Title: Study of a PST-Trained Voice-Enabled Artificial Intelligence Counselor (SPEAC) for Adults With Emotional Distress (Phase 1)

NCT number: NCT04524104

Study of a PST-Trained Voice-Enabled Artificial Intelligence Counselor (SPEAC) for Adults with Emotional Distress (Phase 1)

Principal Investigator:

Jun Ma, MD, PhD
Beth and George Vitoux Professor of Medicine
Director of Vitoux Program on Aging and Prevention
Department of Medicine
University of Illinois at Chicago
Phone: (312) 413-9830
Email: *maj2015@uic.edu*

Multiple Principal Investigator:

Olusola Ajilore, MD, PhD, Associate Professor, Associate Director, Residency Training and Education, Director, Psychiatry Residency Neuroscience Research Tracks, Department of Psychiatry, University of Illinois at Chicago

Co-Investigators:

Ben S. Gerber, MD, PhD, Professor of Medicine, Associate Chief, Division of Academic Internal Medicine and Geriatrics; College of Medicine; Institute for Health Research and Policy, University of Illinois at Chicago

Phillip Yu, PhD, Distinguished Professor, Wexler Chair in Information Technology, Department of Computer Science, University of Illinois at Chicago

Joshua Smyth, PhD, Professor of Biobehavioral Health and Medicine, Department of Biobehavioral Health (BBH) and Stress, Health, and Daily Experiences (SHADE) Lab, The Pennsylvania State University

Jill Johnson, PhD, Assistant Research Professor, Department of Biobehavioral Health (BBH) and Stress, Health, and Daily Experiences (SHADE) Lab, The Pennsylvania State University

Thomas Kannampallil, PhD, Assistant Professor, Department of Anesthesiology and the Institute for Informatics, Washington University School of Medicine in St Louis

Lan Xiao, PhD, Research Biostatistician, Department of Epidemiology and Population Health, Stanford University.

Study Location(s): University of Illinois at Chicago, UI Health
Washington University in St. Louis (WashU)
Pennsylvania State University (PSU)
Stanford University

Sponsor: National Institute of Mental Health

Version: **12**

Date: **8/16/2021**

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

COI	Conflict of Interest
DHHS	Department of Health and Human Services
DMC	Data Monitoring Committee
DSMB	Data and Safety Monitoring Board
DSMP	Data and Safety Monitoring Plan
FERPA	Family Educational Rights and Privacy Act
FDA	Food and Drug Administration
GCP	Good Clinical Practice
HIPAA	Health Insurance Portability and Accountability Act
IBC	Institutional Biosafety Committee
ICD	Informed Consent Document
ICH	International Conference of Harmonization
IDE	Investigational Device Exemption
IDS	Investigational Drug Service
IND	Investigational New Drug
IRB	Institutional Review Board
LAR	Legally Authorized Representative
OHRP	Office of Human Research Protections
OPRS	Office for the Protection of Research Subjects
PHI	Protected Health Information
PI	Principal Investigator
PST	Problem Solving Therapy
PPRA	Protection of Pupil Rights Amendment
PRO	Patient-reported Outcome
QA/QI	Quality Assurance/Quality Improvement
SAE	Serious Adverse Event
SOP	Standard Operating Procedure

SUMMARY OF EDITS

V12 on 2021.8.16

- 6.0 Study design and procedures Aim 2 Pilot RCT: Clarification that the post-intervention interviews are conducted via Zoom, and a certificate of completion is provided to intervention particpants.

V11 on 2021.6.15

- Appendix 4: Event Windows. Rescreening requirement modified from 28 to 90 days in order to attend Visit 1.
- 11.0 Data and Safety Monitoring: In order to access the electronic medical record, as needed, for AE adjudication, consenting participants are asked to provide and/or confirm the minimum necessary identifiers including first and last name, date of birth, mailing address, and whether they have received care at UI Health.

V10 on 2021.4.9

- Aim 1 formative research sections: Updated to remove the focus group stage.
- Appendix 2: Schedule of Measures – RCT: updated to include new instrument: UCLA 3-item loneliness survey.

V9 on 2021.3.25

- 5.0 Subject Enrollment Aim 2 Pilot RCT: Recruitment plan was amended to include an option to receive resources for emotional health services, offered to participants, who score 10 or above on GAD-7 and/or PHQ-9 surveys during initial eligibility screening questionnaire.
- Appendix 2: Schedule of Measures – RCT: updated to include new instruments; COVID-19 impact survey, Duke University Religion Index73, COVID impact on social functioning survey and Social Network Survey.

V8 on 2021.3.4

- Appendix 1: Updated Lumen Architecture. The architectural design of the Lumen virtual coach is simplified to leverage the existing Amazon Alexa platform and the UIC REDCap system to deliver the intervention. This affords the benefit of data storage and management within the UIC CCTS HIPAA compliant REDCap database, and on the AWS cloud which is covered by a BAA already in place between UIC/UI Health and Amazon.

V7 on 2021.2.15

- 4.0 Eligibility Aim 2 Pilot RCT: Exclusion criteria are updated to ensure participants will have, and are willing to use, the technology necessary to utilize the Lumen intervention; and based on data from previous trials, we have updated the body weight limit in order to comfortably fit participants into the MRI scanner.
- 5.0 Subject Enrollment Aim 2 Pilot RCT: Recruitment plan was amended to include use of UIC listservs to email recruitment invitations.
- 6.0 Study design and procedures Aim 2 Pilot RCT: Addition of the Lumen Orientation to improve participant engagement and the intervention experience. Assessment visit (Visits 1 and 2) duration and procedures are updated to reflect moving of study iPad distribution to the Lumen Orientation visits, and to account for time involved in training participants for the Nightly Mood Check-ins (i.e., the ecological daily assessments.)
- 7.0 Expected Risks/Benefits Aim 2 Pilot RCT:

    - Self-harm protection protocol is submitted. It reflects the Lumen intervention design capabilities, and specifies the self-harm alert messages presented to participants when the PHQ-9 is completed with and without staff present.
    - Clarification that participants may explicitly choose whether to authorize data sharing for NIMH Data Archived (NDA) during informed consent process.
    - Data collection and storage is updated per new Lumen design.
    - Updated protections against accidental recording per new Lumen design.
- 8.0 Data collection and management procedures Aim 2 Pilot RCT: Data collection and storage is updated per new Lumen design.
- 11.0 Safety monitoring Aim 2 Pilot RCT: Updated to remove involvement of a study internist as back-up for the study psychiatrist as this is deemed unnecessary given the low enrollment target for the pilot trial.
- Appendix 2. Schedule of Measures – RCT: Replaced the self-reported measure of functioning instrument with the Work productivity and activity impairment questionnaire (WPAI).
- Appendix 4. Study Event Windows Aim 2 Pilot RCT: Added to clarify how the addition of the Lumen Orientation necessitated adjustment of the event window for the final assessment from 14-weeks to 16-weeks post-randomization.

V6 on 2021.1.11

- Study Locations: Added Pennsylvania State University and Washington University in St. Louis as official non-UIC study sites.

V5 on 2020.11.6

- Study Locations: Updated status as 'Pending' for Pennsylvania State University and Washington University in St. Louis as official non-UIC study sites.
- 5.0 Subject Enrollment Aim 1 Formative User Study: Amended recruitment plan to remove active recruitment calling.

V4 on 2020.10.28

- 4.0 Eligibility Aim 1 Formative User Study: Updated eligibility criteria.
- 5.0 Subject Enrollment Aim 1 and Aim 2: Clarified the eConsent procedures.
- 6.0 Study design and procedures Aim 1 and Aim 2: Removed references to verbal consent to describe eConsent.
- 12.0 Regulatory Requirements: Clarified the Aim 1 and Aim 2 consent processes.

V3 on 2020.10.16

- 4.0 Eligibility Aim 1 Formative User Study: Updated enrollment target to 30.
- 5.0 Subject Enrollment Aim 1 Formative User Study: Amended screening plan to include access to technology necessary to conduct remote interviews.
- 6.0 Study design and procedures Aim 1 Formative User Study: Provided details on user interview sessions.
- 8.0 Data Collection and Management Aim 1 Formative User Study: Modified audio recording procedures.
- 12.0 Regulatory Requirements: Clarified the Aim 1 and Aim 2 consent processes.

V2 on 2020.8.3

- 4.0 Eligibility: Clarified participants are not withdrawn post-randomization if they begin pharmacotherapy or psychotherapy during the study.

1.0 Project Summary/Abstract

Depression and anxiety are the leading causes of disability and lost productivity, and are often underdiagnosed and undertreated owing to access, cost, and stigma barriers. Novel and scalable psychotherapies are urgently needed. Advances in artificial intelligence (AI) offer a transformative opportunity to develop intelligent voice assistants as virtual health agents accessible on personal devices. Meanwhile, major advances in human neuroscience have fueled a paradigm shift to study brain mechanisms underlying behavioral health interventions. Leveraging emerging science in these transdisciplinary areas, this project aims to develop and rigorously test a novel voice-enabled, AI virtual agent named Lumen, trained on Problem Solving Therapy (PST), for patients with moderate, untreated depressive and/or anxiety symptoms. The project will investigate the effect of Lumen on engagement of a priori neural targets—amygdala for emotional reactivity and dorsal lateral prefrontal cortex (DLPFC) for cognitive control—as putative mechanisms. The project has 2 phases: R61 and R33 (see Figure 1). Phase 1 is the R61 phase (years 1-2) and will be the sole focus of this protocol. Focusing on Lumen development and refinement as well as pilot testing, Phase 1 has 2 specific aims. Aim 1 on Lumen development and refinement will proceed in 2 stages: (1) scenario-based clinician evaluations, and (2) a formative user study (n=30 participants). Aim 2 is to pilot test Lumen in a 2-arm randomized clinical trial (pilot RCT), among 60 participants with moderate, untreated depression and/or anxiety who are randomized in a 2:1 ratio to receive PST with Lumen (n=40) on a secure study iPad or be on a waitlist (n=20). At weeks 0 and 16 participants will complete functional magnetic resonance imaging (fMRI) to assess neural target engagement as well as validated surveys of patient-reported outcomes (PROs) (e.g., depressive and anxiety symptoms, functioning, quality of life). In addition, participants will complete ecological daily assessments of mood, stress, appraisal and coping for 7 days every 2 weeks during the 16-week follow-up period. The Phase 1 milestones are (1) establishing the functionality, usability, and treatment fidelity of Lumen; and (2) demonstrating feasibility, acceptability, and neural target engagement. Achieving these milestones will provide the basis for the future R33 phase (Phase 2) focused on examining target engagement and PROs in a larger 3-arm RCT by comparing Lumen with a waitlist control arm and an in-person PST arm. This protocol only focuses on the R61 phase.

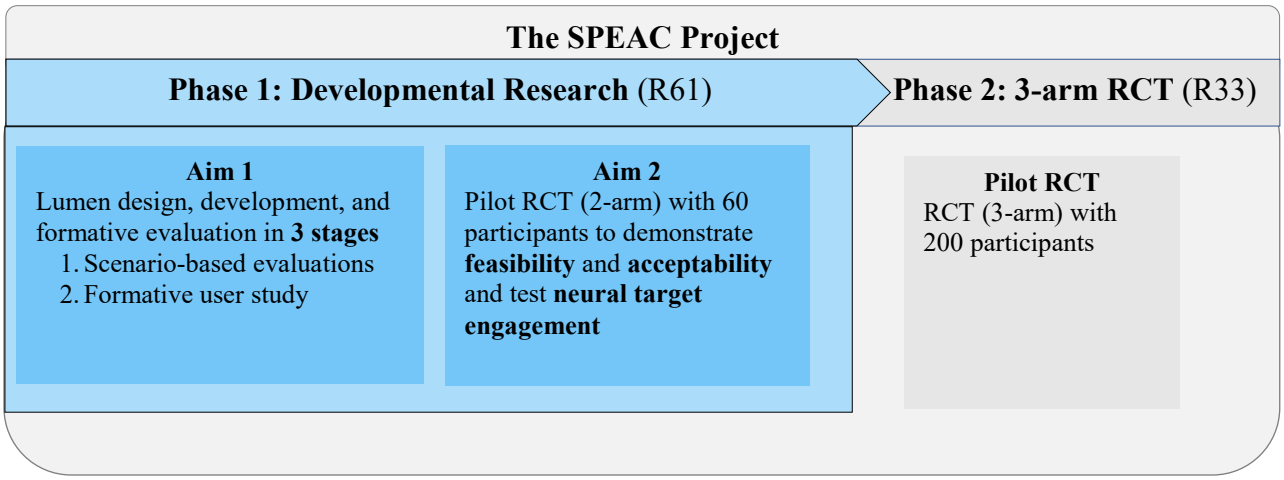

Figure 1. SPEAC Project Overview

2.0 Background/Scientific Rationale

Depression and anxiety cost an estimated $201 billion annually in the United States (US).[1,2] Approximately 16 million US adults (6.7% of the population) experience major depression and 40 million (18.1%) experience anxiety disorders,[3,4] and these disorders are often comorbid.[5] Moreover, subclinical symptomology not meeting the diagnostic criteria for major depression or anxiety is similarly common and is also associated with high disability and disease burden.[6-8] These conditions often go undiagnosed and untreated. Alarmingly, half of Americans with depression and 63% of those with anxiety do not receive any treatment, with the lowest rates among racial/ethnic minorities and persons of low socioeconomic status.[5,9-11] People with mental illness often prefer psychotherapy to medication.[12-14] Yet, the reach and adoption of proven psychotherapies in mental health or general medical settings are limited, owing to barriers such as low reimbursement, provider shortage, patients' lack of time and transportation, and stigma.[15-18] As such, there is a critically unmet need for empirically validated psychotherapies that have low cost, avoid stigma, and can be scaled to help address public health and health equity.

Technology-based interventions have quickly grown as a viable option for treatment delivery to address cost, access, and stigma barriers for traditional mental health services. The general public is highly receptive (76% reporting interest) to mental health monitoring and counseling using internet and mobile technologies,[19] and individuals with mental health concerns often prefer to seek help online rather than in person.[20] Studies have tested a broad range of technologies for delivering bona fide psychotherapies.[21,22] The strength of the evidence on early-generation technology interventions led to depression and anxiety treatment guidelines recommending computerized cognitive behavioral therapies using web-based or interactive voice response systems.[23] Digital mental health interventions have since been evolving with technological advances. Several meta-analyses have documented growing evidence on the utility of mobile interventions for the management of depression and anxiety.[24,25] One latest area of prolific technological growth in AI is around voice-based personal assistants, which are now nearly ubiquitous in personal home or mobile devices—with recent reports[26,27] describing that worldwide sales of Alexa devices crossing 100 million and that of Google Home devices crossing 50 million.

With the advances in AI technology, creating persuasive systems that mimic human-like behaviors has become a realistic endeavor.[28,29] The acceptability of AI-based systems in pragmatic applications is bolstered by studies that have shown that people are more willing to disclose personal information to a virtual agent than when they believe it is human-operated.[30] Empirical and theoretical research in the field of human-computer interface has shown that people anthropomorphize computers and complex technology[31,32] and form social attitudes and behaviors and emotional responses towards them. Experiments showed conversational partners whether perceived to be text-based chatbots or humans were equally effective at creating emotional, relational, and psychological benefits,[33] and establishing strong therapeutic alliances.[34] Importantly, older adults and people with low reading and/or computer literacy find virtual agents approachable and usable,[35,36] making these tools particularly relevant for addressing mental health disparities. Much of the development efforts to date have been centered around *text-based* conversational agents,[37] which have been shown to be feasible in a variety of settings,[38,39] and for depression and anxiety counseling.[40] Although promising, text-based intelligent conversational agents have limitations including a lack of personalized interaction and difficulty for use among older adults or people with low literacy.[37,41] In contrast, AI-powered personal assistants in dedicated voice devices (e.g., Amazon Echo, Google Home) and voice assistant applications (e.g., Alexa, Google Assistant, Siri) have advanced features to engage in human-like conversations. By utilizing automatic speech recognition, natural language processing, and deep learning algorithms, these voice assistants can be transformed from performing routine tasks (e.g., reporting current weather) to more sophisticated and context-specific health intervention tools. Research to develop and test such interventions is in its infancy. The focus of

this protocol is to develop a *voice-based* AI agent that has both pragmatic viability and therapeutic utility for PST.

The *Study of a PST-Trained Voice-Enabled Artificial Intelligence Counselor (**SPEAC**) for Adults with Emotional Distress* protocol is designed to enhance and rigorously test an already prototyped, voice-based AI agent, **Lumen**. In accordance to the Lumen design and training plan (Appendix 1), a mature product will be developed through iterative, user-centered design-evaluation cycles based on the current prototype. The project will test the effect of Lumen on engaging empirically supported neural targets as well as the relationship of change in neural targets to change in PROs using validated surveys and ecological assessments (in natural living). This project addresses the National Institute of Mental Health's strategic research priorities[42] and recommendations for developing novel information technologies in behavioral and social science research.[43] The public health impact lies in the potential to meaningfully improve the reach and impact of psychotherapy, mitigating access, cost, and stigma barriers for people with depression and/or anxiety who cannot or will not seek traditional mental health services or who desire more personalized, connected care. This may be particularly relevant to medically underserved populations (the target population in this study), thereby addressing health disparities.

3.0 Objectives/Aims

In Phase 1 of the SPEAC project, the specific aims are to: (1) establish the functionality, usability, and treatment fidelity of Lumen using iterative, user-centered design, development, and formative evaluation; and (2) demonstrate feasibility, acceptability, and target engagement in a 2-arm pilot RCT.

Aim 1: Formative Research
To achieve aim 1, we will conduct (1) scenario-based evaluations of fidelity by 2 PST master trainers; and (2) a formative user study with 30 eligible patients with depression and/or anxiety (defined below) who complete PST sessions with Lumen. The objective of the scenario-based evaluations is for 2 PST experts to assess the treatment fidelity of Lumen, and the completeness, consistency, and quality of the solution plan, and perceptions of Lumen. The objective of the formative user study is to test participant interaction with Lumen (e.g., what worked, what did not work, breakdowns in the interaction, understanding of developed solution implementation plan, perceptions regarding the viability of their plan, and general comments regarding the interaction) in order to evaluate usability,[44] user experience,[45] and therapeutic alliance[46-49] and complete Lumen development according to the Lumen design and training plan (Appendix 1).

Aim 2: Pilot RCT
To achieve aim 2, we will randomize 60 patients who score 10-19 on the Patient Health Questionnaire-9 (PHQ9) and/or 10-14 on the Generalized Anxiety Disorder-7 (GAD7), and randomize them in a 2:1 ratio to the Lumen treatment arm (n=40) or the waitlist control arm (n=20) for 16 weeks. The objectives of the pilot RCT are to determine feasibility and acceptability of Lumen as well as neural target engagement. Feasibility is assessed by being able to randomize 60 participants and retain ≥85% at week 16. Acceptability is assessed by having ≥95% Lumen participants complete≥1 PST session and ≥80% complete ≥4 sessions. To determine neural target engagement, the requirements are that both (a) the neural target for either nonconscious threat-related reactivity (amygdala) and/or cognitive control (DLPFC) improves meaningfully from week 0 to 16 for Lumen vs. waitlist control; and (b) change in either neural target significantly correlates with changes from weeks 0 to 16 in validated self-report measures corresponding with respective PST theory-based constructs for emotional reactivity or cognitive control.

4.0 Eligibility

Aim 1: Formative Research
The scenario-based evaluations are conducted by 2 independently contracted PST master trainers. The formative user study includes ENGAGE-2 (intervention and control) participants (n=30) screened for access to Lumen compatible Internet-enabled devices (iPad and/or smartphone.)

Aim 2: Pilot RCT
Racially/ethnically and socioeconomically diverse patients with moderate, untreated depression and/or anxiety symptoms who meet the eligibility criteria are included. Study participants are recruited from UI Health outpatient clinics. Eligibility criteria are detailed in Section 4.1 and 4.2. Participant eligibility is confirmed according to a multistep enrollment process detailed in Section 5. Trained research staff will assess eligibility per protocol and document findings in the REDCap (Research Electronic Data Capture) study database.

4.1 Inclusion Criteria
- Age: ≥ 18 years
- Emotional distress defined by elevated depressive (PHQ9 scores 10-19) and/or anxious symptoms (GAD7 scores 10-14)
- Willing and able to provide informed eConsent and HIPAA authorization

4.2 Exclusion Criteria
- Unable to speak, read, or understand English for informed consent
- Current pharmacotherapy or psychotherapy (individual or professionally led group therapy) for depression or anxiety (note: participants are not withdrawn post-randomization if they begin pharmacotherapy drugs or start psychotherapy during the study.)
- Suicidal ideation per PHQ9 with active plan
- Bipolar or psychotic disorder, or current psychiatric treatment
- Weight ≥325 pounds due to brain scanner constraints, MRI contraindications, traumatic brain injuries, and tumor or any other known structural abnormality in the brain (Aim 2 pilot RCT only)
- Severe medical condition (e.g., myocardial infarction or stroke or new cancer diagnosis in the past 6 months, end-stage organ failure, terminal illness) or residence in a long-term care facility
- Diagnosis of cancer (other than non-melanoma skin cancer) that is/was active or treated with radiation or chemotherapy within the past year
- Active alcohol or substance use disorder (including prescription drugs) based on the CAGE Questionnaire Adapted to Include Drugs (CAGE-AID)
- Cognitive impairment based on the Callahan 6-item screener
- Current or planned pregnancy or lactating (<6 months postpartum)
- Participation in other investigational treatment studies that would significantly affect participation in this study, raise safety concerns, and/or confound outcomes (participant may be asked to provide the informed consent of the other study for final decision on exclusion by a study psychiatrist)
- Family/household member of an already enrolled participant or of a study team member
- Plan to move out of the Chicago area during the study period
- Does not have reliable Wi-Fi Internet at home
- Unwillingness to user personal mobile device to receive study text messages
- Investigator discretion for clinical safety or protocol adherence reasons

4.3 Excluded or Vulnerable Populations
Children are excluded from the study as the target age group is 18 years and older. Psychotherapies of depression and anxiety and related risk protections (including protection of patient privacy and confidentiality) for persons under the age of 18 years differ from those for adults. Adults are the focus in this stage of research to develop a neural mechanism-validated, voice-based AI agent for PST. In

addition, participants unable to speak, read and understand English for informed consent are excluded from the study as the research instruments and intervention are administered in English only.

5.0 Subject Enrollment

The procedures for participant recruitment, screening, and teleorientation with eConsent are the same for participants in Aim 1 formative user study and Aim 2 RCT, except where noted below. Subject enrollment follows a multi-step process as follows:

Prescreening. Aim 2: Pilot RCT only.
The primary recruitment strategy uses UI Health's EHR database to identify patients meeting the basic prescreening criteria (e.g., age, absence of exclusionary medical or psychiatric comorbidities, etc.). Primary care providers (PCPs) have the option to review and identify patients who may be inappropriate for the study because of serious medical or psychiatric illness before recruitment invitations are sent. The secondary recruitment strategy uses in-clinic referrals or passive recruitment using advertising brochures and flyers at the UI Health outpatient clinics. Study brochures are distributed in clinic, and providers may refer patients during routine office visits and direct interested patients to the study website or telephone/text line for more information and screening. In addition, UIC listservs are used to distribute recruitment announcements to UIC employees. Based on previous experience using these recruitment strategies, it is anticipated that EHR-based recruitment will be the predominant source, accounting for >95% of enrolled participants. Reviews of patient EHR may be performed to screen for diagnostic exclusion criteria. This step is permissible under a waiver of informed consent. A waiver of informed consent and HIPAA preparatory to research are being requested for this process (medical record screening is minimal risk, will not affect patients' rights or welfare, and without it, subject recruitment would be impracticable).

Recruitment and Screening.
Recruitment invitations are sent to patients by email, if an email address is available in EHR, or by texting a link to the "open for recruitment" announcement on the study website. The recruitment email and online announcement describe the study in lay language and contain a secure web link to REDCap where patients may give online consent to screen for eligibility, if interested, or decline further contact if they choose to opt out. Patients may also opt out by directly replying to the invitation email or text or by calling the study recruitment phone line.
For the Formative User Study patients will receive recruitment invitations (via email and/or text) up to three (3) times and will not be contacted via the phone if no response is received. For the Pilot RCT, patients who do not self-screen or opt-out within 2 weeks of the recruitment invitation send date, a trained study coordinator calls to assess interest and conduct screening by phone. Based on previous experience, it is anticipated that the ratio of patients screened eligible to patients fully eligible and randomized to be around 5:2 (in screening new patients for the RCT). A participant tracking database will be developed in REDCap, which supports Microsoft Excel exports for analysis and reporting on enrollment and screen failures at each step.
During initial eligibility screening participants, who score 10 or above on GAD-7 and /or PHQ-9 surveys of recent symptoms of depression or anxiety, are offered the option to receive resources for emotional health services via email, if interested.

Teleorientation with eConsent. Screened eligible participants are invited to schedule a teleorientation session with a study coordinator via their choice of telephone or online video conference. Teleorientation begins with the coordinator providing the REDCap link to the online consent document and leading the informed consent discussion, as outlined in Section 12, including study procedures, risks and benefits, the voluntary nature of the research, and the pros and cons of being randomly assigned to Intervention and

waitlist control groups (pilot RCT only). Willing participants document their consent electronically using the secure REDCap eConsent framework and receive a copy of their signed consent in PDF form directly through REDCap. The study copy is automatically archived in the REDCap file repository. For the pilot RCT, eConsent is followed by the coordinator completing the final eligibility interview questions (see Appendix 2 screening measures).

Scheduling study procedures.
Aim 1: Formative Research
Formative User Study. After teleorientation, eligible participants are scheduled to attend the formative user study; a total of 30 participants will be scheduled.

Aim 2: Pilot RCT
Pilot RCT. After teleorientation, RCT participants are scheduled for a baseline assessment visit.. During the baseline visit, a trained study coordinator performs fMRI screening and data acquisition and collects baseline data per standardized measurement protocols. Females of child-bearing age who indicate any possibility that they may be pregnant must take a point-of-care urine pregnancy test. At the end of the baseline visit participants receive instructions to complete the initial ecological daily assessment for the next 7 days. Following successful completion of the baseline measures, fully eligible, consented participants are randomized to either Group A (Lumen Intervention) or Group B (waitlist control), at a 2:1 allocation.

6.0 Study Design and Procedures

Aim 1: Formative Research
Scenario-based evaluations do not include human subjects. These evaluations are carried out by 2 independent PST master trainers as contracted members of the research team, each of whom will assume the role of patients and interact with Lumen to complete 8 scenarios of PST encounters. These scripted scenarios will be based on previously recorded encounters with live counselors to broadly exploit various aspects of PST. After each scenario, they will provide feedback on quality and appropriateness of their interactions with Lumen.

The **formative user study** includes 30 participants with low, medium, or high digital health literacy based on the Digital Health Literacy Instrument.[51] After consent is obtained qualified study staff, each participant will complete a full PST session with Lumen in up to each of four 1-hour sessions. In addition, participants will download the paired application on their smartphones to test notifications, surveys, and ecological daily assessments. The formative research version of the Lumen PST session utilizes mock scores of recent psychological symptoms (PHQ-9 and GAD-7). Participants will only complete user-experience assessments and provide open-ended qualitative feedback on the interactive demo. Designated members of the study team (e.g. a researcher and a Lumen developer) will unobtrusively observe and listen to participant interactions with Lumen and take observation notes regarding critical incidents during the interaction (e.g., breakdowns in the conversation) and technical issues that were encountered. The modality of these unobtrusive observations (e.g. Zoom if conducted remotely, or from behind a one-way mirror if conducted in-person) will be determined by following the latest COVID-19 safety guidelines and precautions. Post-completion, participants perform a retrospective cognitive walkthrough to verbally describe their interaction with Lumen with the researcher. Cognitive walkthrough methods are commonly used for design and help in ascertaining the "cognitive fit" between the system and its users.[52] Specific emphasis during the walkthrough will be on participant interaction with Lumen: what worked (and what did not), breakdowns in the interaction, understanding of developed solution implementation plan, perceptions regarding the viability of their plan, and general comments regarding the interaction. The researcher and developer may prompt participants with additional questions based on their observation

notes. Following cognitive walkthroughs participants complete 3 user surveys: (1) NASA Task Load Index (NASA-TLX)[44] of user workload along with 6 dimensions of mental, physical, temporal, performance, effort and frustration, (2) User Experience Questionnaire-Short version (UEQ-S),[45] and (3) the Working Alliance Inventory adapted for digital interventions (WAI-Tech).[48,49] Survey data is collected via REDCap. Lastly, the 2 PST master trainers will evaluate audiorecorded Lumen PST sessions on: (1) treatment fidelity based on the 7-item PST Adherence and Competence Scale (PST-PAC),[53,54] (2) completeness, consistency, and quality of the solution plan, and (3) perceptions of Lumen.

Aim 2: Pilot RCT

The objective of the pilot RCT is to demonstrate (1) the feasibility of participant recruitment and retention, (2) participant acceptability of Lumen for PST, and (3) engagement of neural targets subserving emotional reactivity and cognitive control (i.e., target engagement).

Randomization and blinding. A designated staff person performs randomization using an online system that Dr. Ma and team published.[55] The system is a modern implementation of Pocock and Simon's minimization, a covariate-adaptive method[56] to achieve better-than-chance marginal balance between study arms across multiple key baseline characteristics. Minimization accommodates a greater number of balancing covariates than does stratified randomization.[57] The system's computational algorithm automatically adjusts the randomization probability based on the characteristics of all the previously randomized participants, thus minimizing the total covariate imbalance between arms after each new participant is randomized. In this study, randomization covariates include sex, age, race/ethnicity, education, digital health literacy,[51] PHQ9, and GAD7. Note that, aside from eligibility screening, PHQ9 and GAD7 are also process measures used to monitor treatment progress throughout PST sessions. Their inclusion among randomization covariates helps avoid accidental bias at baseline. The online system applies Efron's biased-coin method[58] to protect allocation concealment with the use of nonextreme randomization probabilities, and the staff person(s) using the system has/have no deep knowledge of the backend coding or execution and no ability to manipulate it. By design, treatment assignments are identifiable to participants, but blinding of outcome assessment, event adjudication, and data analysis will be enforced. To accomplish the 2:1 allocation, participants are randomized into 3 sets, and then 2 of these sets will be combined to form the Lumen group. The remaining set forms the waitlist control group. This method both protects blinding and preserves the allocation ratio at every allocation as per Kuznetsova and Tymofyeyev.[59] The same method is used in the ongoing ENGAGE-2 study.[60]

Assessments. All participants complete two study assessment visits, at baseline (week 0) and at 16 weeks, and self-complete ecological daily assessments for 7 days every 2 weeks. The fMRI scan visits are conducted at the 3T MR Research Program located at the UIMC Advanced Imaging Center, a BioRAFT registered facility, and SPEAC study staff work in compliance with all COVID-19 safety guidelines employed by facility management. When scheduling assessment visits, study coordinators instruct participants in the current COVID-19 safety precautions prior to and at the visits.

Assessments(120 minutes)

- fMRI assesses neural target engagement and treatment outcomes. fMRI is a standard technique for measuring and mapping brain activity that is noninvasive and safe. In this study, it is being used simply to gather neuro imaging data, and not to test the safety/efficacy of a device or software. Afterwards, the patient completes blood pressure, height, and weight measurements. (90 minutes)
- Self-report surveys of PST theory-based constructs of emotion (affect, worry) and cognition (problem solving, dysfunctional attitudes) as well as patient outcomes (e.g., depressive and anxiety symptoms, functioning, quality of life). Participants may complete these surveys during or outside their fMRI visits. (30 minutes).

Self-completed Ecological Daily Assessments (<3 minutes each day)
Using the Lumen companion application participants complete ecological end-of-day assessments of mood, stress, appraisal, and coping for 7 days every 2 weeks (over 16 weeks).

Intervention Orientation. Participants in the intervention arm attend a Lumen orientation visit (60 minutes) during which they will be given a Lumen intervention tutorial and receive a study iPad, configured to limit access to only the Lumen intervention enabled on the device. Before leaving participants are scheduled for the first PST session with Coach Lumen. As needed, Lumen intervention orientations can be done remotely.

Intervention data collection. Secure study iPads used by participants in the intervention arm are enabled with Lumen PST. Participants complete 8 PST sessions beginning with 4 weekly and then 4 biweekly intervals over 12 weeks on their assigned iPad. At the end of each PST session, participants are prompted to schedule their next session with Coach Lumen. They receive automated reminder notifications 1 day prior to their next session and on the session day and have the opportunity to make up missed sessions. Intervention participants also complete the PHQ9 and GAD7 and user experience surveys at all PST sessions. Upon completion of the intervention participants will receive a certificate of completion signed by Dr. Jun Ma. They may also complete a post-intervention interview via Zoom to characterize participant perspectives regarding their Lumen use experience.

Intervention fidelity assurance. All PST sessions that participants complete with Lumen are audiorecorded. Session recordings and transcripts are stored on the Amazon Web Services (AWS) in study configured accounts, with coded identifiers (see Section 7.). These materials will be used to evaluate treatment fidelity and user-Lumen interaction.

7.0 Expected Risks/Benefits

This is a minimal risk study with prudent measures in place to protect the health and well-being of research participants and their privacy and confidentiality. Risks associated with participation in this study may include the potential for the following:

- Potential for self-harm
- MRI-related injury, discomfort, and distress (Aim 2: pilot RCT only)
- Patient privacy and confidentiality breach
- Accidental recording.

These risks are largely associated with the characteristics of the patient population to be studied and the procedures involved in the research. The target population includes patients with moderate, untreated depressive and/or anxiety symptoms. The risks are reasonable in relation to the anticipated benefits and are minimized by using procedures that are consistent with sound research designs and established research and clinical protocols. The following measures are implemented to minimize potential risks to participants in the study.

Protection against risks of self-harm. Some of the questions about depression, thoughts of death and other psychological symptoms and conditions as a part of study assessments may cause discomfort for some participants. However, in general the questions are not particularly intrusive or distressing, and stress is likely transient. In addition, participants are free to refuse to answer any questions. It is widely accepted that asking questions about thoughts of death or suicide does not lead to increased risk of suicide. Nevertheless, in the event that a patient is identified as being suicidal during the screening or follow-up phase of the study, the following self-harm protection protocol (adapted from the UIC IRB

approved ENGAGE-2 study #2018-1174), is in place to immediately alert the study supervising psychiatrist to assess the patient's suicidal thoughts by telephone, followed by notification of the participant's PCP and appropriate clinical action if necessary. If a participant responds "1" ("several days"), "2" ("more than half the days"), or "3" ("nearly every day") to the PHQ-9 question "Over the last 2 weeks, how often have you been bothered by thoughts that you would be better off dead or thoughts of hurting yourself in some way?", we further assess the participant's level of risk by asking "*Do you have a plan for how you would commit suicide?*" and then follow the protocol based on the assessed level of risk. Individuals reporting suicidal ideation on PHQ9 with an active plan at eligibility screening are excluded from participation. Nonetheless, the risk of emergent suicidality still exists after enrollment. Risk for suicide may be detected when a participant completes the PHQ9 at the beginning of each PST session with Lumen. Participants who score 1-3 (see above) on item 9 of the PHQ9, will be offered the option to call an emergency contact, the national 800-SUICIDE/800-273-TALK hotline, or 911 at the time of detection. This information is a pre-programmed script for Lumen. The supervising study psychiatrist, Dr. Ajilore, the intervention manager, Mrs. Ronneberg, receive immediate notifications, as outlined in the table below. The study psychiatrist calls the participant within 3-4 days to conduct an assessment, assess need for further referral, and discuss referral options depending on urgency, needs and preferences, available supports in place, insurance status and routine source of care. The study psychiatrist may refer the participant for immediate comprehensive evaluation (e.g., at a local emergency department), ambulatory psychiatric services, or community resources, with personal accompaniment depending on the evaluation outcome.

<table>
<tr><td colspan="2">[If PHQ9 completed with SPEAC study staff by phone]</td></tr>
<tr><td>If YES, participant has active plan for self-harm:

• Explain to participant: I am concerned for your safety and therefore need to call for help right now.
• Get participant's location
• Stay on the phone with participant and use another phone to call 911.

[Script: "I want to report a self-harm alert for a UI Health research participant who has just endorsed suicidal ideation to me (by phone) - I want to provide his/her name, location, and phone number (any relevant detail the pt provided.)"]
• Send High-priority Page to on-call Study Physician.

CRC Note: You do NOT need participant's consent to call 911 if you feel there is a possibility of immediate risk of harm to self or others.</td><td>If NO active plan or DECLINE TO STATE, explain to participant:

I am not a clinician; however, our study has clinicians who speak with any participant who tells us they've been feeling this way recently. I will have a study doctor call you within the next 3-4 days. I would also like to give you 2 national helpline and 1 text line numbers that you may find helpful. All numbers are available 24 hours/7 days a week. We will work together to get you feeling better.

National Hopeline Network: 1-(800)-SUICIDE or 1-(800)-784-2433

National Suicide Prevention Lifeline: 1-(800) 273-TALK or 1-(800)-273-8255
Crisis Text line: text START to 741741</td></tr>
<tr><td colspan="2">[If PHQ9 completed by Participant online]

[The following pop-up message appears if a participant responds "1" ("several days"), "2" ("more than half the days"), or "3" ("nearly every day") to the 9th question, regardless of active</td></tr>
</table>

action plan or not.]

Please note: We do not monitor this screener in real time, if this is an emergency call 911.

For more immediate attention, because you have been bothered by thoughts that you would be better off dead or of hurting yourself in some way in the last 2 weeks, you should call your physician or other healthcare professional right away or go to the emergency room.

You may also call the National Suicide Hotline at 800-SUICIDE / 800-784-2433 or the National Suicide Prevention Lifeline at 800-273-TALK / 800-273-8255. You may also text START to Crisis Text line number 741741. All these numbers are available 24 hours every day.

We will have a study doctor contact you within 3-4 days. In the meantime, do not delay seeking medical attention.

Protection against injury, discomfort, and distress with brain imaging. (Pilot RCT only)
MRI is non-invasive, widely used, and safe. Routine contraindications to MRI include presence of any metal implants (pacemaker, aneurysm clips, neurostimulators, cochlear, eye implants, old or very fresh tattoos). Participants are thoroughly assessed for MRI eligibility during screening using the well-established procedures as in the ENGAGE-2 study (protocol #2018-1174). A small number of people may feel claustrophobic inside the MRI machine. The study can be immediately stopped via button press if this occurs. Sometimes subjects report a temporary, slight dizziness or light-headedness when they come out of the scanner. Study personnel and technicians are on site during all acquisitions to address any such discomfort. To minimize fatigue, sufficient breaks are provided to participants during the scanning procedure, and the study coordinator conducting the scan regularly inquires as to whether there is anything that s/he can do to facilitate the participant's comfort. In the event of adverse effects related to MRI scanning, study personnel and medical staff are on-site for consultation and assistance. The UI Hospital Emergency Room is less than a 5-minute drive from the scanning center. If there are any adverse events at any time during the MRI procedure, the study coordinator terminates the scanning session, provides a debriefing, and contacts study psychiatrist (Dr. Ajilore) for assistance and follow-up for the participant.

According to the National Institute of Mental Health (NIMH) Council Workgroup on MRI Research and Practices (September, 2005), "there is no known risk of MR brain scanning of a pregnant woman to the developing fetus for scanning at 4T or less, and no known mechanism of potential risks under normal operating procedures." Notwithstanding, subjects are warned about potential risks not yet discovered in the informed consent form. Before each scan, female participants are asked if they are or are trying to become pregnant and when they had their last menstrual period. Any woman who indicates that she is pregnant will not be scanned. Any woman who indicates that she is trying to become pregnant or who is experiencing a late menstrual period is asked to complete a urine pregnancy test, which if positive will preclude the subject from being scanned.

If study personnel observe any unusual features in the MRI scan at the time of acquisition or an incidental finding, or abnormality on MRI scans, staff requests a clinical neuroradiological report on the scan from the neuroradiologists at the Center for MR Research who are on call to provide these reports as required, as part of the infrastructure of the Center's facility. During the consenting process, all participants are informed about the potential risks of discovering an incidental finding or abnormality on their MRI scan. If an abnormality is found in a participant's MRI scan, PI Dr. Ajilore contacts the participant and refers him/her for medical follow-up for the problem if the participant requests, including a referral to a PCP. If

a participant has a PCP, the PI contacts the PCP, at the request of, and with verbal permission from the participant, to inform him/her of the finding on the MRI scan and to help him/her get the participant appropriate follow-up. The decision as to whether to proceed with further examination and/or treatment lies solely with the participant and his/her PCP.

Protection against breaches of participant privacy and confidentiality. All investigators and their staff are adequately trained to protect participant privacy and sign an agreement to do so. All information obtained from research participants during the study are considered strictly confidential and are only used and disclosed as permitted under the HIPAA regulations. All eligible participants must sign a HIPAA authorization as part of the informed consent form in order to participate. They also have the option to agree to sharing of their de-identified data through the NIMH Data Archive (NDA). Only aggregate data will be included in scientific presentations and publications resulting from this study.

To ensure data security, security protocols are incorporated at multiple levels. First, participants are assigned unique anonymous IDs ("Study Participant ID") that are used as a proxy for mapping all PHI and PII (personally identifiable information). A mapping document that links Study Participant IDs to the corresponding PHI and PII is maintained separately from all study-related data of participants and stored in a secure manner (e.g., encrypted REDCap database on a secure server behind the university's fire wall). Only a designated subset of study personnel who have a need to know (e.g., PIs and recruitment personnel) have access to the mapping document.

Aim 1: Formative Research

Formative user study sessions: present minimal risk to research participants and prudent measures are taken to protect their privacy and confidentiality. Participants are informed in advance about the discussion topics so that they may make an informed decision to participate. For the formative user study sessions, participants are made aware of study team observers present during the sessions. Study personnel alerts participants before beginning audio-recording of sessions. Transcriptions of audio-recordings do not identify participants, instead individuals are referred to as "Respondent #." If names or locations of attendees are recorded during a session, that information will not be included in the transcripts. Federal regulations require research records be retained for at least 3 years after completion of the research.

Aim 2: Pilot RCT

In addition to the protection measures described above, randomized participants who begin intervention per group assignment, receive study a iPad which utilizes the in-built encryption and 6-digit passcode protection and is configured with their first name, (as per informed consent.) Only first name will be used by Coach Lumen during intervention to facilitate participant engagement for treatment alliance. Additionally, the study iPads are in a "lock down" mode where participants cannot use them for any purposes but for tasks associated with the study.

No study data resides locally on the iPad. In order to maintain anonymity of the intervention participants, no personal identifiers (other than first name) are used during Lumen PST session data acquisition or storage. All Lumen PST session recordings and transcripts acquired by the Amazon Alexa application using the study iPad are stored on the AWS cloud within study configured accounts. The study-established AWS accounts are assigned coded "AWS IDs" (which are paired with "Study Participant ID") are accessible only by study staff with a need to know. In case of a potential loss or theft of the study iPad, participants are instructed to immediately notify study staff and the AWS ID account access will be disabled.

The SPEAC study is a registered HIPAA account under a standardized Business Associate Addendum (BAA) held by UIC/UI Health with (AWS) which enables the covered entities to be HIPAA compliant.

Finally, participants are able to keep their study iPad for personal use after their AWS ID account access is disabled. Participants will be provided with instructions to remove the Lumen skill and to unlock the iPad to have a fully enabled device to keep. Alternatively, participants may choose to redeem the iPad for $100. Participants in the waitlist control arm in the pilot RCT may choose to attend a Lumen Orientation visit to receive training and a Lumen PST-enabled iPad after the end-of-study assessments at 16 weeks post randomization. These options are explained to study participants during the informed consent and reiterated to waitlist control participants who choose to receive Lumen PST after the end of the study.

Protection against accidental recording. Accidental recording is a commonly raised concern regarding voice-enabled technologies. In the case of Lumen, the application is installed on the locked-down and encrypted study iPad. In order to prevent accidental recording, the following protections are implemented: First, the iPad has a keypad lock. Recording takes place only in the unlocked mode, after the participant opens the Alexa app, and gives permission verbally (using specific "wake phrases" such as "Launch Lumen Session," or by tapping the option on the iPad screen. Second, participants are informed that all Lumen sessions are recorded and give explicit permission for this as part of their informed consent process. Third, if the study iPad is left unattended during an active session, the iPad locks after 10-seconds, and after 30 seconds of no verbal input from the participant, the Lumen session shuts down (preventing any further recording). Once the participant returns, they must go through the process of unlocking iPad and waking Lumen to actively resuming their session.

Potential benefits to participants. Eligible participants in the study have mild or moderate depressive and/or anxiety symptoms that are not treated. They receive PST by interacting with Lumen on a secure study iPad. Waitlist controls have the option to receive Lumen after their 16-week assessments. PST is a brief skill-enhancing psychotherapy with robust evidence for efficacy in treating depression and anxiety. PST focuses on improving one's skills involved in solving personal real-life problems and is easy for most people to understand. Its stepwise, patient-driven approach is also easy to follow. Lumen is a voice-enabled, PST-trained agent who will undergo iterative design development and refinement based on information gathered from evaluation cycles and rigorous testing. Through interacting with Lumen, participants in the study may benefit from having PST at their fingertips and experience improvements in psychological symptoms, functioning and quality of life. If shown successful in this developmental R61 phase of the project, Lumen will undergo further confirmatory testing (in the R33 phase) and may downstream carry the potential to be scaled for use by people with depression and/or anxiety who cannot or will not seek professional help or who desire as-needed, personal help beyond what conventional treatment models can offer.

8.0 Data Collection and Management Procedures

Aim 1: Formative Research

Scenario-Based Evaluations by the 2 independent PST master trainers are completed and stored on REDCap at UIC.

Formative User Study data from audiorecorded Zoom conference session, with participant's explicit permission.

Aim 2: Pilot RCT

The study coordinators are trained on data collection according to standardized protocols, and their performance is continuously monitored. The following types of data are collected.

1. Data on diagnoses, prescriptions, clinical encounters, and hospitalizations are abstracted from EHR for eligibility screening and baseline characterization.
2. fMRI data and physical measurements are collected at the UIMC Advanced Imaging Center.
3. Survey data and ecological daily assessments are collected using self- and interviewer-administered questionnaires on REDCap, a HIPAA-compliant server.
4. With participant's explicit permission, PST sessions are audiorecorded and are stored in study configured accounts on the AWS cloud, accessible only by study staff.

See Appendix 2 for the list of study measures and assessment schedule.

8.1 Data Management

Neuroimaging data management. There are Linux machines dedicated to various aspects of image transfer and storage available to the MPI Dr. Ajilore and team within the UIC Psychiatric Institutes. The Institute also houses an imaging processing server for data storage and analysis (Dell PowerEdge R900, 2.13 GHz, Xenon Four Cores, 128GB RAM with 6 TB of data storage). These machines are networked with a T3 line to facilitate image transfer and back-up of data from the Center for MR Research. The T3 line is connected to the University network and firewall shielded from the outside. These machines are for the sole use of the research team and as such the research team has its own log in for the system with direct access from our dedicated research computers.

Non-imaging data management. REDCap study database on a HIPAA-compliant server, occur automatically according to a preset schedule (no participant action required). Data are immediately available for inspection and reporting of recruitment and retention status and any missing data. The data analyst performs weekly quality controls.

All datasets are cleaned, verified, and archived. One official copy of all study data and a master data dictionary are maintained and updated regularly. All analytic and tracking databases are stored on a HIPAA-compliant server with continuous backups. For the protection of participant confidentiality, unique anonymous study IDs are used for data storing, tracking, and reporting. PHI is stored separately from all other study data and will be used and disclosed in accordance with the HIPAA regulations. Regular reports are produced on (1) patient accrual and follow-up completion/retention in relation to goals and timeline; (2) the randomization process and group comparability on the balancing variables; (3) key baseline characteristics of the sample, by (blinded) group, related to the primary and secondary outcome variables; (4) intervention exposure and adherence; and (5) protocol violations. Any observed delays in these processes or data irregularities shall be followed up and resolved in a timely manner.

Data sharing. De-identified fMRI data and other study data are shared with collaborators at UIC and other collaborating institutions via a UIC Box Health Data Folder. All collaborating institutions are listed as data users under the HIPAA regulations and authorization provided by participants. PHI (e.g., name, address, phone number) and the link between study ID and the patients' identities will not be shared. They can download data from the UIC Box Health Data Folder and store them on a HIPAA-compliant server at their institution for analysis according to the study protocol. We follow the latest industry standards for data encryption, server authentication, and client authentication to ensure secure data transmissions at all times. In addition, all the investigators and staff maintain up-to-date trainings and certifications in human subject's protection, HIPAA, and Good Clinical Practice (GCP). Furthermore, we will partner with the NIMH and use all NIMH data preparation and sharing policies as a guide to ensure the deidentified data and results, along with all associated documentation, are submitted to the NIMH

Data Archive (NDA), to be specifically deposited to the National Database for Clinical Trials Related to Mental Illness (NDCT). We will ensure we have permission from our participants to disclose deidentified participant-level data collected as part of this study to researchers who meet all the NDCT requirements for requesting use of the dataset. We will strictly comply with the HIPAA and IRB regulation requirements regarding research use of PHI and appropriate safeguards for sharing deidentified data.

9.0 Data Analysis and Statistical Considerations

Aim 1: Formative Research

Formative User Study. We will conduct the following analyses: (1) analysis of verbal interactions with Lumen to measure interaction efficiency, including breakdowns in conversations (e.g., stoppages) and pauses that break the flow and continuity, and consequently, the efficiency of the conversational interaction; (2) analysis of Digital Health Literacy Instrument,[51] task load (NASA Task Load Index (NASA-TLX)[44] of user workload along with 6 dimensions of mental, physical, temporal, performance, effort and frustration), user experience (User Experience Questionnaire-Short version (UEQ-S),)[45] and therapeutic alliance (Working Alliance Inventory adapted for digital interventions (WAI-Tech)[48,49] and (3) content analysis[61] of audiorecordings of the walkthrough sessions. Quantitative survey data are summarized with descriptive statistics using SAS software (version 9.4; SAS Institute Inc.). Qualitative data are analyzed using the content analysis method descried above.

Aim 2: Pilot RCT

Pilot RCT sample size calculation using a confidence interval approach: To obtain a precision interval with a standardized half-width of 0.50 (akin to a medium effect) with 90% assurance, we have planned a sample size of 60 (n_{Trt}=40, n_{Con}=20), assuming ≥85% retention at 16 weeks. This provides reliable estimates of between-treatment differences in change in amygdala activation for nonconscious threat-related reactivity and change in DLPFC activation for cognitive control from week 0 to 16 in Criterion 1. In Criterion 2, a sample size of 60 will also be sufficient to detect a correlation coefficient of r=0.4 with 0.80 power and 2-sided α=0.05. We are not focused on smaller correlations (r<0.4) because their implications are likely to have more limited clinical relevance.

The criteria to evaluate neural target engagement are as follows:

Criterion 1. The level of task-evoked activation for either the amygdala for nonconscious threat-related emotional reactivity or for the DLPFC for cognitive control improves meaningfully from week 0 to 16 in the Lumen group compared with the waitlist control group.

Criterion 1 is tested via *t* test that compares the change from week 0 to 16 in the Lumen group with that change in the waitlist control group. Participants are analyzed based on the group to which they are assigned using all available data. The *t* test described above can be readily expanded for exploratory subgroup analyses by applying to each subgroup (e.g., female vs. male).

Criterion 2. Change in the level of task-evoked activation of a neural target that meets Criterion 1 correlates with temporally concurrent change from week 0 to 16 in self-reported measures within the corresponding PST theory-based construct for emotional reactivity or cognitive control. We consider correlations of r≥0.4 meaningful.

To test Criterion 2, we examine the regression change of the amygdala activity from week 0 to 16 on change in self-reported measures of emotional reactivity (worry, affect) from week 0 to 16, controlling for baseline values of the self-reported measures and group assignment. We expect that the coefficient for the change in amygdala activation is significant and positive in the direction of improvement in its

corresponding, validated emotional reactivity self-report measures and that this relationship may be significantly modified by group assignment. Similarly, we examine the regression change of the DLPFC activity from week 0 to 16 on change in self-reported measures of cognitive control (problem solving, dysfunctional attitudes) from week 0 to 16, controlling for baseline values of the self-reported measures and group assignment. We expect that the coefficient for the change in DLPFC activation is significant and positive in the direction of improvement in its corresponding, cognitive control validated self-report measures and that this relationship may be significantly modified by group assignment.

In addition, we will also conduct exploratory and secondary analyses. We will conduct an exploratory whole-brain analysis (1) to verify our *a priori* specification of neural targets, namely, amygdala for emotional reactivity and DLPFC for cognitive control, and (2) to identify any additional voxel clusters that are activated by our tasks and are associated with treatment response. Secondary analyses compare differences in discrete changes in PROs from week 0 to 16 and trajectories of change based on intensive time-series ecological measures between the Lumen and waitlist control groups. Analyses of the treatment effects on PROs employ the same *t* test approach as described above for testing Criterion 1 for target engagement. Furthermore, dosing analyses are conducted within the Lumen treatment group. The primary dosing analysis examines the relationship of neural target activation in the amygdala and DLPFC to the 4 measures of dose (number and length of PST sessions, fidelity-weighted number, and length of PST sessions). The secondary dosing analysis examines the relationship of patient-report outcomes, ecological end-of-day assessments, and measures of treatment progress (e.g., PHQ9 and GAD7 across PST sessions) to the same 4 measures of dose. Given neural target activation measures and PROs are obtained at weeks 0 and 16, these analyses focus on change over the full 16-week span of the study, whereas ecological end-of-day assessments and treatment progress measures are available for a total of 8 occasions each, and thus will be considered longitudinally.

Please see Appendix 3 for detailed power analysis and statistical analysis plan.

10.0 Quality Control and Quality Assurance

REDCap built-in quality control features (e.g., data type validation, valid values/range rules, required responses) and extensive data rules are used to allow for complex real-time quality control (QC) checks (e.g., of missing values, out of range values, logic comparisons, cross form/event consistency). When issues are identified through weekly and monthly scheduled QC checks, the data analyst will use the "REDCap data resolution workflow" feature, a built-in data query tool, to communicate the issues to study coordinators for resolution. Study coordinators can add notes or corrections within the data query tool and have editing rights to alter the source data, but all edits must be approved by the project manager. REDCap tracks all correspondence and decisions of the query process, and it maintains an extensive audit trail of access, entry, and edits of stored data. REDCap also allows for form and record locking capability with e-signature management.

11.0 Data and Safety Monitoring

The following Data and Safety Monitoring Plan (DSMP) will be followed to ensure the safety of study participants and the validity and integrity of data in compliance with NIMH requirements.

Independent Oversight. Because the risks for adverse events (AEs) and data breaches are minimal in the pilot RCT, and the study is not a Phase III clinical trial, it is determined that a Data and Safety Monitoring Board (DSMB) is not needed. Instead a safety monitor *independent of* the study is responsible for overseeing the implementation of this DSMP to ensure (1) the protection and safety of human subjects and (2) the validity and integrity of the trial. Dr. Bernice Man, MD, o at the University of Illinois at Chicago (UIC), serves in this role. The safety officer is responsible for examining aggregate data on

recruitment and retention, adverse events (unexpected or serious), and subject complaints, but he will not be reviewing individual, identifiable, subject data. If unexpected or serious adverse events occur, these will be reported to the IRB in the form of Prompt Reports. See further details below.

Contact PI Dr. Ma and senior biostatistician Dr. Xiao, both of whom will be blinded during the pilot RCT, meet with Dr. Man every 6 months during active participant recruitment and follow-up to present and discuss data and safety monitoring information (below). For the annual continuing renewal application to the IRB and the annual progress report to the sponsor each year, Dr. Man will provide a summary of her findings and recommendations regarding the following:

- ascertainment and any actions to be taken in response to AEs and SAEs reported during the study
- reports related to study operations and the quality of the data
- possible modifications in the study protocol concerning recruitment, participant retention, data quality, or trial operations more generally.

Safety Monitoring. As in any clinical trial, it is not possible to anticipate all possible AEs. Staff will undergo extensive training in ascertaining, monitoring, and documenting AEs—serious or not. The study investigators have extensive experience in clinical trial organization and management, including data and safety monitoring for single site and multisite trials. Established procedures for rendering first aid and life-threatening emergencies will be monitored by Dr. Ajilore (psychiatrist).

An AE is defined as any untoward medical or psychological event experienced by a patient during or as a result of his/her participation in the study that represents a new symptom or an exacerbation of an existing condition whether or not considered study-related based on appropriate medical judgment. SAEs are any adverse experience that results in any of the following outcomes:

- Death
- Life-threatening event/illness
- Inpatient hospitalization or prolongation of existing hospitalization
- Persistent or significant disability/incapacity
- Pregnancy resulting in a congenital anomaly/birth defect
- Any event requiring medical or surgical intervention to prevent permanent impairment or damage. In this study, this is defined as physician confirmed diagnosis of any of the following: angina pectoris, heart attack, stroke, transient ischemic attack, heart failure, coronary angioplasty or bypass surgery, peripheral vascular disease, , any other serious injury to the bone or muscle, liver failure, kidney failure, and cancer (except for non-melanoma skin cancer).

Non-serious AEs are all adverse events that do not meet the above criteria for "serious." To ensure unbiased ascertainment, AEs will be systematically identified by querying participants at 16-week follow-up visits using the AE Patient Query Form. In order to access the electronic medical record, as needed, for AE adjudication, consenting participants are asked to provide and/or confirm the minimum necessary identifiers including first and last name, date of birth, mailing address, and whether they have received care at UI Health.

Blinded Reporting. In the pilot RCT, safety information will be monitored while keeping the true identity of the study groups masked. In his role as the study physician, Dr. Ajilore may need to become unblinded. But Contact PI Dr. Ma and biostatistician Dr. Xiao as well as the independent safety monitor Dr. Man will be blinded throughout the study and will review summaries of the numbers and rates of all AEs by blinded treatment group. Proper blinding of the other investigators and outcome assessors will be enforced as well.

Requirements for Adverse Event reporting. The Contact PI , Dr. Ma, will inform the Director of IRB panel and all relevant oversight committees at the university within 5 business days of learning of an unanticipated AE or major protocol deviation. All relevant information will be reported to the IRB for each unexpected SAE including information about the event and its outcome, dosing history of a suspect medication/treatment, concomitant medications, the participant's medical history and current conditions, and all relevant laboratory data. Within 15 business days of the PIs becoming aware of changes in risk/benefit or events requiring report to the sponsor, these will be reported. This timeline satisfies the NIMH reporting requirements for AEs and unanticipated problems. An annual report will be submitted to the IRB and to the sponsor summarizing all AEs, serious or not.

12.0 Regulatory Requirements

Role of Collaborating Institutions in the Oversight and Conduct of Human Subjects Research. The proposed study is a collaboration between UIC (primary institution), Pennsylvania State University (PSU), Washington University in St. Louis (WashU), and Stanford University. All study participants are patients of UI Health, and the UIC IRB is the IRB of record. Dr. Jun Ma, MD, PhD, and Dr. Olusola Ajilore, MD, PhD, both of UIC are co-principal investigators. Also, at UIC, Dr. Philip Yu, PhD brings expertise on machine learning and AI development, and Dr, Ben Gerber, MD, MPH provides expertise on implementation and evaluation of digital health behavioral interventions in disadvantaged populations of diverse racial and ethnic backgrounds. As the Study Physician, Dr. Ajilore (psychiatrist) oversees the clinical aspects of the study, involving patient recruitment and safety, and as needed, coordinate care with participants' PCP. For more details, see Section 7. Dr. Joshua Smyth, PhD at PSU is a renowned expert in stress resilience and recovery and ecological momentary assessment (EMA). Dr. Smyth collaborates with the UIC team and is responsible for the PSU research team. Dr. Thomas Kannampallil, PhD at WashU brings expertise on Lumen design and development, human-computer interaction, and bioinformatics. Dr. Lan Xiao, PhD at Stanford is an experienced biostatistician and is responsible for developing and executing the statistical analysis plan of this project.

Informed Consent. All study participants involved in both the formative research and the pilot RCT must provide eConsent. The eConsent is obtained remotely via telephone or online video conferenced teleorientation session with trained study staff who leads the participant through the consent process and answer any questions about the study. Study personnel obtaining informed consent are experienced in obtaining informed consent and receive standardized training in trial-specific protocols. Risks and benefits and the voluntary nature of the study will be thoroughly explained by the study personnel. The participant will have as much time and information as needed to consider whether or not to participate.

Aim 1: Formative Research

ENGAGE-2 participants (enrolled but not active in the 2018-1174 protocol) who first screen eligible for access to Lumen compatible Internet-enabled devices, are scheduled for teleorientation with study staff who lead a full informed consent discussion and use the eConsent framework to provide written informed consent documentation (ICD).

Aim 2: Pilot RCT

The study uses a 2-step consent process, an online screening consent is first obtained prior to initial eligibility screening, which facilitates participant self-screening. Second, screened eligible and interested participants are invited to attend a teleorientation with study staff who lead a full informed consent discussion and use the eConsent framework to provide written informed consent documentation (ICD).

All participants document their consent electronically using the secure REDCap eConsent framework and obtain their pdf copy of the signed consent by directly downloading through REDCap. The study copy is

automatically archived in the file repository in REDCap database, which is a HIPAA-compliant server, accessible to the PIs and study personnel.

Subject Confidentiality. See Protection against breaches of participant privacy and confidentiality in Section 7.0.

Unanticipated Problems. An unanticipated problem (UP) is defined to include any incident, experience, or outcome that meets all of the following criteria:
* Unexpected (in terms of nature, severity, or frequency) given (a) the research procedures that are described in the protocol-related documents, such as the IRB-approved research protocol and informed consent document; and (b) the expected natural progression of any underlying disease, disorder, or condition of the subject experiencing the UP and the subject's predisposing risk factor profile for the UP;
* At least possibly related to participation in the research (possibly related means there is a reasonable possibility that the incident, experience, or outcome may have been caused by the procedures involved in the research); and
* Suggests that the research places subjects or others at a greater risk of harm (including physical, psychological, economic, or social harm) than was previously known or recognized.

Within 5 business days of the PI learning of a UP or major protocol deviation, the PI informs the Director of IRB panel and all relevant oversight committees at the university. Within 15 business days of the PI becoming aware of non-serious AE, changes in risk/benefit, or events requiring report to the sponsor, these will be reported. An annual report will be submitted to the IRB and the sponsor summarizing all AEs, serious or not.

APPENDICES

Appendix 1. Design and Training of Lumen and Participant Interaction

Section A: Lumen Architecture

Lumen is a voice-only virtual coach that delivers Problem Solving Treatment (PST) to counsel participants with depression and/or anxiety using the 7-step problem solving process and the SSTA (stop, slow down, think and act) method of coping. Lumen conducts interactive conversations providing appropriate responses based on participant input. In order to deliver seamless interactions, we developed a robust and parsimonious architecture that combines PST, integrated components to provide context for Lumen conversations, data storage for the interactions, and a software infrastructure for providing security and privacy for these interactions.

Lumen architecture (see Figure 1) was developed with input and consultation from researchers (in computer science, medicine, health service researchers, and psychiatry), software developers, and human computer interaction experts. The architecture was developed on Amazon's Alexa platform, with further integration with a secure REDCap system for ecological momentary assessments (EMAs) and surveys. Lumen's software architecture comprises of two components: a conversation manager module and a context manager module.

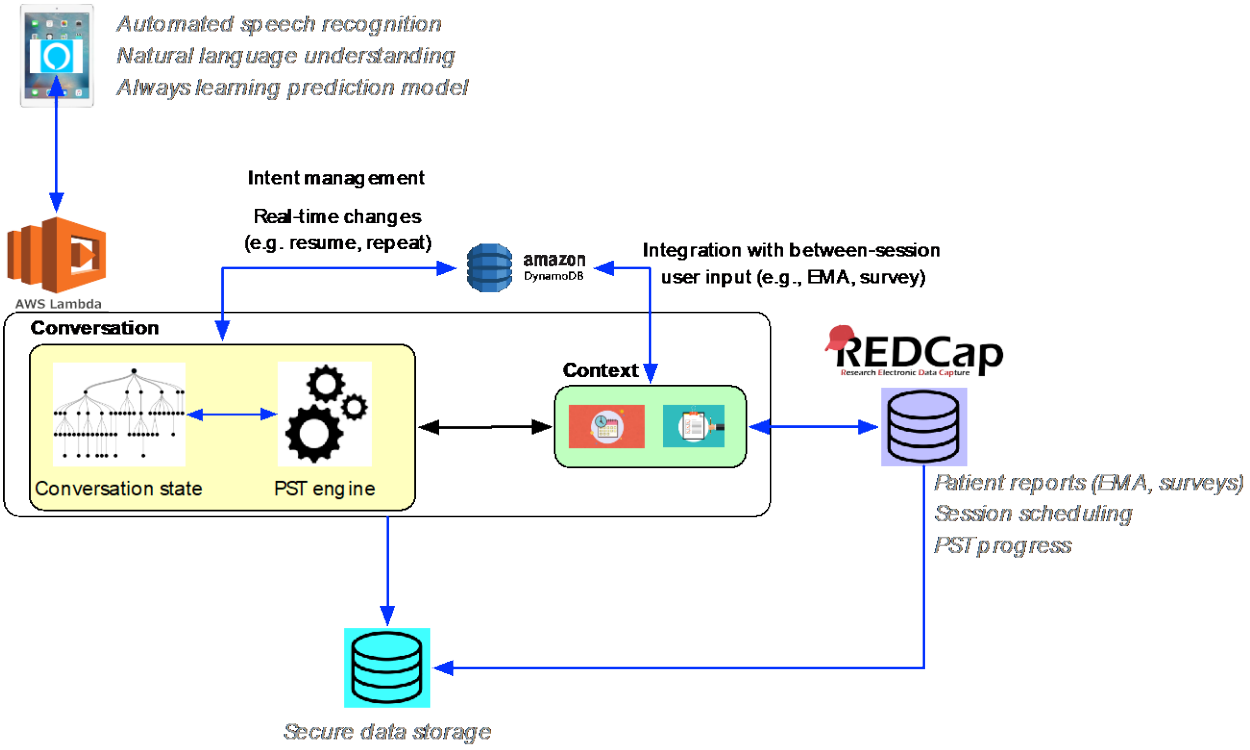

Figure 1. Lumen software architecture that includes the conversation manager and context manager modules.

The **conversation manager** is the voice-interactive component of Lumen and is responsible for delivering PST content. This PST content is aligned with the theoretical constructs and associated treatment guidelines. Towards this end, the conversation manager includes a PST engine that incorporates PST-related and conversational structures. The PST engine is a flexible set up that allows for the adaptation of PST content for other health-related problems (e.g., translating the therapy for weight management or smoking cessation) in the future.

Within the Lumen application, the PST engine tracks the progress of a PST session. For example, in a session, once the participant defines a specific problem to work on (step 1), Lumen will proceed to guide the participant in establishing a realistic goal for solving the problem (step 2). Then, once the participant has defined a set of potential solutions that can meet the goal (step 3), Lumen will guide the participant on evaluation of the pros and cons of each solution (step 4). After the participant chooses his/her preferred solution (step 5), Lumen will guide the participant to develop an action plan (step 6), which Lumen will prompt the participant to implement and evaluate the outcome post the session (step 7). This stepwise structure is the same in all PST sessions as is the case in current practice, and the participant identifies the problem in each session (except session 1, which is an introductory overview session situating the PST process). Importantly, each of these steps is based on established PST theory and practice.

In addition, the conversation manager also manages the "state" of the conversation—including the flow of interactive conversation (e.g., the flow of the above-mentioned steps), and adaptive responses based on user input. The conversational state is adaptively managed based on user input. Towards this end, user "intents" or statements are parsed and mapped into pre-defined categories, which are then mapped to the appropriate state in the PST interaction to deliver relevant content, aligned with the user input. The conversation manager also provides dynamic support for functions such as resume, which allows participants to re-start a previously incomplete PST session. Such stop-restart functions provide flexibility in conversational interactions, affording perceptions of realistic conversational interactions.

The conversation manager also interacts with a **context manager** module that provides situated and contextual information regarding Lumen sessions. The context manager primarily controls three aspects: persistence of therapy content across sessions, scheduling/re-scheduling of sessions, and integrating external content (e.g., surveys, EMAs) into the Lumen sessions. The context manager dynamically tracks user problems (see step 1 above), and potential action plans that were previously developed and evaluates adherence to those plans in ensuing sessions. Such dynamic follow-up increases the persistence and continuity of therapy across sessions, developing trust and confidence in the virtual coach. Similarly, the context manager tracks session progress (e.g., session 5), and helps participants schedule and re-schedule sessions. With its integration with an external scheduling database (in REDCap), follow-up emails and messages are tracked to ensure that the therapy sessions are synchronized with participant needs. Finally, the context manager also interacts with the REDCap database to incorporate survey responses (e.g., completion of the PHQ-9 and GAD-7 surveys) into the Lumen session. For example, prior to the start of each Lumen session, participants are asked to complete the PHQ-9 and GAD-7 surveys; if incomplete (or partially complete), Coach Lumen reminds the user to complete the survey prior to re-starting the session. Further, if a user delays session attendance after having completed the surveys, Lumen checks whether a survey is current (within 3 days) to

the session and directs the user to complete an updated survey, if necessary. Additionally, the context manager assures that PST session intervals are maintained by preventing premature session attendance, enabling sessions after 6 days of the last session completed (for sessions 2-4) or after 13 days (for sessions 5-8) for treatment fidelity.

All sessions and communicative interactions are stored in a secure AWS-based database for analysis. In order to prevent accidental recording, and for pragmatic implementation in a clinical trial, we currently have implemented the entire Lumen infrastructure in a "locked down" mode iPad. Participants can access the Lumen application skill within the Alexa application by stating "Alexa Open Lumen…." A summary of the Lumen components is provided in Table 1.

Table 1. Lumen components, and their associated functions.

Lumen Component	Functions
Conversational manager	Managing conversations based on user intent, aligning with PST constructs, additional functions to manage conversations (e.g., repeat, resume), tracking progress within a session
Context manager	Tracking user problems (from previous PST sessions) and action plans, integrating user responses from PHQ-9 and GAD-7 surveys within sessions, scheduling/re-scheduling sessions,
Lumen REDCap database	Scheduled delivery of Ecological daily assessments, patient surveys, patient calendaring and reminders; Event window checks on session intervals and survey expiration.
Storage/Data Management (AWS & REDCap)	Comprehensive storage of user responses to surveys, user interactions with the Lumen, synchronization across devices, data security and privacy

Section B: Interacting with Lumen

Lumen architecture that is presented in Section A, is realized through user interactions with the Lumen skill embedded within the Alexa application (currently delivered on an iPad device) and user completion of surveys and EMAs with hyperlinks delivered via emails and text messages (on the participant mobile phones) (see **Figure 2**).

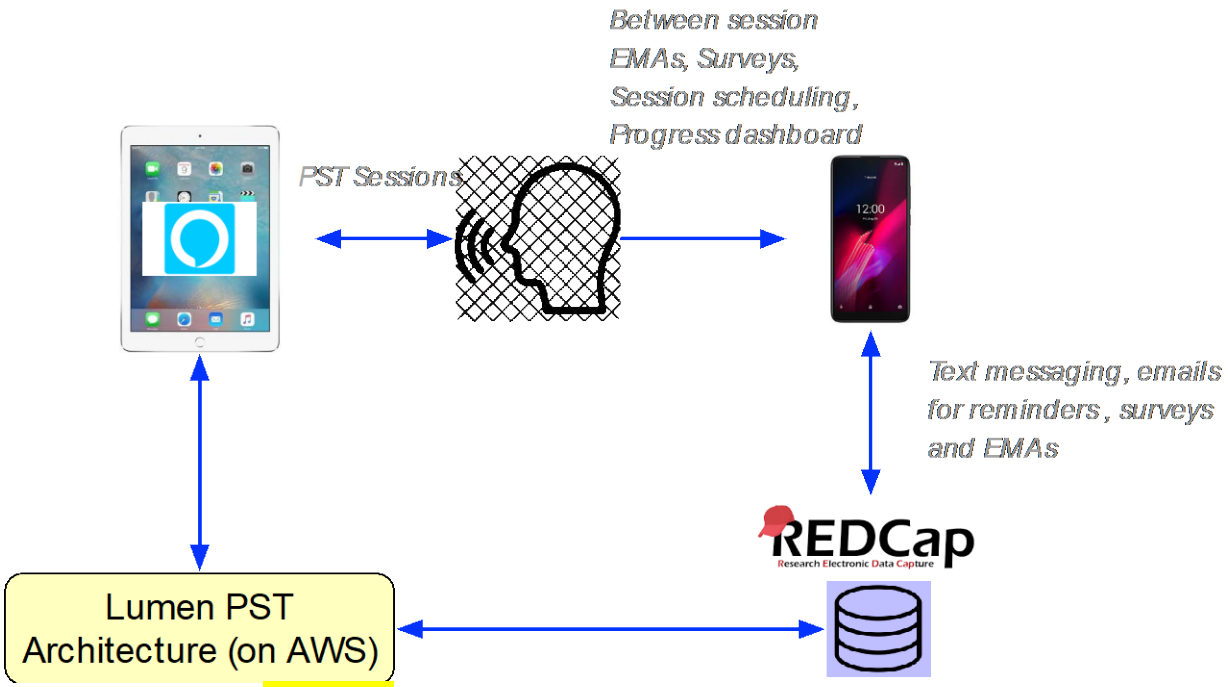

Figure 2. User interaction with Lumen for PST sessions

Participant interaction with the Lumen PST coach involves two primary components: (a) sessions with the coach, and (b) completion of surveys, EMAs, administrative management (e.g., session scheduling, re-scheduling, and monitoring progress).

Participants access their scheduled Lumen session via the Alexa application on their assigned iPad. Sessions are instantiated by the participant by saying "Alexa Open Lumen." Conversations then continue based on the progress that participants have made (e.g., number of sessions completed), current problem(s) being addressed and implementation of previously created goals and action plans. As previously described, these conversations are aligned with PST's treatment protocol and guidelines and with conversational structures and flow that reinforce the patient-centered approach of PST. As with regular conversations, participants can exit, resume or ask the Lumen coach to repeat parts of the conversations that they are unable to follow. At the end of each session, participants schedule their next session with Lumen.

Lumen's PST sessions are delivered on a dedicated iPad, which is encrypted and functions in a "lockdown" mode with no additional functions or local storage. The purpose of using such an approach is for trialing it in a controlled environment without compromising on data safety and privacy of the user. Additionally, such an infrastructure and set-up provide several advantages. First, it creates the perception of a "device as a therapist" mode, assigning a specific role-based purpose (i.e., Lumen as a health coach, that is delivered only on the specific device). Second, embedding the Lumen skill within the Alexa application reduces the potential for accidental recording (as is common with smart speaker devices such as Echo or Google Assistant). Finally, the encrypted mode with no local data storage allows for data privacy protection and remote management, in case of a lost or misplaced device.

Between sessions, participants will also complete surveys and EMAs to track their progress. These administrative tasks and surveys, which provide situated context for Lumen interaction and monitoring of progress will be delivered on the participant's personal devices. The surveys and EMAs are sent as text messages with embedded links. Participants complete these on any browser associated with their mobile phone. Additional messages will include reminders about sessions, ability to schedule/re-schedule planned sessions, and other session related information. All of the survey, EMA and scheduling tasks are instantiated through a dedicated REDCap project. Based on an initial session date a program calendar according to PST guidelines (4 weekly, then 4 biweekly sessions) are set up and used for all session reminders. Future session appointments are confirmed or adjusted per user preference during Lumen sessions and are automatically updated in REDCap to assure that survey are delivered at appropriate times.

Appendix 2. SPEAC Schedule of Measures – Aim 2 Pilot RCT

Measure	Instrument	Collection Method		Time (weeks)[a]		
			Screening	0	16	1-12
Eligibility criteria	Age, sex, current/planned pregnancy/lactation, etc.	Self-report	**X**			
Depression severity	Patient Health Questionnaire-9 (PHQ9), suicidal ideation (item #9)	Self-report	**X**			
Anxiety	Generalized Anxiety Disorder Scale (GAD7)	Self-report	**X**			
Alcohol/substance abuse	CAGE Adapted to Include Drugs (CAGE-AID)	Self-report	**X**			
Brain scan screening	Brain scan screening questions (e.g., self-reported weight)	Self-report	**X**			
Cognitive Impairment	Callahan 6-item screener	Interview	**X**			
Emotional reactivity and cognitive control	fMRI	Measured		**X**	**X**	NA
Emotional reactivity and cognitive control	Penn State Worry Questionnaire[62] Positive and Negative Affect Schedule (PANAS)[63]	self-report		**X**	**X**	NA
	Social Problem Solving Inventory-Revised: Short (SPSI-R:S)[64] Dysfunctional Attitudes Scale[65]	self-report		**X**	**X**	NA
Depression and anxiety symptoms	Hospital Anxiety and Depression Scale (HADS)[66,67]	self-report		**X**[b]	**X**[b]	NA
Functioning	Sheehan Disability Scale,[68] Work productivity and activity impairment questionnaire (WPAI)[69,70]	self-report		**X**	**X**	NA
Quality of life	12-item Short-Form Health Survey (SF12)[71]	self-report		**X**	**X**	NA
Mood	Daily mood (items from circumplex model, capturing affective valence and arousal)	Ecological end-of-day assessments, per Event Window table in Appendix 4.		**X**	**X**	**X**[c]
Stress	Daily stress events (15 categories), perceived stressor severity, stressor-related thoughts			**X**	**X**	**X**[c]
Appraisal	Assessing daily implementation of "problem orientation" (challenge appraisals, optimism, self-efficacy, outcome expectancies)			**X**	**X**	**X**[c]
Coping	Assessing daily implementation of "problem solving-style" (attempts to understand problems, plan effective solutions to coping, impulsivity,			**X**	**X**	**X**[c]

	carelessness, avoidance behaviors)					
Depression and anxiety symptoms	PHQ9 and GAD7	Self-report and data tracking during PST sessions, per Event Window table in Appendix. 4		**X**[b]	NA	**X**[b]
Treatment acceptability	Session metrics: Number and duration of PST sessions completed; End-of-session Lumen user survey: NASA Task Load Index (TLX),[44] User Experience Questionnaire-Short version (UEQ-S),[45] adapted Working Alliance Inventory for digital interventions (WAI-Tech)[48,49]			NA	NA	**X**
Digital health literacy	Digital Health Literacy Instrument[51]	self-report		**X**	NA	NA
COVID impact	COVID impact survey[72]	self-report		**X**	NA	NA
Religious involvement	Duke University Religion Index[73]	Self-report		**X**	NA	NA
COVID impact on social functioning	COVID impact on social functioning survey	self-report		**X**	**X**	NA
Social network	Social Network Index[74,75]	Self-report		NA	NA	**X**
Loneliness	UCLA 3-item loneliness survey[76]	Self-report		NA	NA	**X**
Height	Height	Measured		**X**	NA	NA
Weight	Body weight	Measured		**X**	**X**	NA
Blood pressure	Blood pressure	Measured		**X**	**X**	NA
Sociodemographics	age, sex, race, ethnicity, education, race, ethnicity, income, household size, marital status, employment status, occupation, smoking/vaping	self-report		**X**	NA	NA
Adverse Events (AE)	AE form	Interview		**X**	**X**	NA

[a]To minimize missing data, participants will receive $50 at wk 0, study iPad (unlocked) or $100 (if opting out iPad) at wk 16, and $1 per daily diary.

[b]HADS is used as the independent outcome measure of depressive and anxiety symptoms at weeks 0 and 16. PHQ9 and GAD7 are used for 2 purposes: (1) eligibility screening (before week 0) and (2) treatment progress monitoring across PST sessions (in weeks 1, 2, 3, 4, 6, 8, 10, 12).
[c]In addition to ecological daily assessments that will occur for 7 days at weeks 0 and 16, they also will occur every 2 weeks (on weeks 2, 4, 6, 8, 10, 12). During active treatment in the Lumen arm (Studies 1 & 2) or the in-person PST arm (Study 2), this will occur for 3 days prior to, the day of, and 3 days after each scheduled PST session. For waitlist controls, this will occur for 7 days starting on Sunday of each assigned week.

Appendix 3. Power analysis and statistical analysis plan – Aim 2 Pilot RCT

A. Power Analysis

The primary objective of the Pilot RCT in the R61 phase is to evaluate neural target engagement based on the criteria (specified in B.1 below) in order to transition to the R33 phase. Research to examine neural targets as mechanisms of psychotherapy is an emerging field. No prespecified treatment difference of clinical importance or established effect sizes are available for our proposed neural target measures from fMRI. To address such early stage research, Julious[77] recommended that rather than powering in the traditional fashion to test a formal hypothesis of a (in truth unknown) desirable treatment difference, the sample size be selected in order to provide a given level of precision of the treatment effect estimates using the confidence interval (CI) approach, which Ma et al. have used in a completed NIH-funded pilot RCT.[78] Specifically, Julious[77] and others[79] recommended the use of precision intervals of prespecified standardized width to guide sample size determination. Julious[80] offered a further refinement to provide *Assurance* (akin to power) that a chosen standardized width of the precision interval contains the true mean treatment difference. To obtain a precision interval with a 2-sided standardized half-width of 0.5 (akin to a medium effect size) with 90% assurance, we have planned a sample size of 60 (n_{Trt}=40, n_{Con}=20), assuming ≥85% retention at 16 weeks. This will provide reliable estimates of *meaningful improvement* from week 0 to 16 in amygdala activity for nonconscious threat-related reactivity and in DLPFC activity for cognitive control among Lumen participants, compared with waitlist controls (target engagement criterion (1); see below). Tests of the 2 *a priori* neural targets—amygdala for emotional reactivity and DLPFC for cognitive control—are regarded as separate inferential domains, that is, not within one family of comparisons, and thus are not corrected for multiplicity (or familywise error rate) across domains.[81,82] Importantly, as explained, the objective of the Pilot RCT is not to claim statistical significance of a prespecified key hypothesis but to obtain reliable effect estimates with precision for a decision on transition to the next phase. In this context, multiple testing is not required or recommended.[81,82]

For criterion (2), we will examine correlations of change in amygdala activity with change in validated self-report measures of emotional reactivity (affect, worry), and correlations of change in DLPFC activity with change in validated self-report measures of cognitive control (problem solving, dysfunctional attitudes), among Lumen and waitlist participants. A sample size of 60 will be sufficient to detect a correlation coefficient of r=0.4 with 0.80 power and 2-sided α=0.05. We are not focused on smaller correlations (r<0.4) because their implications are likely to have more limited clinical relevance.

B. Statistical Analysis Plan

B.1. Treatment effects on target engagement

Based on literature and our preliminary data, we define neural targets *a priori* as activation of the amygdala for nonconscious threat-related emotion reactivity and activation of the DLPFC for cognitive control. The activation variables are from fMRI, which will be completed at weeks 0 and 16. Also, validated self-report surveys of emotional reactivity (worry, affect) and cognitive control (problem solving, dysfunctional attitudes) will be obtained at weeks 0 and 16. These self-report measures were chosen because of prior evidence for their role mediating the effect of PST on depressive symptoms.[83]

Note that improvement for emotional reactivity corresponds to a *decrease* in amygdala activity whereas improvement for cognitive control corresponds to an *increase* in DLPFC activity. For consistency in the presentation of analyses, the measure will be constructed such that higher positive values are better; thus, change toward a *higher value is improvement*.

Activation of either neural target satisfies 2 criteria:

i. Criterion 1. The level of task-evoked activation for either the amygdala for nonconscious threat-related emotional reactivity or for the DLPFC for cognitive control improves meaningfully from week 0 to 16 in the Lumen group compared with the waitlist control group.

   The ENGAGE Phase 1 project (UIC IRB protocol #2015-1324) showed significant improvement in amygdala activation for nonconscious threat-related emotional reactivity from baseline to 2 months, and the standardized treatment effect was Cohen's d=0.6. We also underscore that the 2-month time point in that project, which tested an integrated depression and obesity behavioral treatment, coincided with the start of weight loss treatment, which was sequenced after the first 5 PST sessions for depression, so measurements of neural targets at 2 months assessed the effect of PST alone. In ENGAGE, PST was delivered via in-person individual sessions with trained coaches.

   The SPEAC project is at the frontier of research on developing a voice-based AI agent (Lumen) for PST and validating its effect on *a priori* neural targets, and the Pilot RCT (Study 1) is to estimate possible effects of target engagement, which, if promising, will support further investigation to confirm target engagement in Study 2. In this context, we apply a precision interval approach (see Section A above) to determine Criterion 1. We define that, compared with the waitlist control group, the Lumen PST treatment group will demonstrate a *meaningful improvement* in a neural target if the standardized between-group mean difference is at least Cohen's d=0.3 in favor of Lumen (small effect; 50% of that in ENGAGE). At this effect size, the upper limit of the precision interval overlaps with d=0.8 (large effect) given a standardized half-width of 0.5 with 90% assurance that the interval contains the true mean difference based on the power analysis. We consider such a finding to be meaningful and will suffice Criterion 1.

ii. Criterion 2. Change in the level of task-evoked activation of a neural target that meets Criterion 1 correlates with temporally concurrent change from week 0 to 16 in self-reported measures within the corresponding PST theory-based construct for emotional reactivity or cognitive control. As noted, we consider correlations of r=0.4 or greater meaningful. Specifically:
   a. If the amygdala target exhibits a meaningful improvement in the Lumen group vs. the waitlist control (Criterion 1), then we will examine whether that improvement is significantly correlated with improvements in validated self-report measures of emotional reactivity (worry, affect) over the same 0-to-16-week interval.
   b. If the DLPFC target exhibits a meaningful improvement in the Lumen group vs. the waitlist control (Criterion 1), then we will examine whether that improvement is significantly correlated with improvements in cognitive control (problem solving, dysfunctional attitudes) over the same 0-to-16-week interval.

Criterion 1 is tested via *t* test that compares the change from week 0 to 16 in the Lumen group with that change in the waitlist control group. Participants are analyzed based on the group to which they are assigned using all available data. The *t* test described above can be readily expanded for exploratory subgroup analyses by applying to each subgroup (e.g., female vs. male).

To test Criterion 2, we will examine the regression change of the amygdala activity from week 0 to 16 on change in self-reported measures of emotional reactivity (worry, affect) from week 0 to 16, controlling for baseline values of the self-reported measures and group assignment. We expect that the coefficient for the change in amygdala activation is significant and positive in the direction of improvement in its corresponding, validated emotional reactivity self-report measures and that this relationship may be significantly modified by group assignment. Similarly, we will examine the regression change of the DLPFC activity from week 0 to 16 on change in self-reported measures of cognitive control (problem

solving, dysfunctional attitudes) from week 0 to 16, controlling for baseline values of the self-reported measures and group assignment. We expect that the coefficient for the change in DLPFC activation is significant and positive in the direction of improvement in its corresponding, cognitive control validated self-report measures and that this relationship may be significantly modified by group assignment.

B.1.a. Supportive Voxelwise Whole Brain Analysis

We will conduct an exploratory whole-brain analysis (1) to verify our *a priori* specification of neural targets, namely, amygdala for emotional reactivity and DLPFC for cognitive control, and (2) to identify any additional voxel clusters that are activated by our tasks and are associated with treatment response. This cluster-extent based thresholding analysis[84] will be carried out using the same SPM12 software used for neuroimage processing (see the *Study 1 Protocol Synopsis*, Forms E, which contains technical details). The identification of voxel clusters employs 2 criteria: (1) a primary threshold to identify activated voxels with a test level (e.g., $P \leq 0.001$), and (2) an extent threshold applied to clusters of contiguous voxels (neighbors at 5 of 6 voxel faces activated) based on an estimated distribution of cluster sizes, to control the family-wise error rate (FWER) for clusters (e.g., $\alpha\_FWER \leq 0.05$). Recently there has been concern that faulty statistical assumptions have led to high false-positive rates and over-large clusters that cross anatomical boundaries.[85-87] We follow the recommendations of Woo et al.[87] by setting $P \leq 0.001$ and employing nonparametric tests that avoid problematic Gaussian random field assumptions. This specification will provide sufficient sensitivity to allow detection of alternative anatomical clusters, which we can replicate using the same analysis strategy in Study 2.

B.2. Treatment effects on PROs and ecological measures

Secondary analyses will compare differences in discrete changes in PROs from week 0 to 16 and trajectories of change based on intensive time-series ecological measures between the Lumen and waitlist control groups. Analyses of the treatment effects on PROs will employ the same *t* test approach as described above for testing Criterion 1 for target engagement.

Longitudinal mixed modeling will be applied to analyses of the ecological measures (mood, stress, appraisal, and coping) where the design is Group×Time=2×8 as these measures will be obtained (using a validated measurement burst design)[88] daily for 7 days every 2 weeks (on weeks 0, 2, 4, 6, 8, 10, 12, 16). During active treatment in the Lumen group, this will occur for 3 days prior to, the day of, and 3 days after each scheduled PST session. For waitlist controls, this will occur for 7 days starting on Sunday of each assigned week. This allows us rich and detailed information on ambulatory changes over the course of treatment, as well as estimates of proximal responses to treatment within 3 days immediately post a PST session. For example, using the means of ecological measures over the 3 pre-PST days (to avoid confounding estimates with proximal treatment response) in the Lumen group and the first 3 days in the waitlist control group every 2 weeks as DVs, we will characterize their trajectory of change in terms of orthogonal linear, quadratic, and cubic components. We also will examine regression spline versions of these trajectories for visualization and interpretation. In a more exploratory vein, for each set of 7 daily ecological measures, collected biweekly, we will use piecewise linear models to characterize proximal responses to treatment based on the change of the 3 post-treatment days' means from the 3 pre-treatment days' means (omitting treatment day 4) in the Lumen group and the corresponding biweekly change of the last 3 days' means from the first 3 days' means (omitting day 4) in the waitlist control group. As these differences are nested within the 8 biweekly measurement bursts, and thus can be examined longitudinally, we can consider whether there is change over weeks in patients' proximal responses to each PST session (in terms of the ecological measures). In all analyses, all available data for each DV will be used and missing data will be handled directly through maximum-likelihood estimation via mixed modeling, supplemented with imputation and sensitivity analysis (see Handling Missing Data in Section B.4 below).

B.3. Dose effect analysis

B.3.a. Rationale for a standard dose

Participants randomized to the Lumen treatment group will complete 8 scheduled sessions (4 weekly and then 4 biweekly) over 12 weeks using secure study iPads. The essential research question under study is whether the AI agent Lumen as a new form of PST delivery can meaningfully engage *a priori* neural targets as putative mechanisms of treatment for adults with moderate depressive and/or anxiety symptoms. Fairly extensive evidence has been gathered to establish the optimal dose for implementation of PST in general medical and community settings; this typically involves 6-8 sessions that are initially weekly and taper off to every 2-4 weeks over 12 weeks.[89,90] Because this is a proof-of-concept project to develop an AI PST agent, we will design and train Lumen to emulate the standards developed for face-to-face delivery of one standardized, empirically supported dose. Dosing comparisons are thus made relative to appropriate implementation of PST and examination of differences in actual delivery (i.e., primarily number and length of sessions and, secondarily, indicators of intervention delivery fidelity).

B.3.b. Definitions of primary and secondary dose measures

Two primary dose measures will be the number of PST sessions and the length of PST sessions (in minutes) that participants complete with Lumen. Secondary dose measures will be each of the 2 primary measures weighted by the level of treatment fidelity. All Lumen-based PST sessions will be recorded and independently rated for fidelity by 2 PST master trainers using the validated 7-item PST Adherence and Competence Scale (PST-PAC).[53,54]

In psychotherapy outcome research, dose is typically defined as the number of sessions of therapy as it is a natural quantitative unit of treatment applicable across types of psychotherapy.[91-94] In this study, we also include the length of sessions as a measure of dose because (1) it is possible that the number of sessions varies little among participants given the brevity and controlled delivery of the therapy and (2) session length is an important parameter for possible optimization of Lumen to improve participant engagement and efficacy. Note that these dose measures are defined at any stage of the treatment; for example, after 4 sessions have been offered, a participant might have attended 3 sessions (number) for a total of 118 minutes (length). In other words, the dose measures are time-related variables.

Similarly, therapeutic fidelity is another important parameter for optimization in Lumen. In our view, it is a proxy to the strength of medicine and its use in this context adjusts the implied assumption that one session (or the same length of a session) equals one unit or dose, except that the strength (potency) of active ingredients in psychotherapy cannot be measured or standardized in the same way as that of a medication. Although the level of fidelity may be defined differently (e.g., with varying weights of the item scores), the ratio of raw total score to the maximum is simple and easy to interpret and there is no empirical support, to our knowledge, for an alternative. Alternative definitions may be explored using data from this study, especially if they would provide a more nuanced understanding of the need for optimization of Lumen. We will construct fidelity-weighted versions of the dose measures, number and length, in order to adjust doses for conformance to standard PST. Assessing dose in the multiple ways as we propose will allow us to determine which measure is most accurate and to be employed in Study 2. Also, our methodology and results will make an important contribution as assessing dose effects of novel digital health interventions on mechanisms such as brain targets is complex and understudied.

B.3.c. Dosing analyses

Dosing analyses will be conducted **within the Lumen treatment group**. The primary dosing analysis will examine the relationship of neural target activation in the amygdala and DLPFC to the 4 measures of dose (number and length of PST sessions, fidelity-weighted number and length of PST sessions). The secondary dosing analysis will examine the relationship of PROs, ecological end-of-day assessments, and measures of treatment progress (e.g., PHQ9 and GAD7 across PST sessions) to the same 4 measures of

dose. Note that given neural target activation measures and PROs are obtained at weeks 0 and 16, these analyses will focus on change over the full 16-week span of the study, whereas ecological end-of-day assessments and treatment progress measures are available for a total of 8 occasions each, and thus will be considered longitudinally.

Neural targets: The primary dosing analyses will consist in the separate regressions of change in activation from week 0 to 16 for the amygdala and DLPFC on each dose variable. This results in 8 separate regressions: amygdala on number of sessions, amygdala on length of sessions, amygdala on fidelity-weighted number, amygdala on fidelity-weighted length, and also DLPFC on the same 4 dose variables. Assumptions of linear regression will be verified and corrected through data transformation as needed. Our purpose is to identify variants of the dose variables with the strongest association (i.e., highest R^2) with change over 16 weeks in the 2 neural targets. If the coefficient for dose in these regressions is positive and significant, it confirms that the dose variable is associated with higher change in neural target activation.

Patient-reported outcomes: Depression and anxiety symptoms, functioning, and health-related quality of life will be measured at week 0 and 16. Thus, the secondary dose analysis for PROs mirrors the primary analysis for neural targets. That is, change from week 0 to 16 in the PROs will be separately regressed on the 4 dose variants, resulting in separate PRO-dose measure regressions. As in the primary analysis, our purpose is to identify variants of the dose variables with the strongest association (i.e., highest R^2) with change over 16 weeks in the PROs of interest.

Ecological end-of-day assessments: Ecological measures of mood, stress, appraisal, and coping will be available at weeks 0, 2, 4, 6, 8, 10, 12 and 16 (for a total of 8 occasions). Each week's assessment is based on 7 contiguous days' end-of-day assessments of the 4 ecological measures, with day 4 being the day of PST treatment in the Lumen group. Thus, the mean of the 3 pre-treatment days on these variables provides an indication of a participants' status immediately prior to each PST session; the mean of the 3 post-treatment days on the same variables provides an indication of a participants' status immediately post the session. For this secondary dosing analysis, we can mirror the dosing analysis by regressing change in the 7-day mean from week 0 to week 16 for each of the 4 ecological measures on the 4 dose variants. Importantly, there are 6 intermediate measurement occasions in addition to week 16, so we will be able to examine trajectories of change from week 0 to week 2, to week 4, and so on, to week 16. Moreover, for each ecological measure a set of time-related variables may be constructed, such as overall 7 days' means, 3 pre-treatment days' means, and differences between 3 pre-treatment days' means and 3 post-treatment days' means (omitting treatment day 4), so we will be able to examine trajectories of change in these variables. It is also important to note that the 4 dosing variables in these longitudinal regression models reflect the dose received up to that point in the treatment. This provides a view of how accurately the dosing variables perform for the ecological measures over the entire span of the treatment. These analyses will use time-varying mixed models where the DV (e.g., mood) and the dose measure vary over the 16-week period.[95-97] The models for each dose measure (e.g., number of sessions) will include separate polynomial terms for linear, quadratic, and cubic rates of change in the DV over time. Alternatively, we will model the log of the dose measure (e.g., number of sessions) considering that some studies of in-person psychotherapy reported a negatively accelerated relationship between dose and treatment progress, suggesting a pattern of successively diminishing returns with increasing dose.[91,98,99] We will compare model fit for the polynomial and logarithmic models using Schwartz Bayesian Information Criterion (BIC).

Treatment progress measures: To monitor treatment progress, symptoms (PHQ9, GAD7) and treatment acceptability (usability, user experience, and treatment alliance) will be measured in each PST session at weeks 1, 2, 3, 4, 6, 8, 10, and 12 (for a total of 8 occasions). Similar to the analyses of ecological measures as described above, we will implement polynomial and logarithmic models and compare model

fit using BIC. These regressions' R^2's will allow us to examine in detail the relationship of treatment progress and cumulative dose, with fidelity-to-PST taken into account using the fidelity-weighted number and length of PST sessions. Additionally, as is common in psychotherapy in general, it is possible that not all participants will complete the full 8 sessions as some may drop out early owing to either lack of improvement or "enough" improvement based on self-assessment or yet other reasons related or unrelated to the treatment. Using the Good-Enough-Level (GEL) model, some studies of in-person psychotherapy showed that the rate of change varied as a function of treatment duration, namely, the rate of change (progress) was faster for patients attending fewer sessions but slower for those attending more sessions.[99-101] In other words, patients may remain in treatment until they improve. To explore whether a similar pattern is present for Lumen PST, we will apply the GEL modeling approach by extending the polynomial models for treatment progress measures, in particular, PHQ9 and GAD7, to include a fixed effect of total treatment dose (e.g., total number of sessions) and interactions between total treatment dose and linear, quadratic, and cubic rates of change. Essentially, the GEL model is a stratified analysis that examines dose effects across strata of participants with similar treatment dose (e.g., similar number of sessions attended), as opposed to aggregate dose effects across all participants.
Furthermore, we will conduct exploratory analyses of baseline characteristics in relation to the dose effects on neural targets, PROs, ecological measures, and treatment progress measures. We anticipate that baseline characteristics (e.g., severity at presentation, sex as a biological variable, sociodemographics) may be related to dosing estimates (e.g., via actual sessions completed). To explore these relationships, baseline characteristics and their interactions with dose measures will be added to the models as described above.

B.4. Intention to Treat (ITT) and Handling Missing Data
Clinical trials such as the one proposed must face the issues of dropout and missing data. We have put in place a strong plan to ensure maximum retention of participants through the course of the study, with a projected 85% or greater retention (see the *Recruitment and Retention Plan*). Data missing completely at random (MCAR) may be analyzed "as is" without imputation, as it does not introduce bias into estimates and inferences. More commonly, data are missing at random (MAR) owing to covariate-dependent events (e.g., irregular work shifts, children's illness, transportation, and scheduling foul-ups). Our primary analysis technique, linear mixed models, is robust for MAR and provides unbiased estimates of model parameters without imputation. (One may impute as well, but the benefit is usually modestly improved power rather than changes to conclusions.) The more serious concern for behavioral interventions is the possibility that data will be missing not at random (MNAR). This occurs when dropout or missingness is related to the level of the outcome of interest (e.g., a participant's level of depression or anxiety). If treatment is working well and symptoms of depression and anxiety have lessened, then participants may feel continuation in treatment is not necessary or beneficial. They drop out or miss sessions because the treatment seems to have worked. On the other hand, if participants experience no improvement in symptoms, they may drop out or miss sessions because they are discouraged that the treatment seems not to have worked. In either case, the loss of data is related to the outcomes under study and has the potential to bias the study's conclusion about the intervention. Handling the problem of MNAR is difficult. First, there is no test for the presence of MNAR (*vs.* MAR). Second, there are many possible underlying sources of MNAR and there is no guarantee that explorations of missing data patterns will clearly reveal those sources. Third, missing data imputation for MNAR necessarily makes strong and untestable statistical assumptions, so that conclusions from imputed MNAR datasets must remain tentative. This picture reinforces our strong commitment to participant retention (see the Study 1 *Recruitment and Retention Plan*).

Missing data imputation implemented via SAS procedures MI, MIANALYZE, and MCMC will be carried out to reexamine all primary and secondary analysis.[102,103] This effectively implements ITT as traditionally conducted. Specifically, both missing outcomes and covariates will be imputed under the

assumption that data are MAR. The fully conditional specification (FCS) approach allows each variable type to be imputed with an appropriate model: linear, discriminant, logistic, or Poisson regression.[104-108] The algorithm for FCS is implemented in procedure MI: 100 imputed data sets will be created, each imputation is analyzed using MIXED, and finally MIANALYZE combines 100 sets into final estimates. This approach to MAR has been widely adopted given its robust performance.[109] Because linear mixed models estimated via direct likelihood as in MIXED offer unbiased estimates under MAR, the benefit of multiple imputation lies in compliance with the ITT fill-in requirement and modestly increased power. We anticipate these results to differ little from those provided by the primary analysis.

B.4.a. Sensitivity of results to missing data assumptions

These will be conducted by varying the models used for imputation.[110] There is no test for NMAR, but we will examine missing data frequency by condition and in covariate subgroups to detect and diagnose differential dropout. If warranted, we can shift to a Bayesian joint modeling approach to imputation using MCMC. The mathematical statistical underpinnings of sensitivity analysis for missing data imputation have advanced substantially (see, for example, the treatment of selection models in Chapter 8 of Daniels and Hogan[111]). However, the software implementation of these techniques has lagged behind the theory. Simpler approaches have been implemented in SAS at this time; however, the more complex techniques involving semiparametric specification of the Daniels and Hogan extrapolation model are not readily available. During the data collection years, this problem will likely be resolved. Here, we mention 3 approaches that seem feasible now: (1) With just a few data collection waves, the number of frequently occurring missing data patterns is likely 3 to 6 for neural targets and PROs and 8 to 10 for ecological assessments and treatment progress measures (some infrequent patterns can be grouped together); these patterns can be used to specify a pattern mixture model for the missing data. Such models merely include an indicator variable for each pattern and thereby provide a straightforward means to adjust results for missing data.[112] If results are not much altered by inclusion of pattern indicators, then concern about missing data assumptions is diminished; (2) Mallinckrodt et al.[113-115] reviewed simulation and other studies of the impact of MNAR on trial results (primary analyses like those we plan) and concluded that the impact of violations is rarely substantial. The R statistical environments add-on package ‘mice’ now has capabilities to create particular violations of MAR and MNAR assumptions, so that simulation-based exploration of violations using our own data will now be possible; and (3) SAS PROC MI has now incorporated an option to use a SCALE parameter (suggested range between 1.5 and 2.0) to explore the sensitivity of results to imputations relying on the MAR assumptions.

Appendix 4. Study Event Windows

Event windows (in Days)

Preferred windows (in Days)	ES	V1	EMA#1	R														V2
Lumen PST:					Orient.	S1	S2	S3	S4		S5		S6		S7		S8		
Week (indexed from randomization) #: from	-4	-2		0	1	1	2	3	4		6		8		10		12		**14**
to	-2	-1			2	3	4	5	6		8		10		12		14		**18**
IES-eligible to eConsent, max. days	14			Randomization															
eConsent to V1, max. days		14																	
IES-eligible to V1 (rescreen if IES-V1>90d), max. days	28																		
V1 to EMA#1 start (automated), max. days		**1**																	
V1 to Randomization, max. days		**10**																	
Randomization to Lumen Orientation Visit, max. days					14														
Lumen Orientation to Session 1, max. days						7													
Lumen Session to Session, max. days							7	7	7	14		14		14		14			
Randomization to V2, target 112d (16 weeks) +/- allowable 14d window (14-18 weeks)																			112 +/-14

Abbreviations: IES, initial eligibility screening; EMA, ecological daily assessment; PST, problem solving treatment; S1-S8, sessions 1-8 with Lumen; V1, visit 1; V2, visit 2

13.0 References

1. Roehrig C. Mental disorders top the list of the most costly conditions in the United States: $201 billion. *Health affairs (Project Hope).* 2016;35(6):1130-1135.
2. Greenberg PE, Fournier AA, Sisitsky T, Pike CT, Kessler RC. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *The Journal of clinical psychiatry.* 2015;76(2):155-162.
3. National Institute of Mental Health. Major Depression Among Adults. 2015; *https://www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adults.shtml*. Accessed October 11, 2017.
4. National Institute of Mental Health. Any anxiety disorder among adults. 2015; *https://www.nimh.nih.gov/health/statistics/prevalence/any-anxiety-disorder-among-adults.shtml*. Accessed November 21, 2017.
5. Anxiety and Depression Association of America. Facts & statistics. 2017; *https://adaa.org/about-adaa/press-room/facts-statistics*. Accessed November 21, 2017.
6. Rodriguez MR, Nuevo R, Chatterji S, Ayuso-Mateos JL. Definitions and factors associated with subthreshold depressive conditions: a systematic review. *BMC Psychiatry.* 2012;12:181.
7. Meeks TW, Vahia IV, Lavretsky H, Kulkarni G, Jeste DV. A tune in "a minor" can "b major": a review of epidemiology, illness course, and public health implications of subthreshold depression in older adults. *Journal of affective disorders.* 2011;129(1-3):126-142.
8. Haller H, Cramer H, Lauche R, Gass F, Dobos GJ. The prevalence and burden of subthreshold generalized anxiety disorder: a systematic review. *BMC Psychiatry.* 2014;14(1):128.
9. Gonzalez HM, Vega WA, Williams DR, Tarraf W, West BT, Neighbors HW. Depression care in the United States: too little for too few. *Arch Gen Psychiatry.* 2010;67(1):37-46.
10. Kessler RC, Berglund P, Demler O, et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Jama.* 2003;289(23):3095-3105.
11. Cook BL, Trinh NH, Li Z, Hou SS, Progovac AM. Trends in racial-ethnic disparities in access to mental health care, 2004-2012. *Psychiatric services (Washington, DC).* 2017;68(1):9-16.
12. van Schaik DJ, Klijn AF, van Hout HP, et al. Patients' preferences in the treatment of depressive disorder in primary care. *General hospital psychiatry.* 2004;26(3):184-189.
13. Regier DA, Narrow WE, Rae DS, Manderscheid RW, Locke BZ, Goodwin FK. The de facto US mental and addictive disorders service system. Epidemiologic catchment area prospective 1-year prevalence rates of disorders and services. *Arch Gen Psychiatry.* 1993;50(2):85-94.
14. McHugh RK, Whitton SW, Peckham AD, Welge JA, Otto MW. Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: a meta-analytic review. *The Journal of clinical psychiatry.* 2013;74(6):595-602.
15. Hirschfeld RM, Keller MB, Panico S, et al. The National Depressive and Manic-Depressive Association consensus statement on the undertreatment of depression. *Jama.* 1997;277(4):333-340.

16. Fortney JC, Harman JS, Xu S, Dong F. The association between rural residence and the use, type, and quality of depression care. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association.* 2010;26(3):205-213.
17. Sorkin DH, Murphy M, Nguyen H, Biegler KA. Barriers to Mental Health Care for an Ethnically and Racially Diverse Sample of Older Adults. *Journal of the American Geriatrics Society.* 2016;64(10):2138-2143.
18. Leong FT, Kalibatseva Z. Cross-cultural barriers to mental health services in the United States. *Cerebrum : the Dana forum on brain science.* 2011;2011:5.
19. Proudfoot J, Parker G, Hadzi Pavlovic D, Manicavasagar V, Adler E, Whitton A. Community attitudes to the appropriation of mobile phones for monitoring and managing depression, anxiety, and stress. *Journal of Medical Internet Research.* 2010;12(5):e64.
20. DeAndrea DC. Testing the proclaimed affordances of online support groups in a nationally representative sample of adults seeking mental health assistance. *Journal of Health Communication.* 2015;20(2):147-156.
21. Cuijpers P, Marks IM, van Straten A, Cavanagh K, Gega L, Andersson G. Computer-aided psychotherapy for anxiety disorders: a meta-analytic review. *Cognitive behaviour therapy.* 2009;38(2):66-82.
22. Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. *PloS one.* 2010;5(10):e13196.
23. National Institute for Health and Care Excellence. Computerised cognitive behaviour therapy for depression and anxiety. 2013; *https://www.nice.org.uk/guidance/ta97/chapter/1-Guidance*. Accessed April 8, 2018.
24. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry.* 2017;16(3):287-298.
25. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *Journal of affective disorders.* 2017;218:15-22.
26. Bohn D. *Amazon says 100 million Alexa devices have been sold — what's next?* : The Verge;2019.
27. Kinsella B. RBC Analyst Says 52 Million Google Home Devices Sold to Date and Generating $3.4 Billion in 2018 Revenue. 2018; *https://voicebot.ai/2018/12/24/rbc-analyst-says-52-million-google-home-devices-sold-to-date-and-generating-3-4-billion-in-2018-revenue/*.
28. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine.* 2019;25(1):44-56.
29. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* First ed: Basic Books; 2019.
30. Lucas GM, Gratch J, King A, Morency L-P. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior.* 2014;37:94-100.
31. Nass C, Moon Y. Machines and mindlessness: social responses to computers. *Journal of Social Issues.* 2000;56:81-103.

32. Nass C, Steuer J, Tauber ER. Computers are social actors. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 1994; Boston, Massachusetts, USA.
33. Ho A, Hancock J, Miner AS. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *The Journal of communication.* 2018;68(4):712-733.
34. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient education and counseling.* 2005;59(1):21-30.
35. Bickmore TW, Caruso L, Clough-Gorr K, Heeren T. 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers.* 2005;17(6):711–735.
36. Bickmore TW, Pfeifer LM, Byron D, et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. *Journal of Health Communication.* 2010;15 Suppl 2:197-210.
37. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association : JAMIA.* 2018;25(9):1248-1258.
38. Ring L, Bickmore T, Pedrelli P. An affectively aware virtual therapist for depression counseling. 2017; *http://mentalhealth.media.mit.edu/wp-content/uploads/sites/46/2016/04/CHI2016-MentalHealth-lring.pdf*. Accessed October 17, 2017.
39. Bickmore T, Gruber A. Relational agents in clinical psychiatry. *Harvard review of psychiatry.* 2010;18(2):119-130.
40. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health.* 2017;4(2):e19.
41. Pereira J, Diaz O. Using health chatbots for behavior change: a mapping study. *Journal of medical systems.* 2019;43(5):135.
42. National Institute of Mental Health. Strategic research priorities overview. 2019; *https://www.nimh.nih.gov/about/strategic-planning-reports/strategic-research-priorities/index.shtml*. Accessed September 6, 2019.
43. National Institute of Mental Health. Opportunities and challenges of developing information technologies on behavioral and social science clinical research. Year N/A; *https://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/reports/opportunities-and-challenges-of-developing-information-technologies-on-behavioral-and-social-science-clinical-research.shtml*. Accessed September 6, 2019.
44. Hart SG, Stavenland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock P, Meshkati N, eds. *Human Mental Workload.* Amsterdam, The Netherlands: Elsevier; 1988.
45. Schrepp M, Hinderks A, Thomaschewski Jr. Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). *Int J Interact Multimed Artif Intell.* 2017;4(Regular Issue):103-108.
46. Hatcher RL, Gillaspy JA. Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research.* 2006;16(1):12-25.

47. Munder T, Wilmers F, Leonhart R, Linster Hans W, Barth J. Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy.* 2009;17(3):231-239.
48. Kiluk BD, Serafini K, Frankforter T, Nich C, Carroll KM. Only Connect: the working alliance in computer-based cognitive behavioral therapy. *Behaviour research and therapy.* 2014;63:139-146.
49. Bickmore TW, Mitchell SE, Jack BW, Paasche-Orlow MK, Pfeifer LM, Odonnell J. Response to a relational agent by hospital patients with depressive symptoms. *Interacting with computers.* 2010;22(4):289-298.
50. Carroll JM, Rosson MB. Getting around the task-artifact cycle: how to make claims and design by scenario. *ACM Trans Inf Syst.* 1992;10(2):181-212.
51. van der Vaart R, Drossaert C. Development of the Digital Health Literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 Skills. *Journal of Medical Internet Research.* 2017;19(1):e27.
52. Polson PG, Lewis C, Rieman J, Wharton C. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies.* 1992;36(5):741-773.
53. Oxman TE, Hegel MT, Hull JG, Dietrich AJ. Problem-solving treatment and coping styles in primary care for minor depression. *Journal of consulting and clinical psychology.* 2008;76(6):933-943.
54. Hegel MT, Dietrich AJ, Seville JL, Jordan CB. Training residents in problem-solving treatment of depression: a pilot feasibility and impact study. *Family medicine.* 2004;36(3):204-208.
55. Xiao L, Huang Q, Yank V, Ma J. An easily accessible web-based minimization random allocation system for clinical trials. *Journal of Medical Internet Research.* 2013;15(7):e139.
56. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics.* 1975;31(1):103-115.
57. Therneau TM. How many stratification factors are "too many" to use in a randomization plan? *Controlled clinical trials.* 1993;14(2):98-108.
58. Efron B. Forcing a sequential experiment to be balanced. *Biometrika.* 1971;58(3):403-417.
59. Kuznetsova OM, Tymofyeyev Y. Preserving the allocation ratio at every allocation with biased coin randomization and minimization in studies with unequal allocation. *Statistics in medicine.* 2012;31(8):701-723.
60. Lv N, Ajilore OA, Ronneberg CR, et al. The ENGAGE-2 study: Engaging self-regulation targets to understand the mechanisms of behavior change and improve mood and weight outcomes in a randomized controlled trial (Phase 2). *Contemp Clin Trials.* 2020:106072.
61. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qualitative Health Research.* 2005;15(9):1277-1288.
62. Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State Worry Questionnaire. *Behav Res Ther.* 1990;28(6):487-495.
63. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology.* 1988;54(6):1063-1070.

64. D'Zurilla T, Nezu A, Maydeu-Olivares A. *Manual for the Social Problem-Solving Inventory-Revised.* North Tonawanda, NY: Multi-Health Systems; 2002.
65. Weissman AN. *The Dysfunctional Attitude Scale: A Validation Study*. Publicly Accessible Penn Dissertations. 1182., University of Pennsylvania; 1979.
66. Snaith R, Zigmond A. *The Hospital Anxiety and Depression Scale manual.* Windsor, Berkshire (UK): Nfer-Nelson; 1994.
67. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta psychiatrica Scandinavica.* 1983;67(6):361-370.
68. Sheehan KH, Sheehan DV. Assessing treatment effects in clinical trials with the discan metric of the Sheehan Disability Scale. *International clinical psychopharmacology.* 2008;23(2):70-83.
69. Kessler RC, Barber C, Beck A, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). *Journal of occupational and environmental medicine.* 2003;45(2):156-174.
70. Kessler RC, Ames M, Hymel PA, et al. Using the World Health Organization Health and Work Performance Questionnaire (HPQ) to evaluate the indirect workplace costs of illness. *Journal of occupational and environmental medicine.* 2004;46(6 Suppl):S23-37.
71. Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Medical care.* 1996;34(3):220-233.
72. Grasso DJ, Briggs-Gowan MJ, Ford JD, Carter AS. The Epidemic – Pandemic Impacts Inventory (EPII). 2020; *https://health.uconn.edu/psychiatry/wp-content/uploads/sites/51/2020/05/EPII-Main-V1.pdf*. Accessed March 24, 2021.
73. Koenig HG, Büssing A. The Duke University Religion Index (DUREL): A Five-Item Measure for Use in Epidemological Studies. *Religions.* 2010;1(1):78-85.
74. Berkman LF, Syme SL. Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. *Am J Epidemiol.* 1979;109(2):186-204.
75. Ford ES, Loucks EB, Berkman LF. Social integration and concentrations of C-reactive protein among US adults. *Ann Epidemiol.* 2006;16(2):78-84.
76. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A Short Scale for Measuring Loneliness in Large Surveys: Results From Two Population-Based Studies. *Res Aging.* 2004;26(6):655-672.
77. Julious SA, Patterson SD. Sample sizes for estimation in clinical research. *Pharmaceutical Statistics.* 2004;3(3):213-215.
78. Ma J, Strub P, Lv N, et al. Pilot randomised trial of a healthy eating behavioural intervention in uncontrolled asthma. *The European Respiratory Journal.* 2016;47(1):122-132.
79. Sheiner LB. Learning versus confirming in clinical drug development. *Clinical pharmacology and therapeutics.* 1997;61(3):275-291.
80. Julious SA. Sample sizes for clinical trials with Normal data. *Statistics in medicine.* 2004;23(12):1921-1986.
81. Feise RJ. Do multiple outcome measures require p-value adjustment? *BMC medical research methodology.* 2002;2:8.
82. Althouse AD. Adjust for multiple comparisons? It's not that simple. *The Annals of thoracic surgery.* 2016;101(5):1644-1645.

83. Warmerdam L, van Straten A, Jongsma J, Twisk J, Cuijpers P. Online cognitive behavioral therapy and problem-solving therapy for depressive symptoms: Exploring mechanisms of change. *Journal of behavior therapy and experimental psychiatry.* 2010;41(1):64-70.
84. Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic resonance in medicine.* 1995;33(5):636-647.
85. Carter CS, Lesh TA, Barch DM. Thresholds, power, and sample sizes in clinical neuroimaging. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging.* 2016;1(2):99-100.
86. Eklund A, Nichols TE, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America.* 2016;113(28):7900-7905.
87. Woo CW, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage.* 2014;91:412-419.
88. Sliwinski MJ. Measurement-burst designs for social health research. *Social and Personality Psychology Compass.* 2008;2(1):245-261.
89. Hegel M, Barrett J, Oxman T, Mynors-Wallis L, Gath D. *Problem-Solving Treatment for Primary Care (PST-PC): A Treatment Manual for Depression.* Hanover, New Hampshire: Dartmouth University; 1999.
90. Cape J, Whittington C, Buszewicz M, Wallace P, Underwood L. Brief psychological therapies for anxiety and depression in primary care: meta-analysis and meta-regression. *BMC medicine.* 2010;8:38.
91. Howard KI, Kopta SM, Krause MS, Orlinsky DE. The dose-effect relationship in psychotherapy. *The American psychologist.* 1986;41(2):159-164.
92. Hansen NB, Lambert MJ, Forman EM. The psychotherapy dose-response effect and its implications for treatment delivery services. *Clin Psychol Sci Prac.* 2002;9(3):329-343.
93. Ginestet CE, Emsley R, Landau S. Dose-response modeling in mental health using stein-like estimators with instrumental variables. *Statistics in medicine.* 2017;36(11):1696-1714.
94. Robinson L, Delgadillo J, Kellett S. The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy research : journal of the Society for Psychotherapy Research.* 2019:1-18.
95. Diggle P, Heagerty, P., Liang, K.-Y. and Zeger, S. *Analysis of Longitudinal Data.* 2nd ed. Oxford, UK: Oxford University Press; 2002.
96. Fitzmaurice GM, Laird, N.M. and Ware, J.H. *Applied longitudinal analysis.* 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2011.
97. Hedeker D, Gibbons R. *Longitudinal data analysis.* Hoboken, NJ: Wiley-Interscience; 2006.
98. Baldwin SA, Berkeljon A, Atkins DC, Olsen JA, Nielsen SL. Rates of change in naturalistic psychotherapy: contrasting dose-effect and good-enough level models of change. *J Consult Clin Psychol.* 2009;77(2):203-211.
99. Stulz N, Lutz W, Kopta SM, Minami T, Saunders SM. Dose-effect relationship in routine outpatient psychotherapy: does treatment duration matter? *Journal of counseling psychology.* 2013;60(4):593-600.

100. Owen JJ, Adelson J, Budge S, Kopta SM, Reese RJ. Good-enough level and dose-effect models: Variation among outcomes and therapists. *Psychotherapy Research.* 2016;26(1):22-30.
101. Falkenstrom F, Josefsson A, Berggren T, Holmqvist R. How much therapy is enough? Comparing dose-effect and good-enough models in two different settings. *Psychotherapy (Chic).* 2016;53(1):130-139.
102. Sas Institute I. *SAS/STATA 14.1 Users Guide.* SAS Institute, Inc; 2015.
103. Berglund P, Heeringa SG. *Multiple imputation of missing data using SAS.* Cary, NC: SAS Institute, Inc; 2014.
104. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research.* 2007;16(3):219-242.
105. van Buuren S. *Flexible Imputation of Missing Data.* Boca Raton, FL: Chapman and Hall/CRC 2012.
106. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research.* 2011;20(1):40-49.
107. Ibrahim JG. Short course in missing data methods in regression. In:2015.
108. Ibrahim JG. Short course on missing data methods in regression, Parts I and II. In:2016.
109. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation.* 2006;76(12):1049-1064.
110. Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics.* 1996;52(4):1324-1333.
111. Daniels MJ, Hogan JW. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis.* Vol 109: Chapman & Hall/CRC (Taylor & Francis Group); 2008.
112. Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods.* 1997;2(1):64-78.
113. Mallinckrodt CH. *Preventing and Treating Missing Data in Longitudinal Clinical Trials: A Practical Guide.* Cambridge University Press; 2013.
114. Mallinckrodt CH, Lin Q, Lipkovich I, Molenberghs G. A structured approach to choosing estimands and estimators in longitudinal clinical trials. *Pharmaceutical Statistics.* 2012;11:456-461.
115. Mallinckrodt CH, Roger J, Chuang-Stein C, Others. Missing data: Turning guidance into action. *Statistics in Biopharmaceutical Research.* 2013;5(4):369-382.