A Phase-2 RCT of a Voice-based AI Coach for Depression and Anxiety

[1,2]Thomas Kannampallil*, PhD, [3]Olusola A. Ajilore*, MD, PhD, [4]Joshua M. Smyth, PhD, [5]Amruta Barve, MPH, [5]Corina R. Ronneberg, PhD, [5]Nan Lv, PhD, [5]Vikas Kumar, MS, [5]Claudia Garcia, BS, [5]Gbenga Aborisade, BS, [5]Nancy E. Wittels, MS, [5]Zhengxin Tang, BS, [5]Bharathi Chinnakotla, BS, [6]Lan Xiao, PhD, [5]Jun Ma**, MD, PhD

[1]Department of Anesthesiology
[2]Institute for Informatics, Data Science, and Biostatistics
Washington University School of Medicine, St Louis, MO

[3]Department of Psychiatry
University of Illinois at Chicago, Chicago, IL

[4]Department of Psychology
The Ohio State University, Columbus, OH

[5]Department of Medicine
University of Illinois at Chicago, Chicago, IL

[6]Department of Epidemiology and Population Health
Stanford University

*Equal first authors

**Corresponding Author:
Jun Ma, MD, PhD
Beth and George Vitoux Professor of Medicine
Founding director, Vitoux Program on Aging and Prevention
Department of Medicine
University of Illinois at Chicago
1747 W. Roosevelt Rd, Room 586 (MC 275)
Chicago, IL 60608
Email: maj2015@uic.edu
TEL: (312) 413-9830

Number of Tables: 3

Number of Figures: 1

No. of Words (Abstract): 297

No. of Words (Manuscript): 3659

Keywords: problem-solving therapy, digital mental health, artificial intelligence, voice-based coach, depression, anxiety, psychological distress

Research in Context

Evidence before this study

The evidence base for artificial intelligence (AI)-based conversational agents (or "chatbots") for depression and anxiety remains limited. Recent systematic reviews and meta-analyses of randomized clinical trials have generally found these interventions to be modestly effective, although the overall certainty of evidence is low due to substantial heterogeneity and methodological limitations across studies. Moreover, most prior trials have evaluated text-based chatbot interactions, whereas voice-based AI approaches capable of supporting spoken exchanges resembling clinician-patient conversational interactions during therapy remain understudied. For this study, we specifically searched PubMed for randomized clinical trials evaluating voice-based AI-based chatbots for depression and anxiety for the period 2014 to May 13, 2026, without language restrictions, using the following search term: (voice OR vocal OR speak) AND (chatbot OR coach OR agent OR "artificial intelligence" OR "AI") AND (depression OR depressive OR anxiety). This search identified 24 articles, and after initial screening of abstracts, 3 studies were identified, including the phase 1 pilot trial for this study.

Added value of this study

To our knowledge, this is the first 3-arm randomized clinical trial evaluating a rule-based AI coach ("Lumen") delivering problem-solving therapy (PST), a brief cognition-focused psychotherapy, using a commercial voice-based conversational platform. A total of 200 participants with untreated, clinically significant depression and/or anxiety were randomized to Lumen-coached PST, human-coached PST, or waitlist control. Although Lumen did not significantly alter the primary neural target related to cognitive control, secondary measures of patient-reported problem-solving ability and psychological distress improved significantly over 18 weeks, compared with waitlist control. Exploratory noninferiority tests showed that treatment effects of Lumen- and human-coached PST did not differ significantly on these outcomes. Improved problem-solving mediated reduced psychological distress for both modalities, supporting theory-based mechanism of PST.

Implications of the available evidence

Consistent with emerging evidence supporting the therapeutic potential of AI interventions for depression and anxiety, the present findings suggest that voice-based, AI-driven PST may benefit adults with clinically significant depression and/or anxiety who are not receiving care. Further adequately powered trials are needed to confirm clinical noninferiority to human-coached PST, determine the durability of treatment effects, and clarify the neural and behavioral mechanisms underlying treatment response.

Abstract

Background: Clinical evidence regarding artificial intelligence (AI) mental health interventions remains limited. This phase 2 trial investigated the mechanisms and efficacy of a rule-based AI coach, Lumen, delivering problem-solving therapy (PST) via voice for adults with clinically significant symptoms of depression and/or anxiety.

Methods: Participants were randomized to Lumen (n=100), human-coached PST (n=50) or waitlist control (n=50) for 18 weeks. PST was delivered by Lumen on Amazon's Alexa platform via voice or a human coach via videoconferencing in 4-weekly and then 4-biweekly sessions. Change in activation of the right dorsolateral prefrontal cortex (dlPFC) for cognitive control using functional neuroimaging was the primary mechanistic target measure. Patient-reported measures included changes in behavior associated with cognitive control (Social Problem-solving Index-Revised Short Form) and in clinical symptoms (Hospital Anxiety and Depression Scale). Statistical analyses used t-tests and ordinary least square regression

Findings: Participants had a mean age of 36.6 years (SD=11.9), and were 77% women, 25% Black, 29% Latino, and 21% Asian. At 18 weeks, change from baseline in right dlPFC activity did not differ significantly by treatment arm. Compared with waitlist control, Lumen-coached participants had significantly greater improvements from baseline to 18 weeks in overall problem-solving ability (between-group mean difference=1.04, 95%CI [0.23, 1.84]) and in symptoms of psychological distress (between-group mean difference=-3.56, 95%CI [-5.69, -1.43]) due to depression and anxiety. Lumen- and human-coached PST did not differ significantly for any of these measures, and improved problem-solving ability mediated reductions in psychological distress for both modalities. One serious adverse event involving hospitalization, unrelated to the study, was detected.

Interpretation: A rule-based AI coach delivering PST via voice may improve problem-solving abilities and clinical symptoms among adults with clinically significant depression and anxiety. However, these findings are preliminary; further research is warranted to confirm clinical efficacy and to elucidate neural mechanisms.

Funding: R33MH119237

Introduction

Depression and anxiety are among the most common mental health disorders in the United States[1] and among the leading causes of disability worldwide.[2] These conditions are often comorbid, and fewer than half of those reporting symptoms receive treatment,[3] owing to a myriad structural and attitudinal factors.[4] Research has long established that people prefer psychotherapies over medications;[5] however, use of proven psychotherapies is limited by clinician shortages, stigma, cost, and access barriers.[6,7]

Digital mental health tools, ranging from mobile applications to artificial intelligence (AI) interventions, offer alternative modalities for delivering evidence-based psychotherapies at scale. AI-based chatbots have the potential to replicate human-like conversations in a variety of settings. Chatbot-based mental health interventions fall under two broad categories: rule-based applications (relying on structured treatment frameworks for intent detection and therapy content delivery) and large language model (LLM)-based applications (relying on generative frameworks for open-ended dialog and content).[8] Evidence on their promise in mitigating depressive and anxiety symptoms continues to emerge;[9,10] however, systematic reviews have reported mixed results.[8,11] A recent meta-analysis of chatbot-based intervention studies reported that rule-based mental health interventions achieved a significant modest effect on improving depressive symptoms, but not on anxiety symptoms. In contrast, LLM-based interventions showed nonsignificant findings for both depression and anxiety, although the effect sizes did not differ significantly between rule- and LLM-based interventions.[8]

Compared with chatbots relying on text-based interactions, voice-based interactions allow for natural, human-like conversations resembling clinician-patient interactions during therapy.[12,13] However, clinical research on voice-based AI interventions is still nascent.[14-16]

Developed using an iterative user-centered design process,[17,18] Lumen is a voice-based AI coach relying on rule-based natural language processing (NLP) techniques for delivering the clinically proven problem-solving therapy (PST),[19] a cognition-targeted brief psychotherapy, for depression and anxiety. In a phase 1 randomized trial among adults with untreated, clinically significant symptoms of depression and/or anxiety, compared with waitlist control, Lumen-coached PST resulted in meaningful engagement of an a priori neural mechanistic target for cognitive control, right dorsolateral prefrontal cortex (dlPFC), and symptom improvements meeting the prespecified "go criteria" for additional testing.[20]

Aligned with the experimental therapeutics approach prioritizing mechanism-targeted intervention development and testing,[21] the primary aim of this larger, phase 2, 3-arm randomized trial[22] was to confirm that Lumen-coached PST was better (superior to) than waitlist control on target engagement of the right dlPFC for cognitive control. Secondary aims were to investigate (a) whether Lumen-coached PST was not worse than (noninferior to) human-coached PST on right dlPFC engagement; (b) the effects of Lumen-coached PST compared with waitlist control and human-coached PST on additional a priori neural and patient-reported measures of target engagement and on patient-reported symptom and functional outcomes; and (c) the mediating effects of target engagement on outcomes.

Method

Participants

The trial protocol was published.[22] Participants were recruited between January 23, 2023, and December 22, 2024, from outpatient clinics at University of Illinois Hospital and Health Sciences System and employee email listservs at the University of Illinois Chicago (Suppl, Section J).

Adults aged ≥18 reporting clinically significant symptoms of depression and/or anxiety, defined as moderate or moderately-severe depression (Patient Health Questionnaire [PHQ-9] scores 10–19)[23] and/or moderate anxiety (Generalized Anxiety Disorder [GAD-7] scores 10–14),[24] were included. Exclusion criteria included serious medical or psychiatric comorbidities, suicidality, other contraindications or limiting conditions (Suppl, Section A).

Randomization and masking

Participants were randomized to Lumen- or human-coached PST or to the waitlist control arm, using a validated online system[25] based on Pocock's covariate-adaptive minimization.[26] This approach allowed for achieving a better-than-chance marginal balance among the study arms across multiple baseline characteristics, including sex, age, race/ethnicity, digital health literacy,[27] PHQ-9 score, and GAD-7 score. Investigators, safety monitor, outcome assessors, and the blinded data analyst were masked to participants' treatment assignment.

Interventions to be tested

Lumen-coached PST

Developed on Amazon's Alexa-based AI platform, Lumen provides 8 PST sessions (4 weekly, then 4 biweekly) for treating depression and/or anxiety. Luman-coached PST is patient-driven, where Lumen acts as a guide to identify a problem, set a goal, brainstorm and compare solutions, choose a solution, develop an action plan, and to implement and evaluate the plan in each session.[19] These attributes—inherent structure and pragmaticism—make PST an ideal intervention for delivery using a rule-based AI application via a voice-based modality.

Lumen was developed using retrieval-based NLP techniques to align with the manualized PST's stepwise, algorithmic structure, guarding against issues such as hallucinations or non-relevant conversations. Lumen was successfully tested for feasibility, treatment fidelity, and

usability,[28] and showed promising signals of neural target engagement and clinical efficacy in a Phase 1 trial[20] (Suppl. Sections B-D).

As with the phase 1 trial, Lumen was integrated within the Alexa app on an iPad provided to study participants. For each session, participants used a voice invocation – "Open Lumen Session [number]" –and completed the assigned PST session. Between sessions, participants completed online surveys, ecological momentary assessments (EMA; Suppl, Section D), and received reminder notifications for their upcoming or missed sessions.

Human-coached PST

Participants in the human-coached arm received PST from a trained health coach over 4 weekly and then 4 biweekly sessions via Zoom, except for the first session, which was in-person. The structure and content of these sessions were based on the PST treatment manual,[19] similar to the Lumen-coached sessions.

Waitlist control

Participants in the waitlist control arm received automated surveys and EMAs at intervals similar to the Lumen- and human-coached arms. These participants could choose to receive a Lumen-enabled iPad after their end-of-study assessments.

Measures

Blinded outcome assessors conducted standardized assessments at baseline and at 18 weeks. These included neural and patient-reported measures of the hypothesized mechanistic targets related to cognitive control and emotion reactivity, and patient-reported clinical and functional outcome measures.

Neural and patient-reported measures of mechanistic targets

Task-based functional magnetic resonance imaging (fMRI) data were collected utilizing previously established fMRI sequences and parameters (Suppl, Section E). Informed by phase 1 findings,[20,29,30] the primary region of interest (ROI) was the right dlPFC, engaged using a go/no-go task for cognitive control.

Secondary ROIs included the left dlPFC for cognitive control and activation of the bilateral amygdala in the non-conscious viewing condition using threat and sad faces (negative affect) as measures of emotion reactivity. Person-level activation of the ROIs for each contrast of interest for each task (e.g., no-go versus go, threat versus neutral faces) was derived in a manner consistent with prior studies.[31,32]

As a patient-reported measure of behavior associated with cognitive control and a plausible, theory-based mediator of PST on depression and anxiety symptoms, the Social Problem-solving Index-Revised Short Form (SPSI-R:S) assessed overall problem-solving ability and 5 subscales for problem orientation (positive, PPO; negative, NPO) and problem-solving styles (rational problem-solving, RPS; impulsive/careless style, ICS; and avoidant style, AS).[33] Additional patient-reported measures related to cognitive control and emotion reactivity included the Dysfunctional Attitudes Scale (DAS),[34] Penn State Worry Questionnaire (PSWQ),[35] and the Positive and Negative Affect Schedule (PANAS).[36]

Patient-reported symptom and functional outcomes

Clinical symptoms were measured using the Hospital Anxiety and Depression Scale (HADS),[37,38] which assessed depressive and anxiety symptom scores and summative overall psychological distress scores.

Additional patient-reported functional outcome measures included the Sheehan's Disability Scale (SDS)[39] and the Work Productivity and Activity Impairment (WPAI) questionnaire[40] (also see protocol[22])

Statistical analysis

We conducted intention-to-treat (ITT) analyses for superiority tests (Lumen-coached and human-coached vs. waitlist control) and both ITT and per-protocol analyses (i.e., excluding participants who did not complete 8 PST sessions per protocol) for noninferiority tests (Lumen-coached vs. human-coached).[41] Analysis of between-group differences for each neural or patient-reported target or outcome measure included all participants with follow-up data at 18 weeks. Participants were analyzed based on their assigned group.

As the neural targets were defined by standardized activation values based on healthy control norms that were previously validated,[42,43] between-group differences on changes in these measures from baseline to 18 weeks were assessed using Student's *t*-tests.

Between-group differences on changes in patient-reported target, symptom and functional outcome measures from baseline to 18 weeks were tested using ordinary least square (OLS) regression, adjusting for baseline values of the target or outcome measure. We reported model-adjusted between-group mean differences with 95% confidence intervals (CIs) and two-tailed test P values for superiority tests, and with 90% CIs and one-tailed test P values for noninferiority tests. Statistical significance was determined at $\alpha=0.05$. Cohen's d was calculated using the mean difference between two groups divided by the pooled standard deviation (SD). Cohen's d = 0.3 (small-to-medium effect) was the noninferiority margin prespecified for comparing Lumen- and human-coach PST on change in right dlPFC activation. Multiple testing

corrections were not planned and noninferiority margins were not specified for secondary measures given their exploratory nature, with the focus being on effect estimation to inform future research.

Using the approach by Kraemer et al.[44] for mediation analysis, we defined that mediation existed if a potential mediator met two conditions. First, the effect of either Lumen- or human-coached PST vs. waitlist control (X) on the potential mediator (M, X→M Path A) was statistically significant. Second, the potential mediator was significantly associated with the symptom outcome either as a main effect or an interaction effect with the treatment vs. waitlist control (M→Y, Path B). As all data were assessed at baseline and 18 weeks, the focus was on contemporaneous mediation.[45]

All analyses were conducted using SAS, version 9.4 (SAS Institute Inc., Cary, North Carolina).

Sample size calculation

For getting additional data on Lumen as the experimental treatment, a 2:1:1 allocation to the Lumen-coached, human-coached, and waitlist control groups was used. Superiority testing of change in the primary neural target (i.e., right dlPFC for cognitive control) comparing Lumen-coached PST and waitlist control was of primary interest. We calculated that a sample of 150 (100 Lumen, 50 waitlist control) would provide 80% power to detect a medium effect of Cohen's d=0.45 at α=5% (2-sided), assuming ≥85% retention. The human-coached and waitlist control groups were designed to be of equal size, for a total N=200.

Results

Sample characteristics and retention

Of the 3176 individuals who completed screening, 2773 were ineligible and 142 declined or were unable to participate (Figure 1). Ineligible depressive and/or anxiety symptom scores and current psychiatric medication or psychotherapy were primary reasons for exclusion.

Participants (N=200) had a mean age of 36.6 (SD=11.9) years, were primarily female (77.5%), and were from varied racial/ethnic (25.0% non-Hispanic Black, 28.5% Hispanic, 21.0% Asian/Pacific Islander) and socioeconomic groups (62.0% with high school or college education, 47.5% with annual income <$55,000) (Table 1). At baseline, participants had a mean PHQ-9 score of 12.3 (SD=3.5) and a mean GAD-7 score of 10.9 (SD=2.4), indicating moderate severity of symptoms on average; 61.5% of participants had PHQ-9 and GAD-7 scores >10, indicating clinically significant symptoms of both depression and anxiety.

All participants completed baseline assessments, 194 (97%) were assessed at 18 weeks, and 141 (70.5%) had fMRI data passing quality control checks (Suppl. Section E) at both time points for the primary analysis. The number of participants who completed 8 PST sessions was 73 (of 100, 73%) in the Lumen arm and 44 (of 50, 88%) in the human-coached arm.

Neural and patient-reported measures of mechanistic targets

Mean change in right dlPFC activation from baseline to 18 weeks did not differ significantly between either the Lumen- or human-coached group and the waitlist control group (Table 2); therefore, noninferiority analysis between the Lumen- and human-coached groups was not conducted. Secondary neural target measures also did not show any statistical significance (Suppl, Section F).

Compared with waitlist controls, participants had a significantly greater increase from baseline to 18 weeks in overall problem-solving ability in both the Lumen-coached (between-group mean difference=1.04, 95%CI [0.23, 1.84]; Cohen's d=0.34) and human-coached arms (between-group mean difference=1.11, 95%CI [0.18, 2.04]; Cohen's d=0.52). Mean change in overall problem-solving ability did not differ significantly in both ITT and per-protocol noninferiority tests of Lumen- vs. human-coached participants (Table 2 and Suppl, Section G).

Compared with waitlist control, Lumen-coached participants also had significantly greater improvements from baseline to 18 weeks in PPO, NPO, worry, positive and negative affect, with Cohen's d effects ranging 0.24-0.5 (Table 2 and Suppl, Section F). Changes in RPS, ICS, AS, and dysfunctional attitudes did not differ significantly between the Lumen-coached and waitlist control arms.

Similarly, compared with waitlist controls, human-coached participants had significantly greater improvements from baseline to 18 weeks in NPO, worry, and negative affect, with Cohen's d effects ranging 0.49-0.84 (Table 2 and Suppl, Section F). Changes in PPO, RPS, ICS, AS, dysfunctional attitudes, and negative affect did not differ between the human-coached and waitlist control arms.

Patient-reported clinical and functional outcomes

Compared with waitlist controls, participants had a significantly greater reduction from baseline to 18 weeks in symptoms of psychological distress due to depression and anxiety in both the Lumen-coached (between-group mean difference=-3.56, 95%CI [-5.69, -1.43]; Cohen's d=0.49) and human-coached arms (between-group mean difference=-4.81, 95%CI [-7.28, -2.35]; Cohen's d=0.65). Mean change in overall psychological distress did not differ significantly in

both ITT and per-protocol noninferiority tests of Lumen- vs. human-coached participants (Table 3 and Suppl, Section G).

Compared with waitlist controls, Lumen-coached participants also had significantly greater reductions from baseline to 18 weeks in both depressive and anxiety symptoms, disability, percent impairment at work, percent overall work productivity loss, and percent activity impairment, with Cohen's d effects ranging 0.19-0.61 (Table 3 and Suppl, Section F). Changes in the percent work time missed did not differ significantly between Lumen-coached and waitlist control participants.

Similarly, compared with waitlist controls, human-coached participants had significantly greater reductions from baseline to 18 weeks in both depressive and anxiety symptoms, disability, percent impairment at work, percent overall work productivity loss, and percent activity impairment, with Cohen's d effects ranging 0.34-0.76 (Table 3 and Suppl, Section F). Changes in the percent work time missed did not differ significantly between human-coached and waitlist control participants.

Mediation of symptom improvement by change in problem-solving ability

As the changes in the neural targets did not differ significantly between either intervention or waitlist control arms, mediation analysis was not performed using neural targets. Given the theoretical basis of PST, we conducted a mediation analysis on overall problem-solving ability and symptom outcomes. As described, both Lumen- and human-coached PST led to a significantly greater improvement in overall problem-solving ability than waitlist control (Path A).

For Path B, comparing changes from baseline to 18 weeks between the Lumen-coached and waitlist control arms, improvement in overall problem-solving ability correlated significantly with reductions in psychological distress (β=-0.76, 95%CI [-1.46, -0.07]) and anxiety symptoms (β=-0.53 95%CI [-0.98, -0.08]) (Suppl, Section H).

Similarly, comparing changes from baseline to 18 weeks between the human-coached and waitlist control arms, improvement in overall problem-solving ability correlated significantly with reductions in psychological distress (β=-0.84, 95%CI [-1.56, -0.13]) and anxiety symptoms (β=-0.57, 95%CI [-0.99, -0.15]) (Suppl, Section H).

Change in overall problem-solving ability was not significantly correlated with change in depressive symptoms comparing either the Lumen- or human-coached arm with waitlist control.

Adverse events

There were 1 serious (expected, unrelated), involving a hospitalization for pneumonia, and 35 nonserious (5 unexpected, unrelated; 30 expected, unrelated) adverse events (Suppl, Section I). There were no deaths.

Discussion

This 3-arm, phase 2 trial evaluated the mechanisms and potential efficacy of a voice-based AI coach for PST, Lumen, compared with waitlist control and human-coached PST, among 200 adults with untreated, clinically significant depressive and/or anxiety symptoms. Neither Lumen-coached or human-coached PST resulted in significant changes in the a priori primary (right dlPFC) or secondary neural targets related to cognitive control and emotion reactivity, compared with waitlist control. However, similar to human-coached PST, Lumen-coached PST significantly improved the overall problem-solving ability, a patient-reported target measure

related to cognitive control, and the overall psychological distress due to depressive and anxiety symptoms, compared with waitlist control. Moreover, improved problem-solving ability mediated reduced psychological distress for both intervention modalities, supporting the theory-based mechanism of PST. Compared with waitlist control, Lumen-coached PST also led to significantly greater improvements in patient-reported target measures related to emotion reactivity (worry, positive and negative affect) and functional outcomes (disability, work and activity impairments, and work productivity). Lumen- and human-coached PST did not differ significantly on any of the patient-reported target, symptom and functional outcome measures.

This study did not replicate the phase I trial finding of neural target engagement (i.e., right dlPFC) for cognitive control by Lumen-coached PST. Additionally, the findings were also non-significant for human-coached PST. Despite high retention overall, 30% participants were missing acceptable fMRI data due to excessive motion or poor signal-to-noise ratios, which limits the validity of the neural target analyses. Furthermore, it is likely that there are additional neural correlates of improvements in problem-solving ability and clinical outcomes that extend beyond the *a priori* single ROIs examined in the current study. For example, we previously demonstrated that restoration of both activity and functional connectivity in key dorsal prefrontal regions of the cognitive control circuit was associated with improved problem-solving ability and depressive symptoms following human-coached PST.[46] Importantly, these circuit-level changes sustained over time, with effects observed across five assessment points spanning 24 months. Further studies utilizing similar circuit-level analysis and a data-driven, whole-brain network approach may yield insights regarding neural target engagement for Lumen- and human-coached PST.

The secondary findings of the effects of Lumen-coached PST on patient-reported target engagement and symptom and functional outcomes are preliminary. However, these findings provide meaningful signals that warrant further investigation. First, the current study illustrated a plausible mechanism of action by showing PST theory-concordant target engagement (i.e., overall problem-solving ability improved significantly in the Lumen and human interventions compared with waitlist control) and target validation (i.e., improved problem-solving ability was significantly associated with improved symptoms of psychological distress) for both delivery modalities. Second, this study demonstrated Lumen-coached PST's promise as a treatment option for patients with depression and anxiety by replicating the phase 1 trial results of symptom improvements compared with waitlist control, with additional findings of improvements in related measures of worry, affect, and psychosocial functioning. Taken together, these findings emphasize the potential of Lumen for further testing in an efficacy trial.

Moreover, the direct comparison to human-coached PST further suggests that Lumen's therapeutic potential warrants confirmation. This study did not prespecify noninferiority margins for comparing Lumen- and human-coached PST except for the primary neural target measure.

We benchmarked the results on the symptom outcomes against minimal clinically important differences (MCID) for the HADS measures to aid in interpretation for informing future research. The MCIDs for the HADS vary widely by study populations and methodologies, ranging from 2.1-8.5 for the total psychological distress and from 1.1-5.6 for the depressive and anxiety subscales.[47-50] In this trial, the effect estimates with 90% CIs based on the noninferiority tests comparing the HADS total and subscale score changes showed that the magnitudes of differences between Lumen- and human-coached PST were close to the lower bounds of the

MCID distributions. As such, these provide important preliminary data for designing a future confirmatory noninferiority trial to ascertain the potential of Lumen as a clinically acceptable alternative modality for PST delivery.

Consistent with the present findings supporting Lumen's therapeutic potential, recent systematic reviews and meta-analyses have shown that chatbot-based mental health interventions can reduce depressive and anxiety symptoms.[8,14,51,52] However, effects were generally modest and heterogeneous and were limited by variability in intervention design, comparator conditions, and follow-up duration, underscoring the need for more rigorous randomized clinical trials. In additional, few studies have directly examined mediators or mechanisms of therapeutic change, and evidence for voice-based AI interventions remains limited.[14] The current study addressed these gaps.

To the best of our knowledge, Lumen is the first voice-based AI coach for mental health treatment to be evaluated within an experimental therapeutics framework in randomized clinical trials, with an explicit focus on mechanisms of action in intervention development and testing.[21] In line with this framework, the study investigated theory-driven neural and patient-reported behavioral mechanistic targets. The inclusion of human-coached PST as an active comparator was also a notable strength of this study.[53]

The integrated "device-as-coach" architecture and delivery model of Lumen supports scalability, clinical translation, and longitudinal monitoring.[54] Lumen's architecture prioritizes safety through its rule-based, PST-aligned design and its extension of AI capabilities within the Alexa platform, thereby minimizing risks associated with hallucinations and inaccurate responses that have emerged as key concerns with contemporary AI chatbots.[55] A key aspect of

the Lumen development process involved balancing fidelity to PST principles–validated in collaboration with PST trainers–with the integration of AI-enabled features. Accordingly, Lumen leveraged and extended existing Alexa capabilities for PST delivery while incorporating features to support seamless user interactions, such as pausing and rescheduling sessions. Du et al.[8] reported that evidence for rule-based chatbots was comparatively stronger because more studies were available, whereas evidence for LLM-based chatbots remains preliminary but suggests potential advantages in personalization, conversational quality, and user engagement. The extended timeline of Lumen's development and associated clinical trials, spanning approximately six years, limited the feasibility of incorporating newer generative AI approaches during the study period. Given emerging evidence suggesting that generative AI chatbots may improve mental health outcomes,[9,10] efforts to adapt and advance Lumen using these newer technologies are ongoing.

This study has several limitations. The sample size was relatively small and limited to individuals with clinically significant, but non-severe depressive and/or anxiety symptoms. In addition, the study excluded some individuals who may have benefited from the intervention because of design and feasibility constraints, including inability to undergo fMRI scanning, limited internet access, or non-English language status. Furthermore, no adjustments were made for multiple testing across secondary mechanistic targets and clinical outcomes and noninferiority margins were not prespecified for these measures; therefore, the findings should be interpreted with caution and require validation in future studies.

Among adults with untreated, clinically significant depression and anxiety, a rule-based AI coach delivering PST through a voice-based platform did not significantly improve the primary

neural mechanistic target related to cognitive control compared with waitlist control. However, preliminary findings suggesting benefits for problem-solving ability and psychologic distress relative to waitlist control, as well as evidence of potential noninferiority compared with human-delivered PST, are encouraging and warrant further investigation.

Funding

This work was supported by the National Institute of Mental Health (NIMH) [grant number R33MH119237].

Conflict of Interests

Dr. Jun Ma serves as an editor for an Oxford University Press journal, outside of this work. Dr. Olusola A. Ajilore is the co-founder of Keywise AI and serves on the advisory boards of Blueprint Health and Embodied Labs. Dr. Thomas Kannampallil serves as an editor for an Elsevier journal and unpaid member of the research advisory group for Abridge Inc, outside of this work. The other authors report no conflicts of interest.

Data Availability Statement

Data used in the preparation of this manuscript will be submitted to the National Institute of Mental Health (NIMH) Data Archive (NDA). NDA is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in mental health. Those wishing to use this data can contact the corresponding author for the dataset identifier and make a request to the NIMH (visit https://nda.nih.gov/). This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH.

Author Contributions

Drs. Ma and Xiao had full access to all of the data and took responsibility for the data and accuracy of the data analysis. Drs. Kannampallil and Ajilore contributed equally to this manuscript.

Concept and design: Kannampallil, Ajilore, Smyth, Ma.

Acquisition, analysis, or interpretation of data: Kannampallil, Ajilore, Smyth, Barve, Ronneberg, Lv, Kumar, Garcia, Aborisade, Wittels, Tang, Chinnakotla, Xiao, Ma.

Critical review of the manuscript for important intellectual content: Kannampallil, Ajilore, Smyth, Barve, Ronneberg, Lv, Kumar, Garcia, Aborisade, Wittels, Tang, Chinnakotla, Xiao, Ma.

Statistical analysis: Kannampallil, Ajilore, Lv, Xiao, Ma

Obtained funding: Kannampallil, Ajilore, Smyth, Ma

Administrative, technical, or material support: Kannampallil, Ajilore, Lv, Kumar, Wittels, Xiao, Ma.

Supervision: Kannampallil, Ajilore, Ma

References

1. Terlizzi EP, Zablotsky B. *Symptoms of Anxiety and Depression Among Adults: United States, 2019 and 2022. National Health Statistics Reports; no 213*. 2024.

2. World Health Organization. Over a billion people living with mental health conditions – services require urgent scale-up. Updated September 2, 2025. Accessed October 2, 2025. https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up

3. Mental Health America. The State of Mental Health in America. Accessed August 25, 2025. https://mhanational.org/the-state-of-mental-health-in-america/

4. Orozco R, Vigo D, Benjet C, et al. Barriers to treatment for mental disorders in six countries of the Americas: A regional report from the World Mental Health Surveys. *J Affect Disord*. Apr 15 2022;303:273-285. doi:10.1016/j.jad.2022.02.031

5. McHugh RK, Whitton SW, Peckham AD, Welge JA, Otto MW. Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: a meta-analytic review. *J Clin Psychiatry*. Jun 2013;74(6):595-602. doi:10.4088/JCP.12r07757

6. Sorkin DH, Murphy M, Nguyen H, Biegler KA. Barriers to Mental Health Care for an Ethnically and Racially Diverse Sample of Older Adults. *J Am Geriatr Soc*. Oct 2016;64(10):2138-2143. doi:10.1111/jgs.14420

7. Leong FT, Kalibatseva Z. Cross-cultural barriers to mental health services in the United States. *Cerebrum*. Mar 2011;2011:5.

8. Du Q, Ren Y, Meng ZL, He H, Meng S. The Efficacy of Rule-Based Versus Large Language Model-Based Chatbots in Alleviating Symptoms of Depression and Anxiety: Systematic Review and Meta-Analysis. *J Med Internet Res*. Dec 4 2025;27:e78186. doi:10.2196/78186

9. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*. 2025;2(4)doi:10.1056/AIoa2400802

10. Shoshani A, Gurfinkel B, Kor A, et al. Efficacy of a Conversational AI Agent for Psychiatric Symptoms and Digital Therapeutic Alliance: A Randomized Clinical Trial. *JAMA Netw Open*. Apr 1 2026;9(4):e266713. doi:10.1001/jamanetworkopen.2026.6713

11. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J Med Internet Res*. Jul 13 2020;22(7):e16021. doi:10.2196/16021

12. Kannampallil T, Smyth JM, Jones S, Payne PR, Ma J. Cognitive plausibility in voice-based AI health counselors. *NPJ digital medicine*. 2020;3(1):1-4.

13. Steinhubl SR, Topol EJ. Now we're talking: bringing a voice to digital medicine. *The Lancet*. 2018;392(10148):627.

14. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit Med*. Dec 19 2023;6(1):236. doi:10.1038/s41746-023-00979-5

15. Xu S, Ma T. Depression intervention using AI chatbots with social cues: a randomized trial of effectiveness. *J Affect Disord*. Nov 15 2025;389:119760. doi:10.1016/j.jad.2025.119760

16. Sun Y, Xu S, Jin H, et al. Effectiveness of Al-Assisted Patient Health Education Using Voice Cloning and ChatGPT: Prospective Randomized Controlled Trial. *J Med Internet Res*. Mar 19 2026;28:e81387. doi:10.2196/81387

17. Kannampallil T, Ronneberg CR, Wittels NE, et al. Design and Formative Evaluation of a Virtual Voice-Based Coach for Problem-solving Treatment: Observational Study. *JMIR Formative Research*. 2022;6(8):e38092.

18. Lv N, Kannampallil T, Xiao L, et al. Association Between User Interaction and Treatment Response of a Voice-Based Coach for Treating Depression and Anxiety: Secondary Analysis of a Pilot Randomized Controlled Trial. *JMIR Hum Factors*. Nov 6 2023;10:e49715. doi:10.2196/49715

19. Nezu AM, Nezu CM, D'Zurilla T. *Problem-solving therapy: A treatment manual*. Springer Publishing Company, LLC; 2013.

20. Kannampallil T, Ajilore OA, Lv N, et al. Effects of a virtual voice-based coach delivering problem-solving treatment on emotional distress and brain function: a pilot RCT in depression and anxiety. *Transl Psychiatry*. May 12 2023;13(1):166. doi:10.1038/s41398-023-02462-x

21. Nielsen L, Riddle M, King JW, et al. The NIH Science of Behavior Change Program: Transforming the science through a focus on mechanisms of change. *Behav Res Ther*. Feb 2018;101:3-11. doi:10.1016/j.brat.2017.07.002

22. Ronneberg CR, Lv N, Ajilore OA, et al. Study of a PST-trained voice-enabled artificial intelligence counselor for adults with emotional distress (SPEAC-2): Design and methods. *Contemp Clin Trials*. Jul 2024;142:107574. doi:10.1016/j.cct.2024.107574

23. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. Sep 2001;16(9):606-13. doi:10.1046/j.1525-1497.2001.016009606.x

24. Spitzer RL, Kroenke K, Williams JB, Lowe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. May 22 2006;166(10):1092-7. doi:10.1001/archinte.166.10.1092

25. Xiao L, Huang Q, Yank V, Ma J. An easily accessible Web-based minimization random allocation system for clinical trials. *J Med Internet Res*. 2013;15(7):e139. doi:10.2196/jmir.2392

26. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials. a review. *Control Clin Trials*. Dec 2002;23(6):662-74. doi:10.1016/s0197-2456(02)00242-8

27. 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. *J Med Internet Res*. 2017;19(1):e27.

28. Kannampallil T, Ronneberg CR, Wittels NE, et al. Design and Formative Evaluation of a Virtual Voice-Based Coach for Problem-solving Treatment: Observational Study. *JMIR Form Res*. Aug 12 2022;6(8):e38092. doi:10.2196/38092

29. Lv N, Ajilore OA, Xiao L, et al. Mediating Effects of Neural Targets on Depression, Weight, and Anxiety Outcomes of an Integrated Collaborative Care Intervention: The ENGAGE-2 Mechanistic Pilot Randomized Clinical Trial. *Biol Psychiatry Glob Open Sci*. Jul 2023;3(3):430-442. doi:10.1016/j.bpsgos.2022.03.012

30. Goldstein-Piekarski AN, Wielgosz J, Xiao L, et al. Early changes in neural circuit function engaged by negative emotion and modified by behavioural intervention are associated with depression and problem-solving outcomes: A report from the ENGAGE randomized controlled trial. *EBioMedicine*. May 15 2021;67:103387. doi:10.1016/j.ebiom.2021.103387

31. Williams LM, Korgaonkar MS, Song YC, et al. Amygdala Reactivity to Emotional Faces in the Prediction of General and Medication-Specific Responses to Antidepressant Treatment in the Randomized iSPOT-D Trial. *Neuropsychopharmacology*. Sep 2015;40(10):2398-408. doi:10.1038/npp.2015.89

32. Goldstein-Piekarski AN, Ball TM, Samara Z, et al. Mapping neural circuit biotypes to symptoms and behavioral dimensions of depression and anxiety. *Biological Psychiatry*. 2022;91(6):561-571.

33. D'Zurilla T, Nezu A, Maydeu-Olivares A. *Manual for the Social Problem-Solving Inventory-Revised*. Multi-Health Systems; 2002.

34. Weissman AN. *The Dysfunctional Attitude Scale: A Validation Study*. University of Pennsylvania; 1979. http://repository.upenn.edu/edissertations/1182

35. Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State Worry Questionnaire. *Behav Res Ther*. 1990;28(6):487-95. doi:10.1016/0005-7967(90)90135-6

36. 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. doi:10.1037/0022-3514.54.6.1063

37. Snaith R, Zigmond A. *The Hospital Anxiety and Depression Scale manual*. Nfer-Nelson; 1994.

38. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta psychiatrica Scandinavica*. Jun 1983;67(6):361-70.

39. Sheehan KH, Sheehan DV. Assessing treatment effects in clinical trials with the discan metric of the Sheehan Disability Scale. *International clinical psychopharmacology*. Mar 2008;23(2):70-83. doi:10.1097/YIC.0b013e3282f2b4d6

40. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*. Nov 1993;4(5):353-65. doi:10.2165/00019053-199304050-00006

41. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PP. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. *BMJ Open*. Oct 7 2016;6(10):e012594. doi:10.1136/bmjopen-2016-012594

42. Goldstein-Piekarski AN, Ball TM, Samara Z, et al. Mapping Neural Circuit Biotypes to Symptoms and Behavioral Dimensions of Depression and Anxiety. *Biol Psychiatry*. Mar 15 2022;91(6):561-571. doi:10.1016/j.biopsych.2021.06.024

43. Ahn J, Foland-Ross L, Akiki TJ, et al. Developing Clinically Interpretable Neuroimaging Biotypes in Psychiatry. *Biol Psychiatry*. Sep 8 2025;doi:10.1016/j.biopsych.2025.08.019

44. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry*. Oct 2002;59(10):877-83. doi:10.1001/archpsyc.59.10.877

45. Lockhart G, MacKinnon DP, Ohlrich V. Mediation analysis in psychosomatic medicine research. *Psychosom Med*. Jan 2011;73(1):29-43. doi:10.1097/PSY.0b013e318200a54b

46. Zhang X, Pines A, Stetz P, et al. Adaptive cognitive control circuit changes associated with problem-solving ability and depression symptom outcomes over 24 months. *Sci Transl Med*. Sep 4 2024;16(763):eadh3172. doi:10.1126/scitranslmed.adh3172

47. Puhan MA, Frey M, Buchi S, Schunemann HJ. The minimal important difference of the hospital anxiety and depression scale in patients with chronic obstructive pulmonary disease. *Health Qual Life Outcomes*. Jul 2 2008;6:46. doi:10.1186/1477-7525-6-46

48. Lemay KR, Tulloch HE, Pipe AL, Reed JL. Establishing the Minimal Clinically Important Difference for the Hospital Anxiety and Depression Scale in Patients With Cardiovascular Disease. *J Cardiopulm Rehabil Prev*. Nov 2019;39(6):E6-E11. doi:10.1097/HCR.0000000000000379

49. de Filippis R, Mercurio M, Segura-Garcia C, De Fazio P, Gasparini G, Galasso O. Defining the minimum clinically important difference (MCID) in the hospital anxiety and depression scale (HADS) in patients undergoing total hip and knee arthroplasty. *Orthop Traumatol Surg Res*. Apr 2024;110(2):103689. doi:10.1016/j.otsr.2023.103689

50. Longo UG, Papalia R, De Salvatore S, et al. Establishing the Minimum Clinically Significant Difference (MCID) and the Patient Acceptable Symptom Score (PASS) for the Hospital Anxiety and Depression Scale (HADS) in Patients with Rotator Cuff Disease and Shoulder Prosthesis. *J Clin Med*. Feb 15 2023;12(4)doi:10.3390/jcm12041540

51. Villarreal-Zegarra D, Reategui-Rivera CM, Garcia-Serna J, et al. Self-Administered Interventions Based on Natural Language Processing Models for Reducing Depressive and Anxiety Symptoms: Systematic Review and Meta-Analysis. *JMIR Ment Health*. Aug 21 2024;11:e59560. doi:10.2196/59560

52. Lau Y, Ang WHD, Ang WW, Pang PCI, Wong SH, Chan KS. Artificial Intelligence–Based Psychotherapeutic Intervention on Psychological Outcomes: A Meta-Analysis and Meta-Regression. *Depression and Anxiety*. 2025;2025(1)doi:https://doi.org/10.1155/da/8930012

53. Goldberg SB, Sun S, Carlbring P, Torous J. Selecting and describing control conditions in mobile health randomized controlled trials: a proposed typology. *NPJ Digit Med*. Sep 30 2023;6(1):181. doi:10.1038/s41746-023-00923-7

54. Kohane IS. Compared with What? Measuring AI against the Health Care We Have. *N Engl J Med*. Oct 31 2024;391(17):1564-1566. doi:10.1056/NEJMp2404691

55. McBain RK. Teens Are Using Chatbots as Therapists. That's Alarming. *The New York Times*. August 25. Accessed October 2, 2025. https://www.nytimes.com/2025/08/25/opinion/teen-mental-health-chatbots.html#

Figure Legends

Figure 1. Consort chart.

Tables

Table 1. Baseline Characteristics*

Characteristic	All participants (n=200)	Lumen (n=100)	Waitlist (n=50)	Human (n=50)
Age, years†	36.6 ± 11.9	36.1 ± 11.8	37.1 ± 12.4	37.0 ± 11.9
Female, %†	155(77.5)	77(77.0)	39(78.0)	39(78.0)
Race/Ethnicity, %†				
Non-Hispanic White	43(21.5)	21(21.0)	11(22.0)	11(22.0)
Non-Hispanic Black	50(25.0)	28(28.0)	13(26.0)	9(18.0)
Asian/Pacific Islander	42(21.0)	23(23.0)	9(18.0)	10(20.0)
Hispanic	57(28.5)	25(25.0)	13(26.0)	19(38.0)
Other (e.g., declined to state, multirace)	8(4.0)	3(3.0)	4(8.0)	1(2.0)
Education, %†				
High school/GED or less	19(9.5)	11(11.0)	4(8.0)	4(8.0)
College - 1 year to 3 years	42(21.0)	19(19.0)	14(28.0)	9(18.0)
College - 4 years or more	63(31.5)	35(35.0)	14(28.0)	14(28.0)
Post college	76(38.0)	35(35.0)	18(36.0)	23(46.0)
Income, %				
< $35,000	58(29.0)	31(31.0)	15(30.0)	12(24.0)
$35,000- <$55,000	37(18.5)	19(19.0)	10(20.0)	8(16.0)
$55,000- <$75,000	22(11.0)	12(12.0)	4(8.0)	6(12.0)
>=$75,000	83(41.5)	38(38.0)	21(42.0)	24(48.0)
Digital Health Literacy, %†				
Low 1-1.999	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Medium 2-2.999	24(12.0)	13(13.0)	5(10.0)	6(12.0)
High 3-4	176(88.0)	87(87.0)	45(90.0)	44(88.0)
PHQ-9 category, %				
minimal depression 0-4	4(2.0)	2(2.0)	0(0.0)	2(4.0)
mild depression 5-9	28(14.0)	16(16.0)	6(12.0)	6(12.0)
Moderate depression 10-14	116(58.0)	56(56.0)	32(64.0)	28(56.0)
Moderately severe depression 15-19	52(26.0)	26(26.0)	12(24.0)	14(28.0)
GAD-7 category, %				
minimal anxiety 0-4	3(1.5)	1(1.0)	2(4.0)	0(0.0)
mild anxiety 5-9	42(21.0)	21(21.0)	10(20.0)	11(22.0)
moderate anxiety 10-14	155(77.5)	78(78.0)	38(76.0)	39(78.0)
PHQ-9 score†	12.3 ± 3.5	12.1 ± 3.6	12.7 ± 3.1	12.2 ± 3.6
GAD-7 score†	10.9 ± 2.4	10.9 ± 2.4	10.8 ± 2.6	11.0 ± 2.3

PHQ-9 and GAD-9 inclusion criteria				
Met PHQ-9 (10-19)	45(22.5)	22(22.0)	12(24.0)	11(22.0)
Met GAD-7 (10-14)	32(16.0)	18(18.0)	6(12.0)	8(16.0)
Met both criteria	123(61.5)	60(60.0)	32(64.0)	31(62.0)

Abbreviations: GAD-7, Generalized Anxiety Disorder-7; GED, general educational development; HADS, Hospital Anxiety and Depression Scale; PHQ-9, Patient Health Questionnaire-9.
*Values are mean ± SD unless noted otherwise.
†Prognostic factors for randomization: age, sex, race/ethnicity, education, digital health literacy, PHQ-9, and GAD-7.

Table 2. Treatment effects on neural and patient-reported mechanistic targets

				Superiority test						Noninferiority test		
	Unadjusted mean ± SD			Lumen versus Waitlist			Human versus Waitlist			Lumen versus Human		
Neural targets	**Lumen**	**Human**	**Waitlist**	**Mean difference (95% CI)***	**P value ***	**Cohen's d† (Lumen - Waitlist**	**Mean difference (95% CI)***	**P value ***	**Cohen's d† (Human- Waitlist**			
dlPFC R (Cognitive Control Circuit)	n=71	n=36	n=34							**Not Applicable**		
baseline	0.26 ± 0.49	0.30 ± 0.44	0.14 ± 0.47									
change at 18 weeks	-0.01 ± 0.61	0.10 ± 0.77	-0.03 ± 0.60	0.02 (-0.23, 0.27)	0.89	0.03	0.13 (-0.2, 0.46)	0.44	0.18			
	Unadjusted mean ± SD			Lumen versus Waitlist			Human versus Waitlist			Lumen versus Human		
Patient-reported targets	**Lumen**	**Human**	**Waitlist**	**Model-based mean difference (95%CI)‡**	**P value (2-sided)‡**	**Cohen's d† (Lumen - Waitlist**	**Model-based mean difference (95%CI)‡**	**P value (2-sided)‡**	**Cohen's d† (Human- Waitlist**	**Model-based mean difference (90%CI)‡**	**P value (1-sided)‡**	**Cohen's d† (Human-Lumen)**
SPSI-R:S raw score§	n=96	n=48	n=49									
baseline	12.01 ± 2.91	11.08 ± 2.75	11.40 ± 2.98									
change at 18 weeks	1.39 ± 2.63	1.77 ± 2.53	0.52 ± 2.26	1.04 (0.23, 1.84)	0.01	0.34	1.11 (0.18, 2.04)	0.02	0.52	-0.07 (-0.75, 0.61)	0.45	0.15
PPO raw score§	n=96	n=48	n=49									
baseline	11.22 ± 3.98	10.18 ± 3.78	10.06 ± 4.61									
change at 18 weeks	1.91 ± 3.86	2.19 ± 4.25	0.94 ± 4.27	1.46 (0.21, 2.71)	0.02	0.24	1.26 (-0.18, 2.7)	0.09	0.29	0.2 (-0.86, 1.25)	0.60	0.07
NPO raw score§	n=96	n=48	n=49									
baseline	9.60 ± 4.62	10.38 ± 4.05	10.94 ± 4.37									
change at 18 weeks	-2.32 ± 4.09	-3.10 ± 4.21	-1.08 ± 4.06	-1.71 (-3.02, -0.39)	0.01	0.30	-2.25 (-3.77, -0.74)	0.004	0.49	0.54 (-0.56, 1.65)	0.28	0.19
RPS raw score§	n=96	n=48	n=49									

baseline	11.11 ± 4.40	9.90 ± 4.23	11.74 ± 4.49									
change at 18 weeks	1.48 ± 4.38	1.56 ± 4.32	0.04 ± 3.51	1.07 (-0.17, 2.31)	0.09	0.35	0.5 (-0.96, 1.95)	0.5	0.39	0.57 (-0.48, 1.63)	0.73	0.02
ICS raw score§	n=96	n=48	n=49									
baseline	5.69 ± 4.27	6.44 ± 4.00	5.88 ± 4.01									
change at 18 weeks	-0.46 ± 3.98	-0.40 ± 3.99	-0.18 ± 4.01	-0.36 (-1.57, 0.86)	0.56	0.07	0.04 (-1.37, 1.45)	0.95	0.05	-0.4 (-1.43, 0.63)	0.68	-0.02
AS raw score§	n=96	n=48	n=49									
baseline	7.01 ± 4.47	7.88 ± 4.87	7.96 ± 4.42									
change at 18 weeks	-0.77 ± 4.10	-1.58 ± 4.09	-0.37 ± 4.64	-0.77 (-2.05, 0.52)	0.24	0.09	-1.24 (-2.72, 0.25)	0.1	0.28	0.47 (-0.62, 1.56)	0.30	0.20

Abbreviations: AS, avoidant problem-solving style; CI, confidence interval; ICS, impulsive/careless problem-solving style; NPO, negative problem orientation; PPO, positive problem orientation; RPS, rational problem-solving style; SPSI-R:S, Social Problem-solving Index-Revised Short Form.

*_t_ tests.

†The mean difference between two groups divided by the pooled standard deviation.

‡Model-based mean differences with 95% CI and p values (2-sided) for superiority tests and with 90% CI and p values (1-sided) for noninferiority tests were generated from ordinary least square regression models, adjusted for baseline value of the outcome of interest.

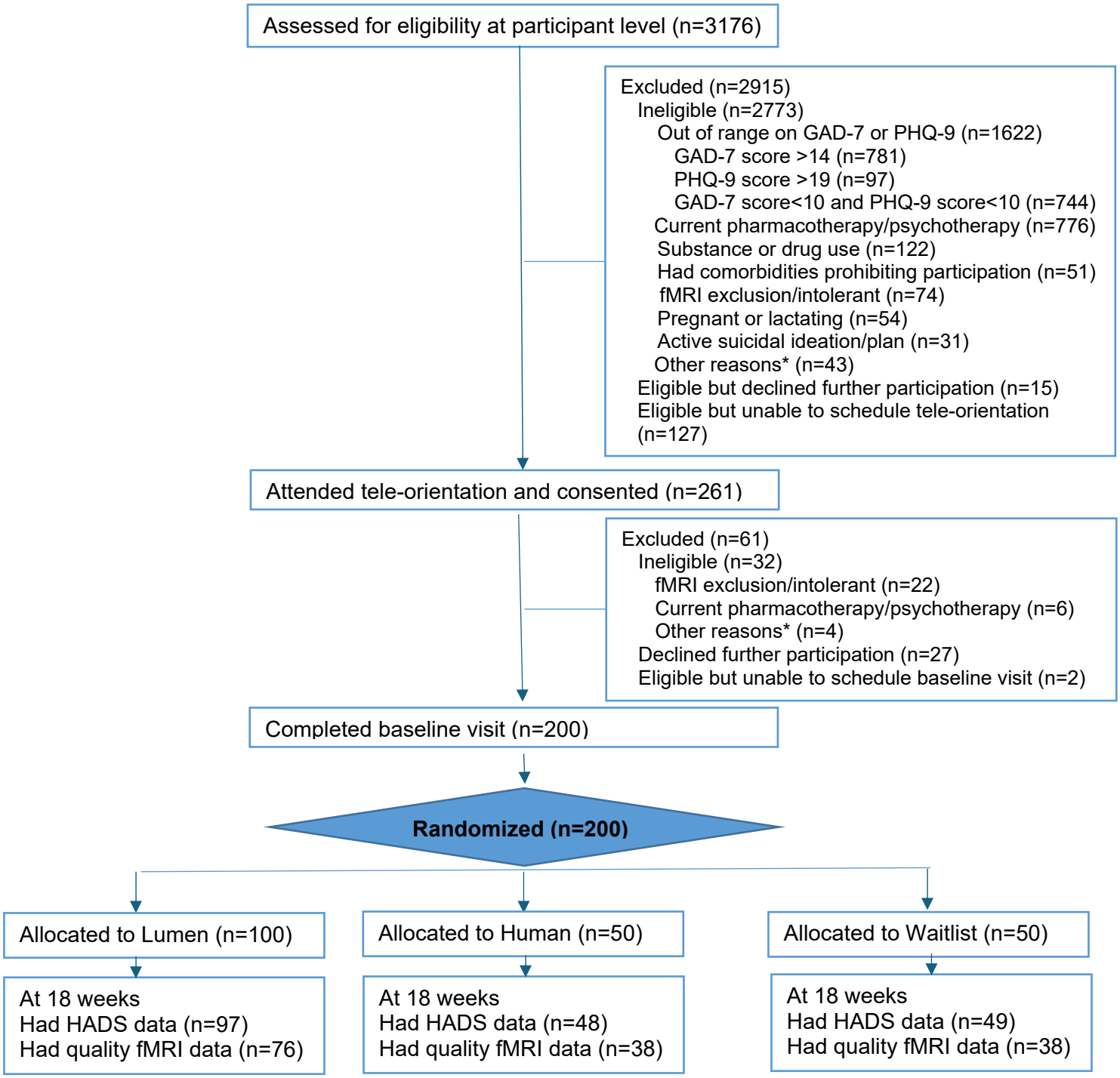
§SPSI-R:S score=(PPO raw score/5)+(20- NPO raw score)/5+ (RPS raw score/5)+(20- ICS raw score)/5+(20- AS raw score)/5; the higher the score the more productive overall problem-solving orientation and skills. Subscales (PPO, NPO, RPS, ICS, and AS) are raw scores without reversal.

Table 3. Treatment effects on patient-reported symptom outcomes

	Unadjusted mean ± SD			Superiority test						Noninferiority test		
				Lumen versus Waitlist			Human versus Waitlist			Lumen versus Human		
Measure	Lumen	Human	Waitlist	Model-based mean difference (95%CI)*	P value (2-sided)*	Cohen's d† (Lumen - Waitlist)	Model-based mean difference (95%CI)*	P value (2-sided)*	Cohen's d† (Human - Waitlist)	Model-based mean difference (90%CI)*	P value (1-sided)*	Cohen's d† (Human -Lumen)
Psychological distress‡	n=97	n=48	n=49									
baseline	18.38 ± 5.61	18.98 ± 4.52	19.68 ± 5.75									
change at 18 weeks	-5.85 ± 6.38	-7.33 ± 8.04	-2.76 ± 5.99	-3.56 (-5.69, -1.43)	0.001	0.49	-4.81 (-7.28, -2.35)	<.001	0.65	1.26 (-0.54, 3.05)	0.23	0.21
Depression§	n=97	n=48	n=49									
baseline	7.09 ± 3.32	7.84 ± 3.05	8.34 ± 3.01									
change at 18 weeks	-2.53 ± 3.22	-3.27 ± 4.47	-1.92 ± 3.33	-1.15 (-2.25, -0.04)	0.04	0.19	-1.6 (-2.87, -0.33)	0.01	0.34	0.45 (-0.47, 1.38)	0.28	0.20
Anxiety§	n=97	n=48	n=49									
baseline	11.29 ± 3.59	11.14 ± 3.10	11.34 ± 3.59									
change at 18 weeks	-3.32 ± 4.10	-4.06 ± 4.49	-0.84 ± 3.96	-2.43 (-3.72, -1.14)	<.001	0.61	-3.22 (-4.72, -1.73)	<.001	0.76	0.79 (-0.3, 1.88)	0.22	0.18

Abbreviations: CI, confidence interval; HADS, Hospital Anxiety and Depression Scale.
*Model-based mean differences with 95% CI and p values (2-sided) for superiority tests and with 90% CI and p values (1-sided) for noninferiority tests were generated from ordinary least square regression models, adjusted for baseline value of the outcome of interest.
†The mean difference between two groups divided by the pooled standard deviation.
‡HADS includes 7 questions about anxiety symptoms and 7 questions about depressive symptoms. Each item on the questionnaire is scored from 0-3. Total score for the entire scale ranges from 0 to 42, with higher score indicating more psychological distress due to symptoms of depression and anxiety.
§Scores range from 0-21, with 0-7 = Normal; 8-10 = Borderline abnormal (borderline case); 11-21 = Abnormal (case).

Figure 1. CONSORT Chart



* Other reasons include relocation, not fluent in English, no Wi-Fi, no mobile, no text, family member in study, did not complete baseline survey.