Mechanisms and Efficacy of a Voice-based AI Coach for Depression and Anxiety: A Phase 2 Randomized Trial

[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: 4

Number of Figures: 1

No. of Words (Abstract): 350

No. of Words (Manuscript): 3036

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

Key Points

Question: Can an Artificial Intelligence (AI)-driven, voice-based coach, Lumen, delivering an evidence-based problem-solving treatment (PST), improve neural and patient-reported measures of mechanistic targets and patient-reported clinical and functional outcomes better than (superior to) waitlist control and not worse than (noninferior to) human-coached PST?

Findings: In a phase 2, 3-arm randomized trial among 200 participants with untreated, clinically significant depression and/or anxiety, Lumen-coached PST did not show changes in neural targets but resulted in significantly greater reductions in patient-reported target, clinical, and functional measures, compared with waitlist control. Treatment effects of Lumen- and human-coached PST did not differ significantly on any of these measures.

Meaning: Although the neural targets did not differ significantly by treatment arm, the findings suggest that Lumen-coached PST may be a viable option for patients with clinically significant depression and/or anxiety who receive no treatment. However, clinical efficacy needs to be confirmed in a phase 3 trial.

Abstract

Importance: Artificial intelligence (AI) voice applications may address unmet depression and anxiety treatment needs. However, their therapeutic utility is largely unknown.

Objective: To investigate the mechanisms and efficacy of a voice-based AI coach, Lumen, delivering problem-solving treatment (PST) for depression and anxiety.

Design: Phase 2, 3-arm randomized trial.

Setting: A public university and affiliated medical center in Chicago, IL.

Participants: Adults with clinically significant depression and/or anxiety were randomized to Lumen-coached PST (n=100), human-coached PST (n=50), or waitlist control (n=50) for 18 weeks.

Interventions: PST was delivered by Lumen on Amazon's Alexa platform or a human coach via videoconferencing in 4-weekly and then 4-biweekly sessions.

Main Outcome(s) and Measure(s): Change in activation of the right dorsolateral prefrontal cortex (R dlPFC) for cognitive control using functional neuroimaging was the primary mechanistic target measure. Symptom outcomes were assessed using Hospital Anxiety and Depression Scale. Other secondary measures included left dlPFC and bilateral amygdala activation and patient-reported problem-solving abilities, dysfunctional attitudes, worry, positive and negative affect, disability, and work productivity. Statistical analyses used t-tests and ordinary least square regression.

Results: Participants were on average 36.6 years old (SD=11.9), 77% women, 25% Black, 29% Latino, and 21% Asian. At 18 weeks, change from baseline in R dlPFC activity did not differ significantly by treatment arm. Compared with waitlist control, Lumen-coached PST resulted in significantly greater improvements in symptoms of psychological distress (between-group mean difference=-3.56, 95%CI [-5.69, -1.43]), depression (-1.15, 95%CI [-2.25, -0.04]), and anxiety (-2.43, 95%CI [-3.72, -1.14]). Similarly, Lumen-coached participants had significantly greater improvements in problem-solving abilities, worry, positive and negative affect, disability, and work productivity, but not in the secondary neural targets or dysfunctional attitudes, compared with waitlist control. Lumen- and human-coached PST did not differ significantly for any of the measures; additionally, improved problem-solving abilities correlated with improved symptoms for both modalities. One serious adverse event involving hospitalization, unrelated to the study, was detected.

Conclusions: Although differences in neural target changes were null, a voice-based AI coach improved measures of symptoms, behavior, affect, and functioning. Although Lumen showed promise compared to human delivery, confirmatory noninferiority testing is warranted.

Trial Registration: NCT05603923

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. For example, AI-based chatbots have the potential to replicate human-like conversations in a variety of settings, with studies showing their promise in mitigating depressive and anxiety symptoms;[8] however, systematic reviews have reported mixed results.[9] With advances in generative AI, such chatbots have gained additional attention, primarily related to safety concerns regarding their use and the lack of rigorous outcomes research.[10,11] However, a recent randomized trial using a generative AI-based chatbot, Therabot, found high engagement, acceptability, and meaningful improvements in clinical-level symptoms for major depressive disorder, generalized anxiety disorder (GAD), and eating disorders.[8] Compared with chatbots, 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]

Developed using an iterative user-centered design process,[15,16] Lumen is a voice-based AI coach delivering the clinically proven problem-solving therapy (PST) for depression and anxiety.[17] 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 target for cognitive control (right dorsolateral prefrontal cortex, R dlPFC) and symptom improvements meeting the prespecified "go criteria" for additional testing.[18]

Aligned with the experimental therapeutics approach prioritizing mechanism-targeted intervention development and testing,[14] the objectives of this larger, Phase 2, 3-arm randomized trial[19] were to investigate (a) whether Lumen-coached PST was better (superior to) than waitlist control and not worse than (noninferior to) human-coached PST on measures of neural mechanistic targets and on patient-reported target, clinical and functional outcome measures; and (b) the mediating effects of neural and patient-reported target measures on clinical symptom outcomes.

Method

Participants

The trial protocol was published.[19] 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 I).

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)[20] and/or moderate anxiety (Generalized Anxiety Disorder [GAD-7] scores 10–14),[21] 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[22] based on Pocock's covariate-adaptive minimization.[23] 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,[24] 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. PST is patient driven, with the coach acting 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.[17] These attributes—inherent structure and pragmaticism—make PST an ideal intervention for delivery using an AI-driven platform via a voice-based modality.

Lumen's design is rule-based, without any generative AI components, and aligned 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,[25] and showed promising signals of neural target engagement and clinical efficacy in a Phase 1 trial[18] (Suppl. Sections B-D).

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,[17] 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 constructs related to cognitive control and emotion reactivity and patient-reported clinical and functional outcome measures.

Neural target measures

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,[18,26,27] the primary region of interest (ROI) was the right dlPFC, and a secondary ROI was the left dlPFC, both engaged using a go/no-go task for cognitive control. Other secondary ROIs as a measure of emotion reactivity included bilateral amygdala engaged by the viewing of threat and sad faces (negative affect) in the non-conscious viewing condition. 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.[28,29]

Patient-reported target measures

As a behavioral measure of potential mediators of the treatment effect on 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).[30] Additional patient-reported measures related to cognitive control and emotion reactivity included the Dysfunctional Attitudes Scale (DAS),[31] Penn State Worry Questionnaire (PSWQ),[32] and the Positive and Negative Affect Schedule (PANAS)[33] (see protocol[19]).

Patient-reported clinical and functional outcome measures

Clinical symptom outcomes were measured using the Hospital Anxiety and Depression Scale (HADS),[34,35] which assessed depressive and anxiety symptom scores and summative overall psychological distress scores. Psychosocial functioning outcomes were measured using the Sheehan's Disability Scale (SDS)[36] and the Work Productivity and Activity Impairment (WPAI) questionnaire.[37]

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).[38] 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,[39,40] 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, clinical and functional outcomes 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 inferiority tests. Statistical significance was determined at $\alpha=0.05$. Multiple testing corrections were not planned for any of the secondary measures, given their exploratory nature. Cohen's d was calculated using the mean difference between two groups divided by the pooled standard deviation (SD).

Using the approach by Kraemer et al.[41] 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.[42]

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 powered the trial for detecting a medium effect of Cohen's d = 0.45, and a sample of 150 (100 Lumen, 50 waitlist control) had 80% power at α=5% (2-sided), assuming at least 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 being on treatment regimens 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. Of the 100 Lumen-coached participants, 89 (89%) completed at least 4 PST sessions (i.e., minimum therapeutic dose) and 73 (73%) completed all PST sessions; of the 50 human-coached

participants, 46 (92%) completed at least 4 PST sessions and 44 (88%) completed all PST sessions.

Neural targets

No mean differences in changes from baseline to 18 weeks in activation of the bilateral dlPFC and amygdala in either the Lumen- or human-coached group, compared with waitlist control, reached statistical significance (Table 2). Thus, noninferiority tests of the Lumen- and human-coached groups were not performed.

Patient-reported target measures

Compared with waitlist controls, Lumen-coached participants had significantly greater increase from baseline to 18 weeks in overall problem-solving ability (between-group mean difference=1.04, 95%CI [0.23, 1.84]) and PPO (1.46, 95%CI [0.21, 2.71]), and a significantly greater reduction in NPO (-1.71, 95%CI [-3.02, -0.39]). Similarly, compared with waitlist controls, Lumen-coached participants had significantly greater improvements in worry (-6.23, 95%CI [-10.32, -2.14]), positive affect (2.47, 95%CI [0.03, 4.91]), and negative affect (-3.47, 95%CI [-5.97, -0.97]), from baseline to 18 weeks. Changes in RPS, ICS, AS, and dysfunctional attitudes did not differ significantly between the Lumen-coached and waitlist control groups (Table 3).

At 18 weeks, human-coached participants had a significantly greater increase from baseline in overall problem-solving ability, and greater reductions in NPO, worry, and negative affect, than waitlist control participants. Changes in PPO, RPS, ICS, AS, dysfunctional attitudes, or positive affect from baseline to 18 weeks did not differ between the human-coached and waitlist control participants (Table 3).

Moreover, Lumen- and human-coached participants had similar magnitudes of changes and the between-group differences did not reach statistical significance across all patient-reported target measures in both ITT (Table 3) and per-protocol noninferiority tests (Suppl, Section G).

Patient-reported clinical and functional outcome measures

At 18 weeks, Lumen-coached participants had significantly greater reductions from baseline in overall psychological distress (between-group mean difference=-3.56, 95%CI [-5.69, -1.43]) and in both depressive (-1.15, 95%CI [-2.25, -0.04]) and anxiety symptoms (-2.43, 95%CI [-3.72, -1.14]) than waitlist controls. At 18 weeks, Lumen-coached participants also had significantly greater reductions in their disability score (-5.08, 95%CI [-7.77, -2.39]), percent impairment at work (-11.87, 95%CI [-20.57, -3.17]), percent overall work productivity loss (-11.70, 95%CI [-21.59, -1.81]), and percent activity impairment (-9.87, 95%CI [-18.13, -1.62]) than waitlist controls. Changes in the percent work time missed did not differ significantly between Lumen-coached and waitlist control participants.

Similarly, human-coached participants had significantly greater reductions than waitlist control participants in all the patient-reported symptom and functioning outcomes, except for the percent work time missed (Table 4).

Moreover, mean differences in change from baseline to 18 weeks were not statistically significant between Lumen- and human-coached groups for clinical or functional outcomes (Table 4; per-protocol: Suppl, Section G).

Mediation of symptom improvement by change in problem-solving ability

As the changes in the neural targets did not differ significantly between either treatment and waitlist control arms, mediation analysis was not performed using neural targets. Given the

theoretical basis of PST, we conducted a mediation analysis on the SPSI-R:S measures of problem-solving and clinical symptom outcomes.

Comparing changes from baseline to 18 weeks between the Lumen-coached and waitlist control groups, improvements in overall problem-solving ability and NPO correlated significantly with reductions in psychological distress (SPSI-R:S overall, β=-0.76, 95%CI [-1.46, -0.07]; NPO, β=0.58, 95%CI [0.2, 0.95]) and anxiety symptoms (SPSI-R:S overall, β=-0.53 95%CI [-0.98, -0.08]; NPO, β=0.39, 95%CI [0.15, 0.63]). Improvement in PPO correlated significantly with a reduction in depressive symptoms (β=-0.19, 95%CI [-0.38, 0]) (Suppl, Section H).

Similarly, comparing changes from baseline to 18 weeks between the human-coached and waitlist control groups, improvements in overall problem-solving ability and NPO correlated significantly with reductions in psychological distress and anxiety symptoms (Suppl, Section H).

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 F). There were no deaths.

Discussion

This 3-arm RCT evaluated the mechanisms and 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 nor human-coached PST resulted in significant changes in the a priori primary (R dlPFC) or secondary neural targets related to the mechanistic constructs of cognitive control and emotion

reactivity, compared with waitlist control. However, Lumen-coached PST resulted in significant improvements in several patient-reported target measures related to cognitive control and emotion reactivity, including overall problem-solving ability, positive and negative problem orientation, worry, and positive and negative affect, compared with waitlist control. Additionally, Lumen-coached PST also resulted in significant improvements in patient-reported clinical and functional outcomes, including overall psychological distress, depressive and anxiety symptoms, disability, work and activity impairments, and work productivity. Moreover, Lumen- and human-coached PST did not differ significantly on any of the patient-reported target, clinical and functional outcome measures.

This study did not replicate the Phase I trial finding of neural target engagement (i.e., R dlPFC) for cognitive control by Lumen-coached PST, and the findings were also non-significant for human-coached PST. The null findings in this study may be due to the relatively higher baseline values of dlPFC activity in both active treatment groups relative to the waitlist control group. Furthermore, there may be 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 cognitive control circuit function involving multiple cortical regions was associated with improved problem-solving ability and depressive symptoms following human-coached PST.[43] Further studies utilizing a data-driven, whole-brain network approach may yield insights regarding neural target engagement for Lumen- and human-coached PST.

The significantly greater improvements in the patient-reported target measures, particularly overall problem-solving ability and negative and positive problem orientation, in the Lumen-coached versus waitlist control group, and their significant associations with improved overall psychological distress, depressive and anxiety symptoms align with PST's theory-based mechanism of action. Similarly, improved problem-solving ability with a less negative problem orientation also mediated improved overall psychological distress and anxiety symptoms for human-coached PST. Furthermore, the results for the clinical symptom and functional outcomes highlight the superiority of Lumen-coached PST to waitlist control and potential noninferiority to human-coached PST, although the latter needs to be confirmed with prespecified noninferiority margins in full-scale trials. The consistency of these findings across both delivery modalities of the same intervention supports the therapeutic utility of Lumen-coached PST.

To the best of our knowledge, Lumen is the first voice-based AI coach for mental health treatment validated using an experimental therapeutics approach—focusing on mechanisms of action in the development and testing of new interventions—in trials.[14] Towards this end, theory-based, plausible behavioral and neural mechanistic targets were investigated. The use of human-coached PST as an active comparator was a notable strength of this trial.[44] Within the framework of a "device as a coach," the integrated architecture and delivery mechanism of Lumen allows for clinical translation and monitoring.[45]

Lumen's framework prioritizes safety by relying on a rule-based, PST-aligned design and extending the AI-based features on the Alexa platform, to avoid chances of device hallucinations, a major concern among current AI-based chatbots.[11] A key aspect of the Lumen development process was the balance between PST treatment fidelity – validated with PST

trainers – and AI features. As such, Lumen's framework relied on extending the AI features of an existing platform (Alexa) for PST delivery and developing features for seamless user interactions (e.g., pausing and rescheduling sessions). Additionally, the long arc of Lumen development and associated clinical trials – spanning approximately 6 years – prevented the ability to change the framework (e.g., using generative AI). With the recent evidence regarding the potential for generative AI chatbots for improving mental health outcomes,[8] we are exploring changes to Lumen.

This study has several limitations. It had a relatively small sample and was limited to those with clinically significant, but non-severe depression and/or anxiety. This study excluded potential participants who may have benefited due to study design and feasibility considerations (e.g., unable to undergo fMRI scans, internet access, non-English speakers). Despite high retention overall, 24% participants were missing acceptable fMRI data, which limits the validity of the neural target analyses. No multiple testing adjustments were made for secondary target and outcome measures, and the results warrant confirmation in future studies.

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.

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. 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

9. 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

10. Metz C. Are A.I. Therapy Chatbots Safe to Use? *The New York Times*. November 6. Accessed January 3, 2026. https://www.nytimes.com/2025/11/06/technology/ai-therapy-chatbots-ash.html

11. 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#

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. 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

15. 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.

16. 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

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

18. 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

19. 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

20. 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

21. 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

22. 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

23. 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

24. 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.

25. 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

26. 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

27. 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

28. 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

29. 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.

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

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

32. 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

33. 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

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

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

36. 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

37. 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

38. 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

39. 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

40. 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

41. 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

42. 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

43. 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

44. 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

45. 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

Figure Legends

Figure 1. Consort chart.

Tables

Table 1. Baseline Characteristics[a]

Characteristic	All participants (n=200)	Lumen (n=100)	Waitlist (n=50)	Human (n=50)
Age, years [b]	36.6 ± 11.9	36.1 ± 11.8	37.1 ± 12.4	37.0 ± 11.9
Female, % [b]	155(77.5)	77(77.0)	39(78.0)	39(78.0)
Race/Ethnicity, % [b]				
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, % [b]				
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, % [b]				
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 [b]	12.3 ± 3.5	12.1 ± 3.6	12.7 ± 3.1	12.2 ± 3.6
GAD-7 score [b]	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.
[a]Values are mean ± SD unless noted otherwise.
[b]Prognostic factors for randomization: age, sex, race/ethnicity, education, digital health literacy, PHQ-9, and GAD-7.

Table 2. Treatment effects on neural targets[ab]

					Superiority test					
		Unadjusted mean ± SD			Lumen versus Waitlist			Human versus Waitlist		
Circuit	**Target Measure**	**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)**
Cognitive Control	dlPFC R	n=71	n=36	n=34						
	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
	dlPFC L	n=71	n=36	n=34						
	baseline	0.14 ± 0.43	0.07 ± 0.41	0.02 ± 0.38						
	change at 18 weeks	-0.07 ± 0.58	0.06 ± 0.47	0.00 ± 0.46	-0.08 (-0.3, 0.15)	0.5	-0.14	0.06 (-0.17, 0.28)	0.62	0.12
Negative Affect (Non-conscious)	Amygdala R	n=64	n=30	n=32						
	baseline	0.02 ± 0.16	0.02 ± 0.17	0.04 ± 0.19						
	change at 18 weeks	0.01 ± 0.24	0.04 ± 0.27	-0.07 ± 0.27	0.08 (-0.03, 0.19)	0.15	0.32	0.11 (-0.03, 0.24)	0.13	0.39
	Amygdala L	n=64	n=30	n=32						
	baseline	0.02 ± 0.17	0.01 ± 0.20	0.03 ± 0.18						
	change at 18 weeks	0.01 ± 0.29	0.05 ± 0.32	-0.02 ± 0.28	0.03 (-0.09, 0.16)	0.6	0.11	0.07 (-0.08, 0.22)	0.36	0.23

[a]*t* tests.

[b]Values are mean ± SD unless otherwise noted.

Table 3. Treatment effects on patient-reported target measures

				Superiority test						Noninferiority test		
	Unadjusted mean ± SD			Lumen versus Waitlist			Human versus Waitlist			Lumen versus Human		
Measure	**Lumen**	**Human**	**Waitlist**	**Model-based mean difference (95%CI)[a]**	**P value (2-sided)[a]**	**Cohen's d[b] (Lumen - Waitlist**	**Model-based mean difference (95%CI)[a]**	**P value (2-sided)[a]**	**Cohen's d[b] (Human- Waitlist**	**Model-based mean difference (90%CI)[a]**	**P value (1-sided[a]**	**Cohen's d[b] (Human -Lumen)**
SPSI-R:S raw score[c]	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[c]	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[c]	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[c]	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[c]	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[c]	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
Dysfunctional Attitudes Scale (DAS)[d]	n=96	n=48	n=49									
baseline	133.8 ± 35.76	143.4 ± 33.17	140.9 ± 34.24									
change at 18 weeks	-14.7 ± 29.02	-20.5 ± 40.92	-11.1 ± 24.40	-5.48 (-15.45, 4.49)	0.28	0.13	-7.95 (-19.47, 3.57)	0.17	0.28	2.47 (-5.97, 10.92)	0.36	0.17
Penn State Worry Questionnaire (PSWQ)[e]	n=97	n=48	n=49									
baseline	59.99 ± 11.21	59.96 ± 11.19	61.44 ± 11.35									
change at 18 weeks	-8.35 ± 12.87	-9.38 ± 15.52	-2.61 ± 10.62	-6.23 (-10.32, -2.14)	0.003	0.47	-6.99 (-11.72, -2.25)	0.004	0.51	0.75 (-2.7, 4.2)	0.39	0.07
Positive and Negative Affect Schedule (PANAS)[f]												
Positive	n=97	n=48	n=49									
baseline	26.90 ± 6.96	26.10 ± 6.07	26.26 ± 7.44									
change at 18 weeks	6.04 ± 7.55	6.46 ± 8.01	3.61 ± 7.03	2.47 (0.03, 4.91)	0.048	0.33	2.69 (-0.14, 5.52)	0.06	0.38	-0.23 (-2.29, 1.83)	0.44	0.05
Negative	n=97	n=48	n=49									
baseline	27.22 ± 7.39	27.86 ± 7.17	26.68 ± 7.24									
change at 18 weeks	-4.04 ± 7.66	-7.19 ± 9.05	-0.22 ± 7.47	-3.47 (-5.97, -0.97)	0.007	0.50	-6.24 (-9.15, -3.34)	<.001	0.84	2.78 (0.67, 4.89)	0.14	0.39

Abbreviations: AS, avoidant problem-solving style; CI, confidence interval; DAS, Dysfunctional Attitudes Scale; ICS, impulsive/careless problem-solving style; NPO, negative problem orientation; PSWQ, PANAS, Positive and Negative Affect Schedule; PPO, positive problem orientation; PSWQ, Penn State Worry Questionnaire; RPS, rational problem-solving style; SPSI-R:S, Social Problem-solving Index-Revised Short Form.

[a]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.

[b]The mean difference between two groups divided by the pooled standard deviation

[c]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.
[d]DAS (Form A) is a 40-item self-reported scale that measures the presence and intensity of dysfunctional attitudes. Each item is rated a 7-point Likert scale (7 = fully agree; 1 = fully disagree). The higher the sum of the 40-items, the more dysfunctional attitudes an individual possesses.
[e]PSWQ is a self-reported, 16-item, Likert-type scale that measures the trait of worry. Each item is related on a 1-5 scale. The total score of the scale ranges from 16 to 80, with higher scores indicating higher trait worry.
[f]PANAS consists of two 10-item self-reported scales to measure positive and negative affect. Each item asks about the extent one has felt a positive or negative feeling on a 5-point scale of 1 (not at all) to 5 (very much). Positive affect scores can range from 10 – 50, with higher scores representing higher levels of positive affect. Negative affect scores can range from 10 – 50, with lower scores representing lower levels of negative affect.

Table 4. Treatment effects on patient-reported clinical and functional outcome measures

				Superiority test						Noninferiority test		
	Unadjusted mean ± SD			Lumen versus Waitlist			Human versus Waitlist			Lumen versus Human		
Measure	**Lumen**	**Human**	**Waitlist**	**Model-based mean difference (95%CI)[a]**	**P value (2-sided)[a]**	**Cohen's d[b] (Lumen - Waitlist)**	**Model-based mean difference (95%CI)[a]**	**P value (2-sided)[a]**	**Cohen's d[b] (Human-Waitlist)**	**Model-based mean difference (90%CI)[a]**	**P value (1-sided)[a]**	**Cohen's d[b] (Human-Lumen)**
HADS_Total[c]	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
HADS_Depression[d]	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
HADS_Anxiety[d]	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
Sheehan Disability Scale (SDS)[e]	n=96	n=48	n=49									
baseline	17.81 ± 6.75	18.62 ± 6.73	19.42 ± 6.74									
change at 18 weeks	-5.02 ± 9.17	-7.40 ± 9.52	-0.86 ± 7.84	-5.08 (-7.77, -2.39)	<.001	0.48	-7.1 (-10.2, -4)	<.001	0.75	2.02 (-0.24, 4.28)	0.19	0.26
Work productivity and activity impairment questionnaire (WPAI)[f]												
Percent work time missed	n=79	n=39	n=40									

baseline	6.86 ± 12.22	11.66 ± 19.41	6.46 ± 14.06									
change at 18 weeks	0.57 ± 18.37	-6.19 ± 20.55	0.02 ± 19.48	0.56 (-5.16, 6.28)	0.85	0.03	-1.22 (-7.92, 5.48)	0.72	0.31	0.02 (-0.03, 0.07)	0.33	0.35
Percent impairment at work	n=79	n=39	n=40									
baseline	25.47 ± 24.29	28.84 ± 27.19	33.41 ± 25.24									
change at 18 weeks	-9.11 ± 27.28	-12.31 ± 32.88	-0.75 ± 27.77	-11.87 (-20.57, -3.17)	0.008	0.32	-12.81 (-22.87, -2.75)	0.01	0.17	0.01 (-0.06, 0.08)	0.43	0.11
Percent overall work productivity loss	n=79	n=39	n=40									
baseline	29.23 ± 26.77	34.15 ± 31.4	35.99 ± 27.16									
change at 18 weeks	-7.9 ± 31.27	-13.87 ± 38.09	0.55 ± 34.36	-11.7 (-21.59, -1.81)	0.02	0.26	-13.21 (-24.66, -1.76)	0.02	0.40	0.02 (-0.07, 0.1)	0.41	0.18
Percent activity impairment	n=96	n=47	n=49									
baseline	32.3 ± 26.13	36.6 ± 28.04	38.8 ± 26.93									
change at 18 weeks	-11.04 ± 30	-17.66 ± 32.85	-4.69 ± 26.54	-9.87 (-18.13, -1.62)	0.02	0.22	-14.77 (-24.34, -5.19)	0.003	0.44	0.05 (-0.02, 0.12)	0.23	0.21

Abbreviations: CI, confidence interval; HADS, Hospital Anxiety and Depression Scale; SDS, Sheehan Disability Scale; WPAI, Work productivity and activity impairment questionnaire.

[a]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.

[b]The mean difference between two groups divided by the pooled standard deviation

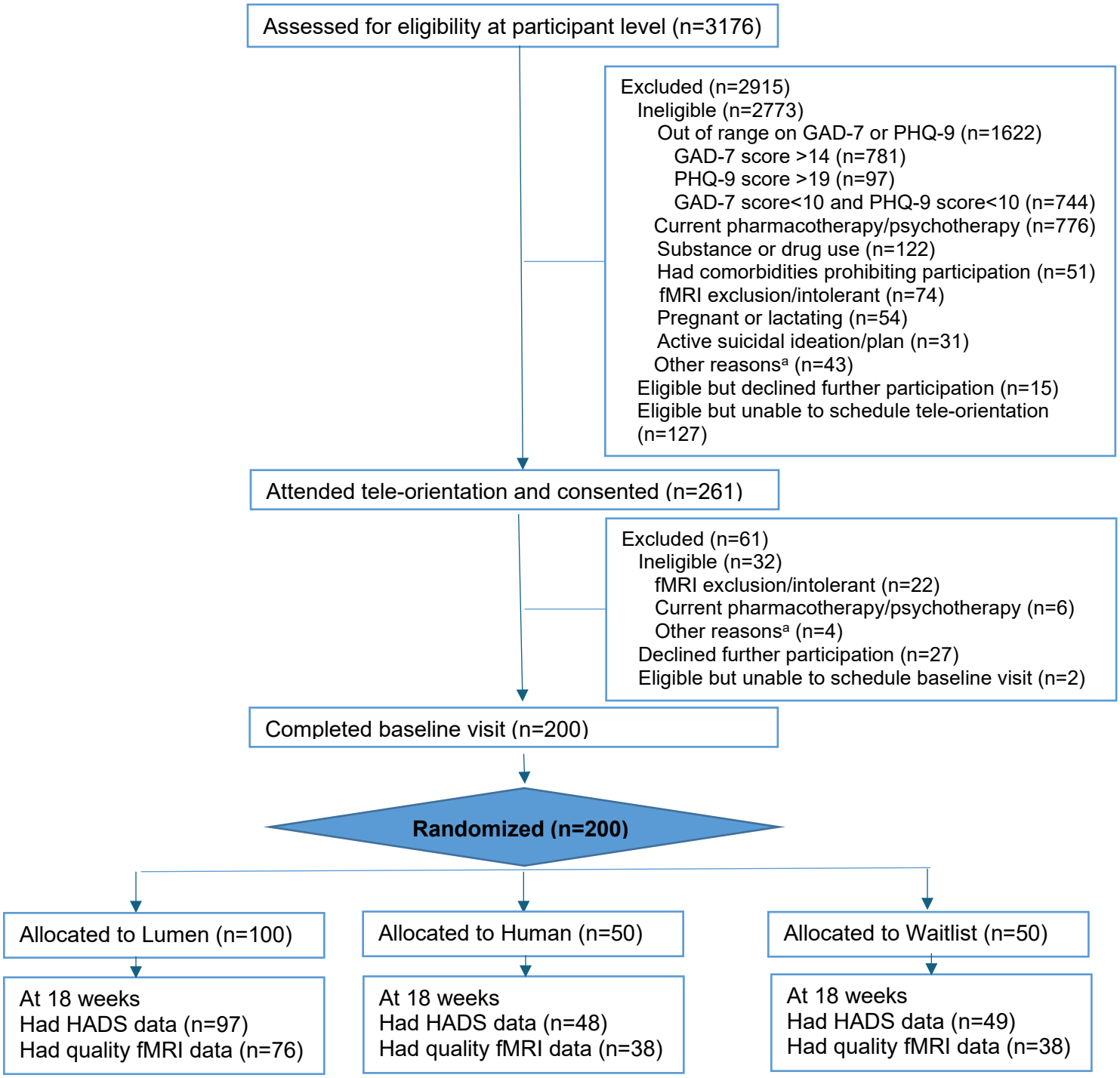
[c]HADS, including 7 questions for anxiety and 7 questions for depression, measures self-reported anxiety and depression in a general medical population of patients. Each item on the questionnaire is scored from 0-3. Total score for the entire scale (emotional distress) ranges from 0 to 42, with higher score indicating more distress.

[d]Scores range from 0-21, with 0-7 = Normal; 8-10 = Borderline abnormal (borderline case); 11-21 = Abnormal (case).

[e]SDS is a validated questionnaire that measures functional disability. Patients rate the extent to which their symptoms impair work/school, social, and family life on a visual analog scale from 0 to 10 and answer the number of days when their symptoms cause them to miss work/school and be unproductive at work/school. Total scores range 0-30, with higher scores indicating greater functional impairment.

[f]WPAI was created as a patient-reported quantitative assessment of the amount of absenteeism, presenteeism and daily activity impairment attributable to general health in the past 7 days.

Figure 1. CONSORT Chart



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