Effect and Mechanisms of a Voice-based Coach using AI on Psychological Distress: 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: 3

Number of Figures: 2

No. of Words (Abstract): 349

No. of Words (Manuscript): 2778

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

Key Points

Question: Can an Artificial Intelligence (AI)-driven, voice-based coach, Lumen, delivering an evidence-based problem-solving treatment (PST), improve psychological distress in adults with depression and/or anxiety better than (superior to) waitlist control and not worse than (noninferior to) human-coached PST?

Findings: We conducted a phase 2, 3-arm randomized clinical trial among 200 participants with untreated, moderate depression and/or anxiety. Compared with waitlist control, Lumen-coached PST resulted in a significantly greater reductions in psychological distress. Treatment effect of Lumen-coached PST was not significantly different from that of human-coached PST.

Meaning: These findings suggest that Lumen-coached PST may be a viable option for patients with moderate 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 have the potential to address the unmet treatment needs among patients with depression and anxiety, but their therapeutic utility is largely unknown.

Objective: To investigate the efficacy and mechanisms of an AI-driven voice-based coach, Lumen, delivering problem-solving treatment (PST) for patients with untreated, moderate depression and/or anxiety.

Design: Phase 2, 3-arm randomized clinical trial.

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

Participants: Adults with a Patient Health Questionnaire-9 score of 10-19 and/or a Generalized Anxiety Disorder Scale score of 10-14 were randomized to Lumen-coached PST (n=100), human-coached PST (n=50), or waitlist control group (n=50) groups.

Interventions: PST was delivered by Lumen on a custom-developed application 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): Hospital Anxiety and Depression Scale was used to assess depression and anxiety symptoms, which summed to an overall psychological distress score (primary outcome), at baseline and at 18 weeks. Measures of mechanisms included problem-solving abilities (using Social Problem-solving Index-Revised Short Form) and a priori neural targets.

Results: Among 200 participants (Mean (SD) age, 36.6 (11.9) years), 77% were women, 25% Black American, and 28% Latino. At 18 weeks, Lumen-coached participants had significantly greater reductions in psychological distress (between-group mean difference=-3.56 [95%CI: -5.69, -1.43], Cohen's d=0.49), depression (-1.15 [95%CI: -2.25, -0.04]; d=0.19), and anxiety symptoms (-2.43 [95%CI: -3.72, -1.14]; d=0.61), compared with waitlist control participants. Mean differences in change from baseline to 18 weeks between Lumen- and human-coached participants were not statistically significant for all 3 symptoms. Both Lumen- and human-coached PST resulted in significant improvements in overall problem-solving ability and negative problem orientation, which were significantly correlated with improved psychological distress and anxiety symptoms. Neural targets did not change significantly with either treatment vs. waitlist control. One serious adverse event involving hospitalization, unrelated to this study, was detected.

Conclusions and Relevance: An AI-driven voice-based coach delivering PST may be a viable treatment option for untreated, moderate depression and/or anxiety. However, clinical efficacy needs to be confirmed in a phase 3 trial.

Trial Registration: clinicaltrials.gov,NCT05603923

Introduction

Depression and anxiety are among the most common mental health disorders in the United States[1] and leading causes of disability worldwide.[2] These conditions are often comorbid, and more than half of those reporting symptoms do not receive any treatment,[3] owing to a myriad economic and socio-cultural factors. Research has long established that people prefer psychotherapies over medications;[4] however, use of proven psychotherapies is limited by low reimbursement, clinician shortage, patients' transportation availability, and stigma.[5,6]

Digital mental health tools, ranging from mobile applications to Artificial Intelligence (AI)-based chatbots, offer new avenues for delivering psychotherapies, with therapeutic fidelity. Text-based chatbots have shown promise in mitigating depression and anxiety symptoms;[7] however, systematic reviews have shown mixed results.[8] Moreover, their safety and long-term sustainability, and challenges for meaningful interactions among subpopulations (e.g., older adults), warrant further study.[9]

Compared with text-based chatbots, voice-based interactions allow for natural, human-like conversations. Developed using an iterative user-centered design process,[10,11] Lumen is a novel AI-enabled, voice-based coach that delivers the clinically proven problem-solving therapy (PST) for depression and anxiety.[12] In a phase 1 randomized trial, compared with waitlist control, Lumen-coached PST resulted in meaningful engagement of an a priori neural target for cognitive control and symptom improvements meeting the prespecified "go criteria" for additional testing.[13]

The objectives of this larger, phase 2, 3-arm randomized trial[14] were to investigate (1) whether Lumen-coached PST is better than (superior to) waitlist control and not worse than (noninferior

to) human-coached PST in improving psychological distress due to symptoms of depression and anxiety, and (2) the plausible behavioral and neural mechanisms of action.

Methods

Participants

The trial protocol was published.[14] 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 moderate or moderately-severe symptoms of depression (Patient Health Questionnaire [PHQ-9] scores 10–19)[15] and/or moderate anxiety (Generalized Anxiety Disorder [GAD-7] scores 10–14)[16] 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 the waitlist control group, using a validated online system[17] based on Pocock's covariate-adaptive minimization.[18] This approach allowed for achieving a better-than-chance marginal balance between the study arms across multiple baseline characteristics including sex, age, race/ethnicity, digital health literacy,[19] PHQ-9 score, and GAD-7 score. Investigators, the 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 platform, Lumen provides 8 PST sessions (4 weekly, then 4 biweekly) for patients with moderate depression and/or anxiety. PST is patient driven, with the coach acting as a guide to identify a problem, set a goal, brainstorm solutions, choose a solution, develop an action plan, and to implement and evaluate the plan in each session.[12] Lumen's design is rule-based, without any generative AI components, and aligned with 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,[20] and showed promising signals of neural target engagement and clinical efficacy in a phase 1 trial.[13] Details regarding Lumen are provided in the supplement (Sections B-D).

Lumen was integrated within the Alexa app on an iPad provided to participants for the study. For each session, participants used a voice invocation – "Open Lumen Session [number]" –and completed the assigned PST session. Between sessions, participants completed surveys, ecological momentary assessments (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,[12] 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

Primary and other patient-reported outcomes

Blinded outcome assessors conducted standardized assessments at baseline and 18 weeks. Measured by the Hospital Anxiety and Depression Scale (HADS),[21,22] total psychological distress was the primary symptom outcome, and depression and anxiety symptoms were secondary outcomes. On the HADS, depression and anxiety symptom scores ranged from 0-21, categorized as normal (0-7), borderline abnormal (8-10), and abnormal (11-21). HADS total scores were the sum of depression and anxiety scores, indicating overall psychological distress.

As a behavioral measure of mediator 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-RS, impulsive/careless-ICS, and avoidant-AS).[23] Each subscale was scored by summing the respective 5 items (each from 0 to 4), and the total score ranged from 0 to 20 (higher the better) by reversing the NPO, ICS, and AS scores and averaging all 5 subscale scores. Additional self-reported measurements are not reported here (see protocol[14]).

Neural target measures

Task-based functional magnetic resonance imaging (fMRI) data were collected at baseline and 18 weeks, utilizing previously established fMRI sequences and parameters (Suppl, Section E). Informed by previous findings,[13,24,25] the primary region of interest (ROI) was the right dorsal lateral prefrontal cortex (dlPFC), and a secondary ROI was the left dlPFC, both engaged using a

go/no-go task for cognitive control. Other secondary ROIs 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., threat versus neutral faces, no-go versus go) was derived in a manner consistent with prior studies.[26,27]

Statistical analysis

We conducted Intention-to-Treat (ITT) analyses for superiority tests (Lumen-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).[28] The analyses of the between-group differences for each primary and secondary outcome included all participants with follow-up data at 18 weeks. Participants were analyzed based on their assigned group.

The between-group differences on changes in HADS scores from baseline to 18 weeks were tested using ordinary least square (OLS) regression with adjustment of baseline values of the outcome measure. Model-adjusted between-group mean differences with 95% confidence intervals (CIs) were reported. Cohen's d was calculated using the mean difference between two groups divided by the pooled standard deviation (SD).

The between-group differences on changes in neural targets from baseline to 18 weeks were assessed using *t*-tests. The neural targets were defined by standardized activation values based on healthy control norms that have been validated.[29,30]

The mediation effects of changes in total SPSI-R:S and its subscale scores on changes in HADS symptom outcomes were tested, employing the approach by Kraemer et al.[31] We defined that mediation exists 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 both the HADS and SPSI-R:S were assessed at 0 and 18 weeks, the focus was on contemporaneous mediation.[32]

Moderation analysis used the same OLS models as described above, plus the main effect of each potential effect modifier (e.g., sex) and its interaction with the treatment vs. waitlist control; the latter, if significant, rejected the null hypothesis of no moderation. These were post-hoc analyses on the potential effect modifiers based on phase 1 trial findings (cite).[13]

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 tests of change in the primary outcome (HADS total psychological distress score) and change in the primary neural target (i.e., right dlPFC for cognitive control) comparing Lumen-coached PST and waitlist control were of primary interest. Multiple testing was not planned as these measures were of distinct types. 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, and thus with a total N=200.

Results

Sample characteristics and retention

Of the 3176 individuals who completed screening, 2773 were ineligible and 142 declined to participate or were unable to participate (Figure 1).

Participants (N=200) had a mean age of 36.6 (SD=11.9) years, 77% were women, 25% were Black American, 28% were Latino, 63% had a high school or college education (1–4+ years), and 47.5% had an annual income less than $55,000 (Table 1). At baseline, randomized 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.

All 200 participants completed baseline data; 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.

Between-treatment differences in psychological distress, depression and anxiety symptoms

At 18 weeks, Lumen-coached participants had a statistically significant decline in total psychological distress scores (-3.56, 95%CI [-5.69, -1.43]), compared with waitlist control participants, with a medium effect size (Cohen's d=0.49). Similarly, Lumen-coached participants had a statistically significant decline in both depression (-1.15, 95%CI [-2.25, -0.04]; d=0.19), and anxiety symptom scores (-2.43, 95%CI [-3.72, -1.14]; d=0.61), compared with waitlist control participants. Human-coached participants had significantly greater declines in all 3 symptom scores than waitlist control participants (Table 2).

At 18 weeks, mean differences from baseline between Lumen- and human-coached participants were not statistically significant for all 3 symptom outcomes (Table 2; per-protocol: Suppl, Section H).

Mediation of symptom improvement by change in problem-solving skills

At 18 weeks, Lumen-coached participants had statistically significant increases in overall problem-solving (1.04, 95%CI [0.23, 1.84]) and PPO scores (1.46, 95%CI [0.21, 2.71]), and a statistically significant decrease in NPO scores [-1.71, 95%CI [-3.02, -0.39]), compared with waitlist control participants. RPS, ICS and AS change scores did not differ significantly by group (Table 3).

Comparing the Lumen-coached and waitlist control groups at 18 weeks, changes in SPSI-R:S total and NPO scores correlated significantly with change in total psychological distress scores (SPSI-R:S -0.76, 95%CI [-1.46, -0.07]; NPO 0.58 95%CI [0.2, 0.95]) and with change in anxiety symptom scores (SPSI-R:S -0.53, 95%CI [-0.98, -0.08]; NPO 0.39, 95%CI [0.15, 0.63]). The PPO score change correlated significantly with change in depression symptom scores (-0.19, 95%CI [-0.38, 0]).

At 18 weeks, human-coached participants had a statistically significant increase in overall problem-solving score (1.11, 95%CI [0.18, 2.04]) and a statistically significant decrease in NPO score (-2.25, 95%CI [-3.77, -0.74]), compared with waitlist control participants. PPO, RPS, ICS and AS change scores did not differ significantly by group (Table 3).

Comparing the human-coached and waitlist control groups at 18 weeks, changes in SPSI-R:S total and NPO scores correlated significantly with change in total psychological distress scores

(SPSI-R:S -0.84 95%CI [-1.56, -0.17]; NPO 0.6 95%CI [0.23, 0.96]) and with change in anxiety symptom scores (SPSI-R:S -0.57, 95%CI [-0.99, -0.15]; NPO 0.4, 95%CI [0.19, 0.61]) (Table 3).

Neural targets

No mean differences in activation changes for the bilateral dlPFC and amygdala between either Lumen- or human-coached group and waitlist controls reached statistical significance (summary in Suppl, Sections E-F).

Moderation analysis

Mean differences in total psychological distress and anxiety symptom scores between the Lumen-coached and waitlist control groups were greater among those <40 years (vs. ≥40), non-white (vs. white) participants, those having college or less education (vs. post-college), and those reporting <$55,000 annual family income (vs.≥$55,000), although none were statistically significant. Similar patterns were observed comparing the human-coached and waitlist control groups, where additionally, males had greater, though statistically nonsignificant, decreases in total psychological distress, depression, and anxiety symptom scores than females (Figure 2).

Additionally, Lumen-coached PST may be less effective than human-coached PST in improving the 3 symptom outcomes among participants with higher total HADS scores at baseline (≥20 vs. <20) (Suppl, Section H).

Adverse events

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

Discussion

In this 3-arm RCT of 200 adults with moderate or moderately-severe depression and/or moderate anxiety, an AI voice-based virtual coach for PST, Lumen, resulted in statistically significant improvements of medium effect in total psychological distress and anxiety symptoms, compared to waitlist control; the effect in the improvement of depression symptoms was smaller, although statistically significant. Moreover, the treatment effects of Lumen-coached PST were not statistically different from that of human-coached PST for all three symptom scores. Both Lumen- and human-coached PST resulted in significant improvements in both overall problem-solving ability and negative problem orientation, which was significantly correlated with improved total psychological distress and anxiety symptoms. Additionally, Lumen-coached PST resulted in significant improvement in positive problem orientation, which correlated significantly with improved depression symptoms. These findings indicate that consistent with the theoretical basis of PST, improved problem-solving skills were the behavioral mechanism explaining improved symptoms for both Lumen- and human-coached delivery.

This trial corroborates the findings from the pilot trial that demonstrated medium-sized benefits of Lumen-coached PST for psychological distress and anxiety, with more favorable improvements among racial and ethnic minorities and those with college or less education.[13] This trial extends the pilot trial, showing comparability of Lumen- to human-coached PST for

symptom management and, more importantly, that improving problem-solving skills is a mechanism of action of PST with both delivery modalities.

The minimum important differences (MID) of the HADS total psychological distress, depression, and anxiety scores varied by study populations and methodologies. We found no MID values for these HADS outcomes specifically in the target study population; however, MIDs of 1.5-1.7 have been reported for patients with chronic cardiovascular or pulmonary diseases, a related population which depression and anxiety are common.[33-35] Accordingly, the mean differences and 95% CIs as compared with waitlist control in this trial suggest that Lumen-coached PST can provide clinically meaningful symptom relief for patients with moderate depression and/or anxiety who otherwise receive no treatment, providing a potentially viable alternative for those with no mental health support. Moreover, the mean differences and 95% CIs as compared with human coaching suggest that Lumen can be an alternative modality for PST delivery, although clinical benefits may be less for patients with higher pre-treatment psychological distress.

Prior findings of engagement of the neural target (i.e., right dlPFC) for cognitive control were not replicated in the present study. Although effect sizes were similar in both studies, this study had stricter statistical criteria for significance. Another factor was the similar activation values for both groups at baseline, seen in the pilot study, were not observed in the current study, which may have affected the ability to detect significant treatment-induced differences in neural targets. Notwithstanding, the null findings do not disprove the intervention's impact on neural targets. Rather than a single ROI-focused approach, data-driven whole-brain network techniques may yield insights of neural target engagement for Lumen- and human-coached PST.[36]

To the best of our knowledge, Lumen is the first AI voice-based application for mental health counseling validated in randomized trials. Moreover, this trial included both human-coached and waitlist control comparator arms, which are often missing in trials of most AI-based digital mental health interventions.[37] Within the framework of a "device as a coach," the integrated architecture and delivery mechanism of Lumen allows for translation for clinical use and monitoring of patient progress.[38] Finally, this integrated framework prioritizes safety over flexibility by using a rule-based PST-aligned design, to avoid chances of device hallucinations or non-therapy related conversations, which is a major concern among current AI-based chatbots.[39]

This study has several limitations. It has a relatively small study sample and is limited to those with moderate, untreated depression and/or anxiety. This study excluded potential participants who may have benefited due to study design and feasibility considerations (e.g., those who could not complete fMRI scans, access to internet, non-English speakers). Despite high retention overall, 24% participants were missing acceptable fMRI data, which limits the validity of the analysis of neural targets.

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, 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. 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

5. 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

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

7. 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

8. 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

9. Enam MDA, Murmu C, Dixon E. "Artificial Intelligence - Carrying us into the Future": A Study of Older Adults' Perceptions of LLM-Based Chatbots. *International Journal of Human–Computer Interaction*. 2025:1-24. doi:10.1080/10447318.2025.2476710

10. 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.

11. 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

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

13. 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

14. 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

15. 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

16. 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

17. 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

18. 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

19. 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.

20. 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

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

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

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

24. 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

25. 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

26. 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

27. 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.

28. 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

29. 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

30. 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

31. 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

32. 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

33. 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(1):46. doi:10.1186/1477-7525-6-46

34. 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

35. Smid DE, Franssen FM, Houben-Wilke S, et al. Responsiveness and MCID Estimates for CAT, CCQ, and HADS in Patients With COPD Undergoing Pulmonary Rehabilitation: A Prospective Analysis. *J Am Med Dir Assoc*. Jan 2017;18(1):53-58. doi:10.1016/j.jamda.2016.08.002

36. 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

37. 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

38. 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

39. 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.

Figure 2. Treatment effect on symptoms, overall and by sub-group.

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)
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
HADS Depression score	7.6 ± 3.2	7.1 ± 3.3	8.3 ± 3.0	7.8 ± 3.1
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)
HADS Anxiety score	11.3 ± 3.5	11.3 ± 3.6	11.3 ± 3.6	11.1 ± 3.1
HADS Total score	18.9 ± 5.4	18.4 ± 5.6	19.7 ± 5.7	19.0 ± 4.5
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)

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. HADS outcomes: Lumen-coached vs. Waitlist control, Human-coached vs. Waitlist control, and Lumen-coached vs. Human-coached

	HADS Total*		HADS Depression†		HADS Anxiety†	
	Unadjusted mean ± SD					
	baseline	change at 18 weeks	baseline	change at 18 weeks	baseline	change at 18 weeks
Lumen	18.38 ± 5.61	-5.85 ± 6.38	7.09 ± 3.32	-2.53 ± 3.22	11.29 ± 3.59	-3.32 ± 4.10
Human	18.98 ± 4.52	-7.33 ± 8.04	7.84 ± 3.05	-3.27 ± 4.47	11.14 ± 3.10	-4.06 ± 4.49
Waitlist	19.68 ± 5.75	-2.76 ± 5.99	8.34 ± 3.01	-1.92 ± 3.33	11.34 ± 3.59	-0.84 ± 3.96
	Lumen versus Waitlist					
Model-based mean difference (95%CI)	-3.56 (-5.69, -1.43)		-1.15 (-2.25, -0.04)		-2.43 (-3.72, -1.14)	
P value	0.001		0.04		<.001	
Cohen's d§	0.49		0.19		0.61	
	Human versus Waitlist					
Model-based mean difference (95%CI)	-4.81 (-7.28, -2.35)		-1.6 (-2.87, -0.33)		-3.22 (-4.72, -1.73)	
P value	<.001		0.01		<.001	
Cohen's d§	0.65		0.34		0.76	
	Lumen versus Human					
Model-based mean difference (95%CI)	1.26 (-0.54, 3.05)		0.45 (-0.47, 1.38)		0.79 (-0.30, 1.88)	
P value	0.25		0.42		0.23	

Abbreviations: HADS, Hospital Anxiety and Depression Scale.
* Scores range from 0-42.
† Scores range from 0-21, with 0-7 = Normal; 8-10 = Borderline abnormal (borderline case); 11-21 = Abnormal (case).
 Regression model adjusted for baseline value of the interest.
§ *t* tests.

Table 3. SPSI outcomes: Lumen-coached vs. Waitlist control, Human-coached vs. Waitlist control, and Lumen-coached vs. Human-coached (Greyed out areas are non-applicable comparisons)

	SPSI-R:S total score*		**PPO score***		**NPO score***		**RPS score***		**ICS score***		**AS score***	
	Unadjusted mean ± SD											
	baseline	change at 18 weeks	baseline	change at 18 weeks	baseline	change at 18 weeks	baseline	change at 18 weeks	baseline	change at 18 weeks	baseline	change at 18 weeks
Lumen	12.01 ± 2.91	1.39 ± 2.63	11.22 ± 3.98	1.91 ± 3.86	9.60 ± 4.62	-2.32 ± 4.09	11.11 ± 4.40	1.48 ± 4.38	5.69 ± 4.27	-0.46 ± 3.98	7.01 ± 4.47	-0.77 ± 4.10
Human	11.08 ± 2.75	1.77 ± 2.53	10.18 ± 3.78	2.19 ± 4.25	10.38 ± 4.05	-3.10 ± 4.21	9.90 ± 4.23	1.56 ± 4.32	6.44 ± 4.00	-0.40 ± 3.99	7.88 ± 4.87	-1.58 ± 4.09
Waitlist	11.40 ± 2.98	0.52 ± 2.26	10.06 ± 4.61	0.94 ± 4.27	10.94 ± 4.37	-1.08 ± 4.06	11.74 ± 4.49	0.04 ± 3.51	5.88 ± 4.01	-0.18 ± 4.01	7.96 ± 4.42	-0.37 ± 4.64
	Path A (treatment effect on SPSI): Lumen versus Waitlist											
Model-based mean difference (95%CI) †	1.04 (0.23, 1.84)		1.46 (0.21, 2.71)		-1.71 (-3.02, -0.39)		1.07 (-0.17, 2.31)		-0.36 (-1.57, 0.86)		-0.77 (-2.05, 0.52)	
P value†	0.01		0.02		0.01		0.09		0.56		0.24	
Cohen's d□	0.34		0.24		0.30		0.35		0.07		0.09	
	Path A (treatment effect on SPSI): Human versus Waitlist											
Model-based mean difference (95%CI)†	1.11 (0.18, 2.04)		1.26 (-0.18, 2.70)		-2.25 (-3.77, -0.74)		0.50 (-0.96, 1.95)		0.04 (-1.37, 1.45)		-1.24 (-2.72, 0.25)	
P value†	0.02		0.09		0.004		0.50		0.95		0.10	
Cohen's d□	0.52		0.29		0.49		0.39		0.05		0.28	
	Path B (SPSI→ HADS Total): Lumen vs. Waitlist											
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**						
Estimate (95% CI)	-0.76 (-1.46,	-0.12 (-0.94,	-0.22 (-0.60,	-0.09 (-0.58,	0.58 (0.20,	0.02 (-0.45,						

	SPSI-R:S total score*		PPO score*		NPO score*		RPS score*		ICS score*		AS score*	
	-0.07)	0.69)	0.17)	0.41)	0.95)	0.48)						
P value	0.03	0.77	0.27	0.73	0.003	0.93						
	Path B (SPSI→ HADS Depression): Lumen vs. Waitlist											
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**						
Estimate (95% CI)	-0.26 (-0.61, 0.10)	-0.17 (-0.58, 0.25)	-0.19 (-0.39, 0)	0.06 (-0.19, 0.30)	0.2 (0, 0.4)	0.01 (-0.24, 0.25)						
P value	0.15	0.43	0.049	0.65	0.06	0.96						
	Path B (SPSI→ HADS Anxiety): Lumen vs. Waitlist											
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**						
Estimate (95% CI)	-0.53 (-0.98, -0.08)	0.08 (-0.45, 0.60)	-0.02 (-0.27, 0.23)	-0.15 (-0.46, 0.17)	0.39 (0.15, 0.63)	0 (-0.30, 0.29)						
P value	0.02	0.77	0.89	0.36	0.002	0.99						
	Path B (SPSI→ HADS Total): Human vs. Waitlist											
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**						
Estimate (95% CI)	-0.84 (-1.56, -0.13)	-0.57 (-1.56, 0.42)	-0.2 (-0.6, 0.2)	-0.42 (-0.99, 0.15)	0.60 (0.23, 0.96)	0.43 (-0.09, 0.95)						
P value	0.02	0.26	0.33	0.15	0.002	0.10						
	Path B (SPSI→ HADS Depression): Human vs. Waitlist											
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**						
Estimate (95% CI)	-0.28 (-0.69, 0.12)	-0.51 (-1.07, 0.04)	-0.18 (-0.41, 0.04)	-0.16 (-0.48, 0.16)	0.20 (-0.02, 0.42)	0.28 (-0.04, 0.59)						
P value	0.17	0.07	0.11	0.32	0.08	0.08						
	Path B (SPSI→ HADS Anxiety): Human vs. Waitlist											
	Main effect	**Interaction**	**Main effect**	**Interaction**	**Main effect**	**Interaction**						
Estimate	-0.57	-0.03	-0.01	-0.26	0.40	0.15						

	SPSI-R:S total score*		PPO score*		NPO score*		RPS score*		ICS score*		AS score*	
(95% CI)	(-0.99, -0.15)	(-0.61, 0.55)	(-0.25, 0.22)	(-0.59, 0.08)	(0.19, 0.61)	(-0.15, 0.44)						
P value	0.01	0.93	0.90	0.13	<.001	0.34						

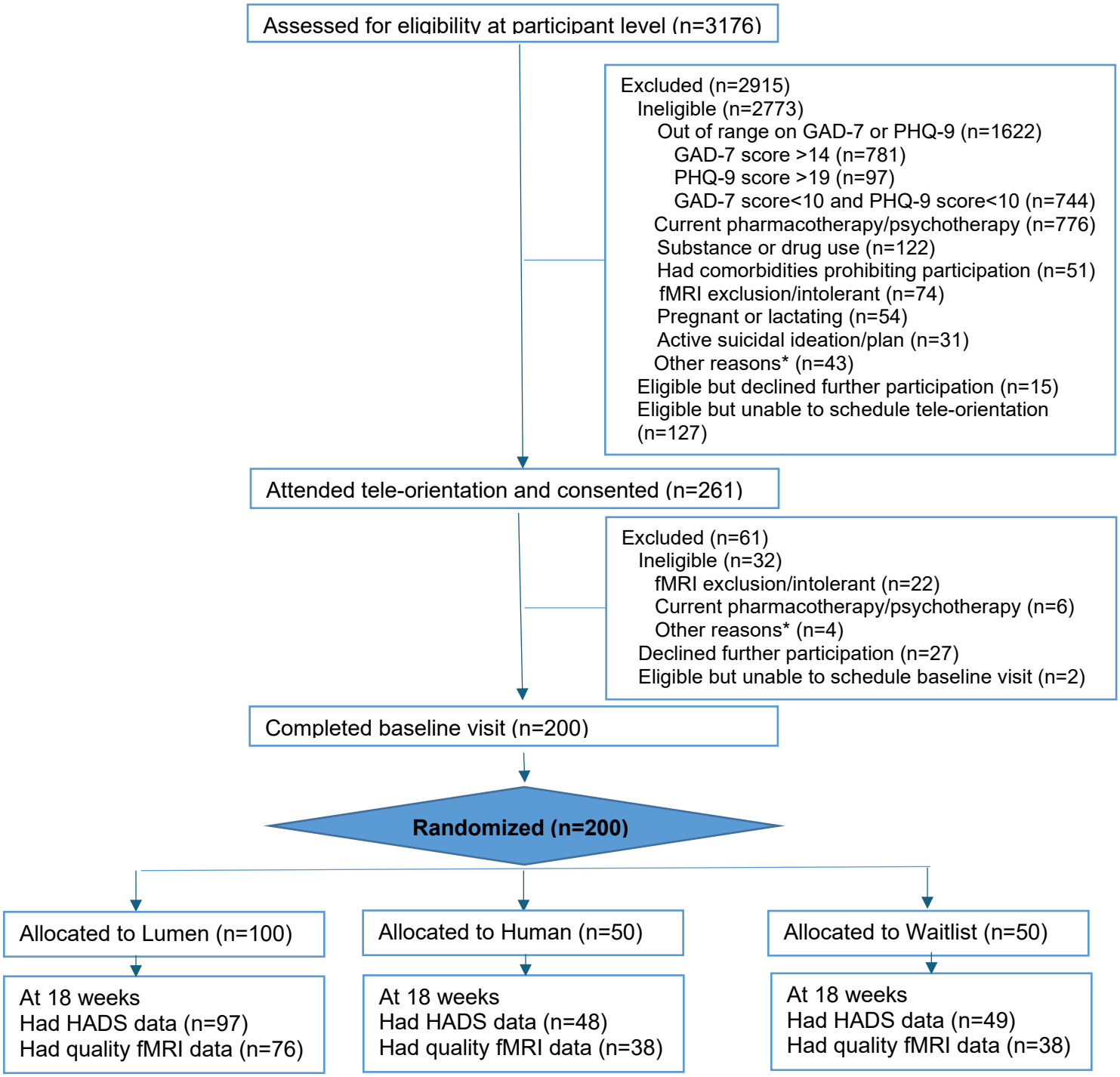
Abbreviations: SPSI-R:S, Social Problem-solving Index-Revised Short Form; PPO, positive problem orientation; NPO, negative problem orientation; RPS, rational problem-solving style; ICS, impulsive/careless problem-solving style; AS, avoidant problem-solving style.

* SPSI-R:S total score=(PPO score/5)+(20- NPO score)/5+ (RPS score/5)+(20- ICS score)/5+(20- AS 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.

† Regression model adjusted for baseline value of the interest.

 t tests.

Figure 1. CONSORT Chart

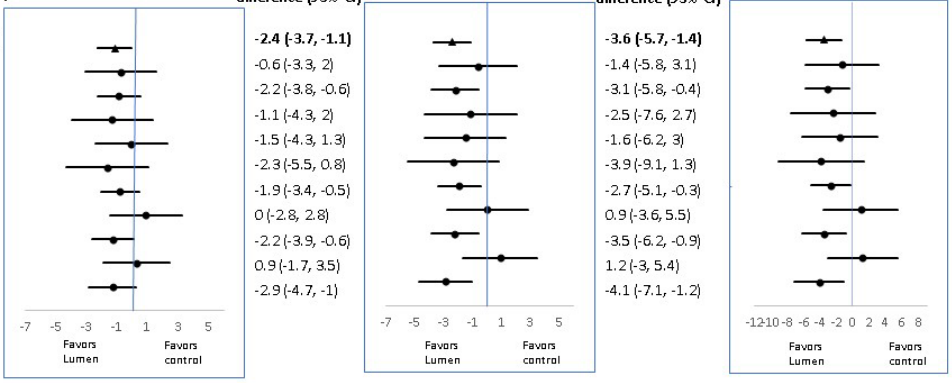


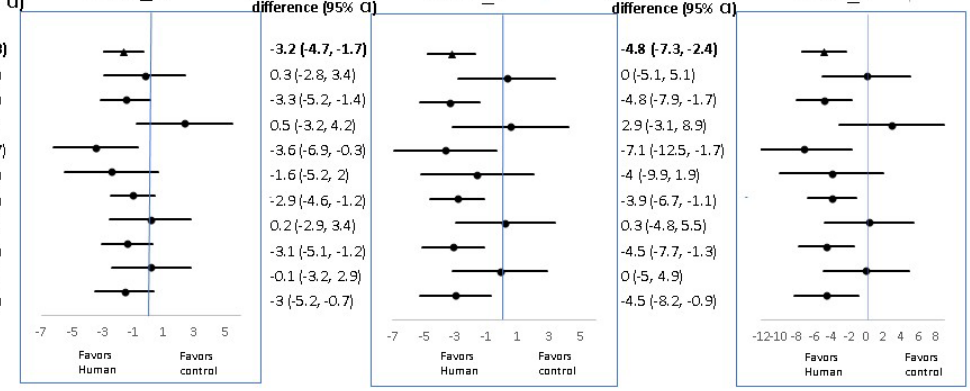
* Other reasons include relocation, not fluent in English, no Wi-Fi, no mobile, no text, family member in study, did not complete baseline survey.

Figure 2. Treatment effect on symptoms, overall and by sub-group

Lumen versus Waitlist

		n	Adjusted mean difference (95% CI) HADS_D	Adjusted mean difference (95% CI) HADS_A	Adjusted mean difference (95% CI) HADS_T
Overall:		**146**	**-1.2 (-2.3, -0.04)**	**-2.4 (-3.7, -1.1)**	**-3.6 (-5.7, -1.4)**
Age:	≥40	52	-0.7 (-3, 1.6)	-0.6 (-3.3, 2)	-1.4 (-5.8, 3.1)
	18-<40	94	-0.9 (-2.3, 0.5)	-2.2 (-3.8, -0.6)	-3.1 (-5.8, -0.4)
Female:	Yes	114	-1.3 (-4, 1.3)	-1.1 (-4.3, 2)	-2.5 (-7.6, 2.7)
	No	32	-0.1 (-2.4, 2.3)	-1.5 (-4.3, 1.3)	-1.6 (-6.2, 3)
Race/ethnicity:	Non-Hispanic White	31	-1.6 (-4.3, 1)	-2.3 (-5.5, 0.8)	-3.9 (-9.1, 1.3)
	Non-White	115	-0.8 (-2, 0.5)	-1.9 (-3.4, -0.5)	-2.7 (-5.1, -0.3)
Education:	Post College	49	0.9 (-1.5, 3.2)	0 (-2.8, 2.8)	0.9 (-3.6, 5.5)
	College or Less	91	-1.3 (-2.6, 0.1)	-2.2 (-3.9, -0.6)	-3.5 (-6.2, -0.9)
Income:	≥$55,000	73	0.3 (-1.9, 2.5)	0.9 (-1.7, 3.5)	1.2 (-3, 5.4)
	<$55,000	73	-1.3 (-2.8, 0.3)	-2.9 (-4.7, -1)	-4.1 (-7.1, -1.2)



Human versus Waitlist

		n	Adjusted mean difference (95% CI) HADS_D	Adjusted mean difference (95% CI) HADS_A	Adjusted mean difference (95% CI) HADS_T
Overall:		**97**	**-1.6 (-2.9, -0.3)**	**-3.2 (-4.7, -1.7)**	**-4.8 (-7.3, -2.4)**
Age:	≥40	37	-0.2 (-2.9, 2.4)	0.3 (-2.8, 3.4)	0 (-5.1, 5.1)
	18-<40	60	-1.5 (-3.1, 0.1)	-3.3 (-5.2, -1.4)	-4.8 (-7.9, -1.7)
Female:	Yes	77	2.4 (-0.8, 5.5)	0.5 (-3.2, 4.2)	2.9 (-3.1, 8.9)
	No	20	-3.4 (-6.2, -0.7)	-3.6 (-6.9, -0.3)	-7.1 (-12.5, -1.7)
Race/ethnicity:	Non-Hispanic White	22	-2.4 (-5.5, 0.6)	-1.6 (-5.2, 2)	-4 (-9.9, 1.9)
	Non-White	75	-1.1 (-2.5, 0.4)	-2.9 (-4.6, -1.2)	-3.9 (-6.7, -1.1)
Education:	Post College	37	0.1 (-2.5, 2.7)	0.2 (-2.9, 3.4)	0.3 (-4.8, 5.5)
	College or Less	57	-1.4 (-3.1, 0.2)	-3.1 (-5.1, -1.2)	-4.5 (-7.7, -1.3)
Income:	≥$55,000	53	0.2 (-2.4, 2.7)	-0.1 (-3.2, 2.9)	0 (-5, 4.9)
	<$55,000	44	-1.6 (-3.5, 0.3)	-3 (-5.2, -0.7)	-4.5 (-8.2, -0.9)