy, and the confusion matrix will also be reported.

### 9.2 Co-primary endpoint 2: continuous MADRS severity estimation

The association between the model-predicted MADRS score and the clinician-administered MADRS total score will be quantified using the Pearson correlation coefficient. The primary estimate will use all valid paired observations from weeks 0 and 4. Spearman correlation will be reported as a sensitivity analysis.

### 9.3 Success criteria

For categorical severity classification, the null hypothesis is AUROC ≤0.60 versus the alternative AUROC >0.60. The anticipated AUROC used for study planning is 0.75.

For continuous MADRS severity estimation, the null hypothesis is a population Pearson correlation ≤0.50 versus the alternative correlation >0.50. The anticipated correlation used for study planning is 0.70, corresponding to a squared correlation of 0.49.

Each co-primary hypothesis will be evaluated at a one-sided alpha level of 0.025. Success requires the lower bound of the two-sided 95% confidence interval to exceed the applicable null value.

Both co-primary endpoints must meet their respective success criteria for the primary study objective to be met. Because this is an intersection-union testing procedure, no additional multiplicity adjustment is required between the co-primary endpoints.

The anticipated values of AUROC = 0.75 and r = 0.70 are planning targets rather than null values. The study will not be described as demonstrating performance at or above these targets unless the observed estimates and confidence intervals support that conclusion.

The robustness of the co-primary results will be evaluated through:

- Separate analyses of week 0 and week 4 observations;
- An analysis restricted to participants contributing both week 0 and week 4 assessments;
- A descriptive comparison of participants with and without an evaluable week 4 assessment.

### 9.4 Agreement, calibration, and prediction error

Correlation alone does not establish agreement between predicted and clinician-administered MADRS scores. Supportive analyses will therefore include:

- Predictive $R^2$, calculated from squared prediction error relative to prediction using the observed sample mean;
- Mean absolute error and root mean squared error;
- Mean prediction error;
- A plot of predicted versus observed MADRS scores, including the identity line and a smoothed calibration curve.

The squared Pearson correlation may also be reported descriptively but will not be interpreted as equivalent to predictive $R^2$ or absolute agreement. No formal pass/fail threshold is specified for these supportive measures.

### 9.5 Secondary and exploratory analysis

Three-category MADRS severity will be defined as lower severity (MADRS <20), moderate severity (MADRS 20-34), and severe symptoms (MADRS ≥35).

Three-category classification will be summarized using the confusion matrix, balanced accuracy, and quadratic weighted kappa. One-versus-rest AUROCs will be reported using the relevant locked class probabilities. A prespecified secondary analysis will evaluate severe symptoms, defined as MADRS ≥35 versus MADRS <35.

Other secondary analyses will include:

- Performance at weeks 0 and 4 separately;
- Correlation between changes in predicted and clinician-administered MADRS scores from week 0 to week 4;
- Prediction of treatment response, defined according to the criteria specified in Section 2;

- Associations with QIDS, GAD-7, AIS, and prespecified symptom-level outcomes.

Analyses incorporating passive sleep or activity data will be exploratory. Secondary and exploratory analyses will be reported with confidence intervals.

**9.6 Sample Size justification**

Up to 100 participants will be enrolled. Assuming that all enrolled participants contribute a baseline assessment and that 50% contribute a week 4 assessment, approximately 150 paired visit-level observations are expected.

To account for repeated observations, the sample-size assessment applies a design effect based on the expected distribution of one or two assessments per participant (mean 1.50) and an assumed within-participant correlation ranging from 0.30 to 0.70. This results in an approximate effective sample size of 111 to 130 independent observations. For the continuous co-primary endpoint, the planning assumptions are:

- Minimum acceptable population correlation: 0.50;
- Anticipated population correlation: 0.70;
- One-sided alpha level: 0.025; and
- Effective sample size: 111 to 130 independent observations.

Under these assumptions, the effective sample size provides approximately 91% to 96% marginal power for the continuous co-primary endpoint.

For the categorical co-primary endpoint, the planning assumptions are:

The minimum acceptable AUROC is 0.60, the anticipated AUROC is 0.75, the expected proportion of pooled assessments with MADRS ≥20 is 30% to 70%, and the one-sided alpha level is 0.025.

Under these assumptions, approximate marginal power for the categorical co-primary endpoint ranges from 72% under the most conservative assumptions to 92% under more favorable assumptions. At an anticipated AUROC of 0.75, the expected half-width of the 95% confidence interval is approximately 0.08 to 0.11, depending on effective sample size and category balance. These are marginal powers; because both co-primary endpoints must meet their success criteria, overall study power will be lower and will depend on the correlation between the two endpoint statistics.

Enrollment of up to 100 participants represents the maximum feasible sample size and prioritizes clinically meaningful performance criteria over weaker null hypotheses such as correlation equal to zero or AUROC equal to 0.50.

If the number of evaluable assessments or the balance between MADRS severity categories is less favorable than anticipated, all analyses will be conducted as prespecified, with correspondingly wider confidence intervals and reduced power.

# Monitoring and Quality Assurance

**Monitoring and Quality Assurance Plan**

**Adverse Event Monitoring and Reporting**
All adverse events (AEs), including any reports of suicidal ideation or behavior, will be assessed and documented at each study interaction, including clinical visits and self-report assessments. AEs will be reviewed by the PI and documented in accordance with institutional and IRB requirements. Serious adverse events (SAEs) will be reported promptly per IRB guidelines.

**Safety Monitoring Procedures**
This study will be PI-monitored, with ongoing oversight by the PI. Formal study safety meetings will occur on a biweekly basis and will include the PI and study coordinators. During these meetings, the team will review participant status, adverse events, and overall study conduct.

Although the use of the study's mobile application is not expected to increase suicide risk, participants will be closely monitored throughout the study:

1. If a participant endorses clinically significant suicidal ideation at the study zoom meetings (as determined by the clinician), the clinician will thoroughly assess the participant to determine their suicide risk. If necessary, care will be escalated to a higher level (e.g., emergency services, inpatient hospitalization).
2. Participants endorsing suicidal ideation via mobile assessments (e.g., score >1 on QIDS item 12) will receive an automated notice to contact their clinician, National Suicide Hotlines (available 24 hours/day, 7 days/week), or go to the nearest emergency room.

Participants who develop active suicidal ideation will be referred for urgent clinical assessment at a facility close to their location. All participants will be provided with the PI's emergency contact information during the consent process and may contact the study team at any time if concerning suicidal ideation arises.

**Outcomes Monitoring**
Study outcomes, including clinical assessments and application-derived measures, will be reviewed on an ongoing basis by the study team to ensure data quality and participant safety.

**Internal Monitoring of Source Data and Protocol Adherence**
The Principal Investigator (PI) and designated research assistants will oversee internal monitoring of source data, protocol adherence, and study recordkeeping. All participant forms and study data will be reviewed on an ongoing basis to ensure accuracy, completeness, and compliance with the study protocol. Data entry and documentation procedures will be periodically reviewed to identify and resolve discrepancies in a timely manner.

Staff at Callyope will be responsible for monitoring and maintaining the integrity and quality of application-derived metrics, including routine checks for data consistency and completeness.

# 10. Data and Research Material Sharing

## A) Sending Data/Materials to Research Collaborators outside Mass General Brigham

The MGH study team will share data with Callyope for the purposes of study-related analysis. Data shared will include de-identified self-report questionnaire data and limited clinical/demographic variables

necessary for analysis. In addition, limited Protected Health Information (PHI) related only to study visit dates may be shared with Callyope to support the timing of app-based activities and assessments tied to specific study visits; no other PHI will be disclosed.

All data will be de-identified prior to transfer. Participants will be assigned a unique study ID, and no direct identifiers (e.g., name, contact information) will be included in the shared dataset. A master linking log connecting identifiers to study IDs will be maintained securely at MGH and will not be shared externally. As such, the recipient will not be able to re-identify participants.

Data will be transferred using Mass General Brigham (MGB)-approved secure transfer methods. Data collected by the MGH team will be stored in REDCap, accessible only to authorized MGH study staff.

Participants will provide informed consent for data sharing with external collaborators. Participants may withdraw from the study at any time by notifying the study team; no new data will be collected after withdrawal. Data collected prior to withdrawal may continue to be used consistent with the consent form.

No clinical data will be stored at external sites, and shared data will not be used to re-identify participants. De-identified data may be stored by the collaborator for study-related analyses and future research use as permitted by the consent form and DUA.

### B) Receiving Data/Materials from Research Collaborators outside Mass General Brigham

The MGH study team will receive data from Callyope collected via the study mobile application. These data will include speech task recordings and self-report questionnaire responses collected as part of study procedures.

Data received by MGH will be de-identified and labeled using study IDs only. Callyope will not provide any direct identifiers or a linking file; therefore, MGH will not be able to re-identify participants using externally received data alone. Re-identification, if necessary, will only be possible through the secure linking log maintained at MGH.

Data collected through the application will be stored on secure AWS cloud servers located in the United States and shared with the MGH study team on a monthly basis using MGB-approved secure transfer methods. Upon receipt, data will be stored in secure, access-restricted MGH systems.

All data transfers will occur under a fully executed Data Use Agreement (DUA), and data use will be limited to the purposes described in the protocol and consent form. No identifiable data will be received or stored outside of MGH systems.

Participants will consent to this data collection and sharing process. As with all study data, participants may withdraw at any time; no additional data will be collected following withdrawal, and previously collected data may be retained and used in de-identified form in accordance with the consent.

## 11. Privacy and Confidentiality

- ☒ Study procedures will be conducted in a private setting.
- ☒ Only data and/or specimens necessary for the conduct of the study will be collected.
- ☒ Data collected (paper and/or electronic) will be maintained in a secure location with appropriate protections such as password protection, encryption, physical security measures (locked files/areas)
- ☒ Specimens collected will be maintained in a secure location with appropriate protections (e.g. locked storage spaces, laboratory areas)
- ☒ Data and specimens will only be shared with individuals who are members of the IRB-approved research team or approved for sharing as described in this IRB protocol.
- ☒ Data and/or specimens requiring transportation from one location or electronic space to another will be transported only in a secure manner (e.g. encrypted files, password protection, using chain-of-custody procedures, etc.)
- ☒ All electronic communication with participants will comply with Mass General Brigham secure communication policies.
- ☒ Identifiers will be coded or removed as soon as feasible and access to files linking identifiers with coded data or specimens will be limited to the minimal necessary members of the research team required to conduct the research.
- ☒ All staff are trained on and will follow the Mass General Brigham policies and procedures for maintaining appropriate confidentiality of research data and specimens.
- ☒ The PI will ensure that all staff implement and follow any Research Information Service Office (RISO) requirements for this research.
- ☒ Additional privacy and/or confidentiality protections

NA

# 13. References

Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. From joyous to clinically depressed: mood detection using spontaneous speech. 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2012. 2012:141-146.

Ando, Angelika, et al. "SLAP: Learning Speaker and Health-Related Representations from Natural Language Supervision." (2025).

Carney, Colleen E., et al. "The relation between insomnia symptoms, mood, and rumination about insomnia symptoms." Journal of Clinical Sleep Medicine 9.6 (2013): 567-575.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. Psychological Science, 15(10), 687–693.

Fava, Maurizio, et al. "Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR* D report." American Journal of Psychiatry 165.3 (2008): 342-351.

Gaynes BN, Warden D, Trivedi MH, Wisniewski SR, Fava M, Rush AJ. What did STAR*D teach us? Results from a large-scale, practical, clinical trial for patients with depression. Psychiatr Serv. 2009;60(11):1439-1445. doi:10.1176/ps.2009.60.11.1439.

Guo T, Xiang YT, Xiao L, et al. Measurement-based care versus standard care for major depression: a randomized controlled trial with blind raters. Am J Psychiatry. 2015;172(10):1004-1013. doi:10.1176/appi.ajp.2015.14050652.

Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017;318(13):1215-1216. doi:10.1001/jama.2017.11295.

Mac Giollabhui N, Madison AM, Lydston M, Miller AH, Quang EL, Liu RT. Systematic Review Examining the Efficacy of Anti-Inflammatory Treatment on anhedonia and depressive symptom severity in individuals with major depressive disorder and an inflammatory phenotype, Brain, Behavior, and Immunity, Volume 122, Supplement, 2024, Pages 12-13. https://doi.org/10.1016/j.bbi.2024.12.058.

Riad R, et al. Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: algorithm development and validation study. J Med Internet Res. 2024;26:e58572. doi:10.2196/58572.

Riley, Richard D., et al. "Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study." BMJ 384 (2024).

Roquefort, F., Ducorroy, A., Riad, R. (2025). In-context learning capabilities of Large Language Models to detect suicide risk among adolescents from speech transcripts. Proc. Interspeech 2025, 414-418. doi:10.21437/Interspeech.2025-2083.

Taguchi T, Tachikawa H, Nemoto K, et al. Major depressive disorder discrimination using vocal acoustic features. J Affect Disord. 2018;225:214-220. doi:10.1016/j.jad.2017.08.038.

Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data-driven smartphone research. JMIR Ment Health. 2016;3(2):e16. doi:10.2196/mental.5165.

Treadway MT, Etuk SM, Cooper JA, et al. A randomized proof-of-mechanism trial of TNF antagonism for motivational deficits and related corticostriatal circuitry in depressed patients with high inflammation. Molecular Psychiatry. 2025;30(4):1407-1417. doi:10.1038/s41380-024-02751-x.

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(1):28-40. doi:10.1176/appi.ajp.163.1.28.

Wilkinson ST, Ballard ED, Bloch MH, et al. The effect of a single dose of intravenous ketamine on suicidal ideation: a systematic review and individual participant data meta-analysis. Am J Psychiatry. 2018;175(2):150-158. doi:10.1176/appi.ajp.2017.17040472.

Zarate CA Jr, Singh JB, Carlson PJ, et al. A randomized trial of an N-methyl-D-aspartate antagonist in treatment-resistant major depression. Arch Gen Psychiatry. 2006;63(8):856-864. doi:10.1001/archpsyc.63.8.856.

# APPENDIX A

## Data Monitoring Committee / Data and Safety Monitoring Board Appendix

- *To be completed for studies monitored by Data Monitoring Committee (DMC) or Data and Safety Monitoring Board (DSMB) if a full DMC/DSMB charter is not available at the time of initial IRB review.*
- *DMC/DSMB Charter and/or Roster can be submitted to the IRB later via Amendment, though these are not required* if this Appendix has been completed and approved by the IRB.

A Data Monitoring Committee (DMC) or Data and Safety Monitoring Board (DSMB) will be convened for safety monitoring of this research study. The following characteristics describe the DMC/DSMB convened for this study (Check all that apply):

☐ The DMC/DSMB is independent from the study team and study sponsor.

☐ A process has been implemented to ensure absence of conflicts of interest by DMC/DSMB members.

☐ The DMC/DSMB has the authority to intervene on study progress in the event of safety concerns, e.g., to suspend or terminate a study if new safety concerns have been identified or need to be investigated.

☐ Describe number and types of (i.e., qualifications of) members:
Click or tap here to enter text.

☐ Describe planned frequency of meetings:
Click or tap here to enter text.

☐ DMC/DSMB reports with no findings (i.e., “continue without modifications”) will be submitted to the IRB at the time of Continuing Review.

☐ DMC/DSMB reports with findings/modifications required will be submitted promptly (within 5 business days/7 calendar days of becoming aware) to the IRB as an Other Event.