Combining AI and human support in mental health: a digital intervention with comparable effectiveness to human-delivered care

Clare E Palmer¹, Emily Marshall¹, Edward Millgate¹, Graham Warren¹, Michael P. Ewbank¹, Elisa Cooper¹, Samantha Lawes¹, Malika Bouazzaoui¹, Alastair Smith¹, Chris Hutchins-Joss¹, Jessica Young¹, Morad Margoum², Sandra Healey², Louise Marshall¹, Shaun Mehew¹, Ronan Cummins¹, Valentin Tablan¹, Ana Catarino¹, Andrew E Welchman¹ and Andrew D Blackwell¹

1ieso Digital Health, The Jeffrey’s Building, Cowley Road, Cambridge, CB4 0DS, UK
2Dorset HealthCare University NHS Foundation, Sentinel House, Nuffield Industrial Estate, Nuffield Road, Poole, UK

Abstract

Escalating global mental health demand exceeds existing clinical capacity. Scalable digital solutions will be essential to expand access to high-quality mental healthcare. This study evaluated the effectiveness of a digital intervention to alleviate mild, moderate and severe symptoms of generalized anxiety. This structured, evidence-based program combined an Artificial Intelligence (AI) driven conversational agent to deliver content with human clinical oversight and user support to maximize engagement and effectiveness. The digital intervention was compared to three propensity-matched real-world patient comparator groups: i) waiting control; ii) face-to-face cognitive behavioral therapy (CBT); and iii) remote typed-CBT. Endpoints for effectiveness, engagement, acceptability, and safety were collected before, during and after the intervention, and at one-month follow-up. Participants (n=299) used the program for a median of 6 hours over 53 days. There was a large clinically meaningful reduction in anxiety symptoms for the intervention group (per-protocol (n=169); change on GAD-7 = −7.4, d= 1.6; intention-to-treat (n=299); change on GAD-7 = −5.4, d= 1.1) that was statistically superior to the waiting control, non-inferior to human-delivered care, and was sustained at one-month follow-up. By combining AI and human support, the digital intervention achieved clinical outcomes comparable to human-delivered care while significantly reducing the required clinician time. These findings highlight the immense potential of technology to scale effective evidence-based mental healthcare, address unmet need, and ultimately impact quality of life and economic burden globally.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Combining AI and human support in mental health

Introduction

Mental health conditions are the economic and healthcare challenge of our time. Globally, one in eight people live with a mental health condition (1), yet only one in four who require treatment receive it (2). Advances in technology and widespread internet access have been pivotal in increasing access to high-quality mental healthcare. However, one-to-one mental healthcare is inherently limited in its ability to meet the rising mental health demand, and there remains a significant shortage of therapists: there are only four psychiatrists per 100,000 people globally (3), and 58% of the US population live within a health workforce shortage area (4). Technology is primed to enable massive scaling of mental health interventions to increase both access and quality of support worldwide (5).

Rapid advances in computing and Artificial Intelligence (AI) in recent years have led to a rise in the development of digital interventions aiming to solve this scalability problem, and there are an estimated 10,000–20,000 smartphone applications available for mental health support (6,7). These solutions have the potential to enable timely access to support when needed, negate the logistical challenges of attending regular appointments, offer greater patient choice, and reduce burden on therapists and healthcare services (8). However, real-world usage, and in turn effectiveness, of many digital mental health solutions – most of which are self-led – has been poor (9–11). Despite a reported willingness of patients to adopt smartphone applications (12), one month retention rates are typically under 6% (13). Moreover, a recent meta-analysis of mental health applications for symptoms of anxiety and depression found a small pooled clinical effect size ($g=0.26$) and highlighted that only 48% delivered content based on Cognitive Behavioral Therapy (CBT) principles (14) – a “gold-standard” evidence-based approach for anxiety and depression (15). Improving access is crucial, but equally vital is ensuring the support available to patients is engaging and effective.

NHS Talking Therapies (NHS TT, formerly IAPT) is a world-leading initiative designed to increase access to and improve delivery of mental health treatment in the UK. Fundamental to the success of NHS TT is systematic outcomes-monitoring, use of evidence-based treatment protocols, and an appropriately trained and supervised workforce (16). The acceleration of telehealth and expansion of care delivery through digital platforms (e.g. typed conversations) has also enabled insights into the relationship between the active components of evidence-based treatments and clinical outcomes (17,18). Combining this approach with the scalable, systematic delivery of evidence-based protocols through digital tools, offers the opportunity to reduce heterogeneity across the provision of mental healthcare worldwide, and accelerate large-scale scientific research to further enhance treatment quality and personalization (19). High-quality, accessible digital mental healthcare has the potential to maximize impact globally by both improving patient quality of life and reducing the growing economic burden of mental health on health systems and society (20,21).

In this study, we evaluated a digital program that uses this approach to alleviate mild, moderate and severe symptoms of generalized anxiety in adults. The program was designed to maximize engagement and effectiveness by using i) a structured evidence-based program drawing on principles from traditional CBT (22) including third wave approaches i.e. Acceptance and Commitment Therapy (ACT) (23), and ii) an AI-powered conversational agent to deliver the program content in a personalized way. In addition, a dedicated human clinical and user support service was designed to wrap around the digital program, following previous research that human support significantly improves engagement with digital interventions (12,24). This service was developed to provide appropriate support while maintaining the scalability of the digital solution.

This study aimed to measure engagement, clinical effectiveness, acceptability and safety of this digital intervention. Evidence of the effectiveness of a digital intervention is often established through the comparison between the intervention and a waitlist control or self-led non-digital treatment only. However, if digital programs are to provide a scalable solution to global mental health need, we should expect them to provide comparable effectiveness to current standards of care. In this pragmatic, prospective single-intervention arm study, we compared the digital program against propensity-matched external control data from three groups of real-world NHS patients: i) a waiting control with no intervention; ii) patients receiving human-delivered face-to-face CBT; and iii) patients receiving human-delivered typed-CBT. While 1:1 face-to-face therapy serves as the gold-standard for comparison, 1:1 typed-therapy provides a more analogous comparison to the digital program under evaluation where content is predominantly delivered through written communication with the conversational agent. This study design allowed us to evaluate the comparative clinical effectiveness of the digital intervention to human-delivered standard care.
Palmer et al., (2024) – Combining AI and human support in mental health

Methods

Study design

This was a pragmatic, single-intervention arm study with multiple external control groups to measure the engagement, clinical effectiveness, acceptability and safety of a digital program to alleviate symptoms of generalized anxiety in a sample of 300 UK participants. This study was conducted by ieso Digital Health ("ieso", https://www.iesogroup.com/), an outpatient service provider within NHS TT delivering 1:1 human-delivered CBT via a typed modality to treat patients with common mental health disorders. The digital program evaluated here (software name: IDH-DP2-001) was developed by ieso as part of a clinical innovation program creating new scalable digital solutions for mental health support. This was an externally controlled trial meaning comparator arms (sometimes referred to as synthetic control arms) were generated through 1:1 propensity-matching of participants with real-world patients. External propensity-matched control groups were generated to evaluate the digital intervention in comparison to no intervention (i.e. waiting control), face-to-face CBT (gold-standard benchmark), and typed-CBT. This latter group provides an important comparator as it is an example of human-delivered care that closely mirrors the written content delivery within the digital program.

The intervention was delivered via a smartphone application (iPhone & Android). Following an initial clinical assessment with a qualified clinician, eligible participants downloaded the software on their personal smartphone and completed the program in their own time and according to a defined schedule. Participants were required to complete the six-module program within nine weeks. Clinical outcomes were collected as part of the program prior to each module (up to six time-points). Additional study endpoints were collected using validated questionnaires prior to the intervention, during the intervention (following completion of module three activities), after the intervention (completion or when nine-week time limit reached) and at one-month follow-up.

At the point of consent, all participants were asked if they were willing to participate in interviews with additional compensation offered. The sub-sample (based on first-come-first-served sign-up for available interview slots) attended a semi-structured interview pre- and post-intervention to gather qualitative insights into the experience, acceptability and perceived safety of the digital program. Findings on acceptability of the digital program from detailed qualitative analysis of these interviews are reported in a separate publication.

The study was pre-registered (ISRCTN ID: 52546704) and obtained ethical approval prior to recruitment (IRAS ID: 327897, NHS Research Ethics Committee: West of Scotland REC 4). The trial design and participant CONSORT flowchart (25) are summarized in Figure 1. In line with the Declaration of Helsinki, all participants provided signed informed consent and were debriefed following the study.

Study endpoints

Anxiety and mood symptoms were measured before and after the intervention, as well as at the beginning of each module within the program (maximum 6 symptom check-ins) using the Generalized Anxiety Disorder-7 scale (GAD-7) (26) and the Patient Health Questionnaire (PHQ-9) (27) scale. The Work and Social Adjustment Scale (WSAS) (28) and the inflexibility scale (30 items) of the Multidimensional Psychological Flexibility Inventory (MPFI) (29) were collected pre-intervention, during (at the program mid-point) and post-intervention, as measures of functioning and psychological inflexibility, respectively. The following validated self-report measures were collected only at post intervention: the User Engagement Scale (UES) (30), the System Usability Scale (SUS) (31), and the Service-User Technology Acceptability Questionnaire (SUTAQ) (32). A qualitative feedback survey was also administered post-intervention and at one-month follow-up. Demographic data were collected at enrolment and are summarized in Table 1. Findings from the SUS, UES, SUTAQ, MPFI, feedback surveys and qualitative data from pre- and post-intervention semi-structured interviews are reported in a separate publication.

Description of Digital Intervention

The intervention consisted of a six-module digital program (‘ieso Digital Program’; software name: IDH-DP2-001) that used a conversational agent to guide participants through a pre-defined set of activities with human clinical oversight and user support. The program was intended as a first-line intervention for people primarily presenting with anxiety symptoms. The program was designed based on cognitive behavioral principles from traditional CBT and third wave approaches, such as ACT (33,34) (see Supplementary Table 1 for module details). All of the cognitive and behavioral processes, analogies and examples within the intervention were selected for their specificity in targeting symptoms of generalized anxiety.
Palmer et al., (2024) – Combining AI and human support in mental health

**Figure 1. CONSORT diagram.** Prior to assessment enrolment avenues differed for external recruits (left) and patients referred to ieso for typed therapy (either from an NHS Provider or via a self-direct referral; right). External recruits signed-up specifically for the study via an external webpage following social media or email advertisements. All potential participants irrespective of enrolment avenue were triaged for suitability based on a Self-Assessment Questionnaire (SAQ). For patients, only those deemed to be potentially eligible were invited to participate. Participants were withdrawn either actively (requested to withdraw), passively (dropped-out or disengaged from study procedures), clinician-led (withdrawn based on clinician recommendation), or other (due to reasons such as technical issues). TAU = treatment as usual.
The six modules consisted of an introduction module, three core modules, and two consolidation modules (Figure 2). The three core modules each consisted of three sessions that followed the pattern of i) learning, ii) activity, and iii) practice. The two consolidation modules consisted of two sessions. There were 16 sessions total. The introduction and consolidation modules consisted of sessions designed for onboarding and learning consolidation, respectively. All modules began with a symptom "check-in" consisting of the GAD-7 and PHQ-9 within the software immediately before the first session within that module. Sessions were made available on a timed schedule subject to completing the prior session (Figure 2).

Within each session, the software used a conversational agent to guide participants through a combination of videos, educational content, conversations, and worksheets written by accredited clinicians. The software used AI models for Natural Language Understanding, specific and tailored elements of Natural Language Generation and a dialogue management system. Part way through enrolment, with agreement from the overseeing NHS Research Ethics Committee, the software was updated to fix bugs, improve the user experience within the introductory module, and update select AI models. The final 60 participants enrolled were offered the updated software. Software version was controlled for in statistical analyses.

To ensure participant safety and maximize engagement and acceptability of the program, a dedicated human user and clinical support service was provided. Prior to enrolment, as part of the screening process, all participants received a standardized clinical assessment by a trained clinician with an accredited postgraduate qualification via typed modality. The clinician assessed the individual's needs, determined if they were eligible for the study and obtained informed consent. Research coordinators provided fortnightly check-in calls to all participants throughout the program and sent weekly emails or SMSs to remind participants only if they deviated from the program schedule. Risk could be flagged through symptom monitoring of GAD-7 and PHQ-9 scores or through interaction with the research coordinators during check-in calls or ad hoc communication. Flagged risk was escalated to a clinician for review. Where appropriate the participant would then be contacted for further risk assessment by a clinician to ensure their safety. Participants could also request an appointment with a clinician at any point to discuss their journey, particularly if they were unsure the program was working for them. At the end of the study, all participants were offered a further discharge appointment with a study clinician to discuss the next steps for their care.

**Figure 2. Schematic of ieso Digital Program with human clinical and user support service and study procedures.** All participants received a clinical assessment prior to enrolment and were offered a discharge appointment with a clinician following the program. Clinicians were available via asynchronous messaging or for a review appointment whenever needed. All participants received email or SMS reminders and fortnightly check-in calls throughout the program to maximize engagement delivered via the research team. The ieso Digital Program included 6 modules with a total of 16 sessions. Each module started with a symptom check-in consisting of the GAD-7 and PHQ-9.
Palmer et al., (2024) – Combining AI and human support in mental health

The support service and study procedures are illustrated in Figure 2. In total, delivering the intervention required an average of 97 minutes (1.6 hours) of clinician time (i.e. time spent in sessions with participants) per participant. This included 299 assessments (mean 66 mins; range 31–105 mins), 46 review appointments (mean 32 mins; 14–60 mins) and 174 discharge appointments (mean 44 mins; range 13–76 mins).

Participants

Adults with mild to severe symptoms of anxiety, consistent with Generalized Anxiety Disorder (GAD), were invited to participate either following referral to ieso’s typed therapy service (either referred to ieso from the NHS Provider or via self-referral direct to ieso) or in response to online advertisements or email invitation through the NIHR BioResource for Translational Research (https://bioresource.nihr.ac.uk/). Only participants with a main problem descriptor of GAD were eligible as established through a clinician assessment in line with the NHS TT manual (16).

All participants met the following eligibility criteria:
- over the age of 18 years at point of recruitment;
- GAD-7 total score > 7;
- PHQ-9 total score < 16;
- access to a smartphone and internet connection;
- registered with a General Practitioner in the UK;
- not currently receiving psychological therapy;
- did not have PTSD, OCD or Panic Disorder;
- did not have a change in psychiatric medication in the past 1 month;
- did not display significant risk of harm to self, to others or from others (as established with the clinical assessment).

Any individuals who had previously participated in user research for the digital program were excluded. Participants were recruited between 10\textsuperscript{th} October 2023 and 2\textsuperscript{nd} February 2024. Financial incentive up to a total of £60 was provided in the form of vouchers based on study assessments and completion of modules within the digital program. For a sub-sample that participated in additional interviews, an additional £15 voucher per semi-structured interview was provided.

Sample size

Previous studies have reported up to a 70% attrition rate when measuring engagement and adherence in mental health digital programs (35–37), therefore we aimed to enroll 300 participants with the expectation of a 40–70% attrition rate, resulting in a final sample of 90–180 participants. A non-inferiority power analysis was conducted prior to retrospective analysis of external control data to estimate the total sample size needed to quantify clinical effectiveness (i.e. change in GAD-7 total score) compared to an active external control. Clinical effectiveness was defined as a change in GAD-7 score over either the course of six treatment sessions or until recovery was reached (if sooner than 6 sessions). A non-inferiority margin of a 1.8 change in GAD-7 total score was chosen based on previous literature ((38–40); see Supplementary Methods for more details). Using data from patients being treated for GAD via typed-CBT, with at least six sessions or recovery, we estimated an expected standard deviation of GAD-7 change of 5.14. To estimate a sample size, we used the following equation:

\[ n = 2 \left( \frac{\bar{z}_a + \bar{z}_b}{\delta + Z_{\beta}/\sigma} \right)^2 \]  

where \(Z_{\alpha}\) and \(Z_{\beta}\) are the standard normal scores for the one-sided significance level of 2.5% (1.96) and power of 90% (1.28) respectively, \(\delta\) is the non-inferiority level 1.8 and \(\sigma\) is the standard deviation 5.14. A sample size of 172 was estimated for the study intervention to enable a non-inferiority analysis of clinical effectiveness compared to human-delivered care.

Patient Public Involvement (PPI)

At ieso, experts-by-lived experience are involved in research and development work as members of a PPI panel and as partners advising on ongoing work. For this study, all participant facing documents were reviewed by members of the PPI panel. In addition, focus groups with members of the PPI panel during study conceptualization aimed to understand participant needs and expectations in the context of “keeping safe” whilst using the digital program, and helped developed recruitment marketing campaigns.

External comparator data source

External comparator data were taken from two NHS TT service providers: i) ieso typed therapy data where a patient receives CBT through 1:1 communication with a qualified therapist using real-time text-based messaging; and ii) Dorset Healthcare University NHS Foundation Trust (DHC) delivering face-to-face routine therapy appointments. The information captured through the dataset of NHS TT is intended to support the monitoring of the implementation and effectiveness of national policy and legislation, policy development, performance analysis and benchmarking, national analysis and statistics, and national audit of NHS TT services. At registration, patients agree to the services’ terms and conditions, including the use of deidentified data for research and audit purposes, including academic publications or conference presentations. External control data were obtained from patients referred to: a) ieso’s typed therapy service between January 2022 and
Palmer et al., (2024) – Combining AI and human support in mental health

December 2023, and b) DHC between January 2017 and December 2021.

Statistical methods

Analyses were conducted in R (42). A statistical analysis plan was defined prior to final analyses being conducted.

Per-protocol vs. intention-to-treat samples

The per-protocol (PP) sample (n=169) was defined as participants who completed the minimum meaningful clinical dose of the program (MMCD) and the final post-intervention GAD-7 and PHQ-9 questionnaires. This dose was defined a priori by three accredited cognitive behavioral therapists who evaluated the content of the program to determine the amount of content required to deliver meaningful clinical improvement on the GAD-7 scale based on their clinical experience (mean experience of 14 years delivering psychological therapy). Based on this evaluation, the MMCD was defined as completing modules 1 to 3 in the digital program and the module 4 check-in.

The intention-to-treat (ITT) sample (n=299) included all participants who completed questionnaires at enrolment irrespective of adherence to the digital program except for one participant who requested that their data be deleted. Due to missing data for the pre-intervention WSAS (external recruits only), the ITT sample for all WSAS analyses was n=295.

Engagement and adherence analyses

Metrics of adherence were primarily assessed with descriptive statistics of in-software usage metrics: median and distribution of time spent in the digital program in hours, days since initialization of the program (defined based on the date that the software was downloaded); and proportion of participants completing each session, module, and check-in. An "engaged" patient is defined as an individual who has received the minimum amount of therapy such that pre- and post-treatment measures can be collected, and clinical outcomes estimated (16). Here we used a comparable definition of engagement based on usage of the program (including time in the program, content delivered, and number of outcomes measured) defined as completing session 1 of module 2 in the program. This is in contrast to the MMCD definition which is defined based on both usage and expected improvement in symptoms.

Effectiveness analyses

Clinical effectiveness was quantified by calculating the change in anxiety symptoms, measured using the GAD-7, from baseline to final score, and estimating a within-subject effect size (Cohen’s d). The threshold for a clinically meaningful reduction in symptoms was defined as a change greater than the reliable change index of the GAD-7 scale (minimum of a 4-point reduction; Toussaint et al., 2020). Clinical outcomes were calculated using the following definitions: a) improvement was defined as a reduction on the PHQ-9 or GAD-7 scales greater than or equal to the reliable change index (≥4 for GAD-7; ≥6 for PHQ-9) and no reliable increase on either measure; b) recovery was defined as reduction on both scales to below the clinical cutoff (GAD-7 score <8; PHQ-9 score <10); c) reliable recovery was defined as having both improved and recovered; d) responder rate was defined as an improvement of either ≥4 on the GAD-7 or ≥6 on the PHQ-9; and e) remission rate was defined as having either a final GAD-7 score <8 or final PHQ-9 score <10 for those only having started above the clinical cut-off. Definitions for improvement, recovery and reliable recovery are equivalent to those used in NHS TT (44). A within-subjects effect-size for mean change in GAD-7 scores from post-intervention to one month follow-up was calculated to determine the short-term durability of any effects of the digital intervention. We also measured effectiveness by calculating the change in PHQ-9 and WSAS between baseline and final score, as well as between comparator groups. For the ITT sample, when calculating GAD-7 and PHQ-9 effectiveness, missing post-intervention scores were imputed using last observation carried forward, such that the final score collected prior to disengagement or withdrawal was used.

Regression models predicting adherence & effectiveness

To determine whether any demographic or study variables were associated with adherence or effectiveness, a series of regression analyses were conducted. All regression models included age, gender, highest qualification, employment status, religion, presence of a chronic physical health condition, ethnicity, reported disability, sexuality, baseline GAD-7 severity, software version, and enrolment path (referred to ieso’s typed therapy service or externally recruited) as predictors. Linear regression models were used to predict continuous dependent variables: i) number of sessions completed; ii) change in GAD-7 score from baseline to final score. A logistic regression model was used to predict non-adherence (i.e. participants who did not complete the necessary program sessions or study assessments to be in the PP sample; non-adherence coded as 1). Due to unequal sample sizes within demographic sub-categories (e.g. sexuality), groups were truncated to aid in the interpretability of findings and power of analyses. To determine if adherence across sessions differed between groups, adherence rates were compared between the digital program, face-to-face CBT.
Palmer et al., (2024) – Combining AI and human support in mental health

and typed CBT by estimating the slope and confidence intervals of the association between the proportion of sample that completed each GAD-7 assessment (either symptom check-in within the digital program or prior to each therapy session as standard within NHS TT) and session number. Sessions were aligned such that each symptom check-in within the program was associated with a treatment session for the control group.

Safety analyses

Safety was assessed using reported serious adverse events, software deficiencies, and number of cases withdrawn based on clinician assessment of suitability to continue with the program. Software deficiencies include malfunctions or errors of the software that could result in issues related to safety or software performance.

Statistical comparison to propensity-matched control groups

Three propensity-matched external control groups were created using real-world historic patient data (see External comparator data source) to compare the clinical effectiveness of the intervention to no intervention and standard of care. All propensity-matched control patients had a main problem descriptor of GAD as established through a clinician assessment.

The control groups consisted of:

i. waiting controls (total available sample n=576); patients referred for typed-CBT with two GAD-7 scores between 4-10 weeks apart without having started treatment during that time (same sample used for PP and ITT analyses),

ii. therapist delivered typed CBT (total available sample n=2,210); patients referred for typed-CBT with at least two scores on the GAD-7, who had completed a course of typed CBT - defined by the discharge code of ‘completed treatment’ - and discharged with a maximum of twelve treatment sessions (PP sample), or any patient who had entered treatment, regardless of completion (ITT sample), and

iii. therapist delivered face-to-face CBT (total available sample n=753); NHS TT patients referred to DHC who received face-to-face CBT and had a minimum of two and a maximum of twelve treatment sessions (PP sample), or any patient who attended treatment (ITT sample). Unlike the typed-CBT comparator, due to unavailability of discharge codes it was not possible to use the ‘completed treatment’ to define the PP sample for this group.

Enrolled participants in the intervention group were propensity-matched to patients from these control groups using baseline GAD-7 scores, baseline PHQ-9 scores, age, and the presence of a chronic physical health condition (yes/no/not known). Propensity-matching was conducted using the ‘MatchIT’ package (45) in R with ‘nearest neighbor’ methodology (average treatment effect in treated patients). For the waitlist control only participants in the PP sample were matched (n=169) due to limited available data for matching. For the human-delivered therapy control groups all participants were matched (n=299). Supplementary Table 2 illustrates the matching of comparator groups to the intervention sample.

In line with the a priori defined statistical analysis plan, a superiority analysis was conducted to test the hypothesis that the clinical effectiveness of the intervention was greater than a propensity-matched waiting control group. A non-inferiority analysis was conducted to test the hypothesis that the clinical effectiveness of the intervention was not inferior to the effectiveness of typed CBT or face-to-face CBT in comparison to waiting-list. Within and between-subject effect sizes were also estimated for the change in total score on the PHQ-9 and the WSAS to estimate the effectiveness of the intervention on low mood and work and social functioning relative to the waiting control.

Results

The final sample for analysis included 299 participants of whom 80% were female (n=240) with a mean age at baseline of 39.8 years (range: 18 – 75 years). Table 1 provides an overview of demographics and baseline severity for participants in the intervention group for both the ITT and PP samples.

Engagement and adherence

Participants (n=299) completed a median of 6.1 hours of program interaction over 53.1 days. This was higher for the PP sample in which participants completed a median of 8.7 hours over 59.6 days. In total, 232 participants (78%) were engaged in the program (i.e. completed session 1 of module 2) involving a median of 2 hours interacting with the program content over 14 days. Out of those engaged participants, 78% (n=180) reached the minimum meaningful clinical dose (i.e. completing up to check-in 4 out of 6 in the program). The overall study attrition rate (defined as the proportion of participants who did not complete the final study questionnaires) was
Palmer et al., (2024) – Combining AI and human support in mental health

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Category</th>
<th>ITT (N=299)</th>
<th>PP (N=169)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, Mean (SD)</td>
<td>-</td>
<td>39.8 (12.8)</td>
<td>41.7 (11.8)</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td>Female</td>
<td>240 (80.3)</td>
<td>137 (81.1)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>46 (15.4)</td>
<td>26 (15.4)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>4 (1.3)</td>
<td>2 (1.2)</td>
</tr>
<tr>
<td></td>
<td>Not Known</td>
<td>9 (3.0)</td>
<td>4 (2.4)</td>
</tr>
<tr>
<td>Ethnicity, n (%)</td>
<td>White</td>
<td>266 (89.0)</td>
<td>155 (91.7)</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>5 (1.7)</td>
<td>2 (1.2)</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>14 (4.7)</td>
<td>6 (3.6)</td>
</tr>
<tr>
<td></td>
<td>Black/African/Caribbean/Black British</td>
<td>3 (1.0)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>2 (0.7)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>9 (3.0)</td>
<td>4 (2.4)</td>
</tr>
<tr>
<td>Highest Qualification, n (%)</td>
<td>Post-graduate degree level qualification</td>
<td>103 (34.4)</td>
<td>65 (38.5)</td>
</tr>
<tr>
<td></td>
<td>Degree level qualification</td>
<td>100 (33.4)</td>
<td>59 (34.9)</td>
</tr>
<tr>
<td></td>
<td>Qualifications below degree level</td>
<td>84 (28.1)</td>
<td>41 (24.3)</td>
</tr>
<tr>
<td></td>
<td>No formal qualifications</td>
<td>2 (0.7)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Don’t know</td>
<td>7 (2.3)</td>
<td>2 (1.2)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1 (0.3)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>2 (0.7)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td>Disability, n (%)</td>
<td>Disability</td>
<td>56 (18.7)</td>
<td>33 (19.5)</td>
</tr>
<tr>
<td></td>
<td>No Perceived Disability</td>
<td>232 (77.6)</td>
<td>132 (78.1)</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>11 (3.7)</td>
<td>4 (2.4)</td>
</tr>
<tr>
<td>Long Term Condition, n (%)</td>
<td>LTC</td>
<td>114 (38.1)</td>
<td>70 (41.4)</td>
</tr>
<tr>
<td></td>
<td>No LTC</td>
<td>167 (55.9)</td>
<td>91 (53.8)</td>
</tr>
<tr>
<td></td>
<td>Not Known</td>
<td>18 (6.0)</td>
<td>8 (4.7)</td>
</tr>
<tr>
<td>Religion, n (%)</td>
<td>No religion</td>
<td>187 (62.5)</td>
<td>104 (61.5)</td>
</tr>
<tr>
<td></td>
<td>Christian</td>
<td>71 (23.7)</td>
<td>45 (26.6)</td>
</tr>
<tr>
<td></td>
<td>Buddhist</td>
<td>1 (0.3)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Hindu</td>
<td>5 (1.7)</td>
<td>3 (1.8)</td>
</tr>
<tr>
<td></td>
<td>Jewish</td>
<td>3 (1.0)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Muslim</td>
<td>5 (1.7)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sikh</td>
<td>1 (0.3)</td>
<td>1 (0.6)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>11 (3.7)</td>
<td>7 (4.1)</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>15 (5.0)</td>
<td>7 (4.1)</td>
</tr>
<tr>
<td>Sexual Orientation, n (%)</td>
<td>Heterosexual</td>
<td>237 (79.3)</td>
<td>132 (78.1)</td>
</tr>
<tr>
<td></td>
<td>Gay/Lesbian</td>
<td>7 (2.3)</td>
<td>5 (3.0)</td>
</tr>
<tr>
<td></td>
<td>Bi-sexual</td>
<td>32 (10.7)</td>
<td>22 (13.0)</td>
</tr>
<tr>
<td></td>
<td>Other sexual orientation not listed</td>
<td>7 (2.3)</td>
<td>2 (1.2)</td>
</tr>
<tr>
<td></td>
<td>Don’t know</td>
<td>11 (3.7)</td>
<td>4 (2.4)</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>5 (1.7)</td>
<td>4 (2.4)</td>
</tr>
<tr>
<td>Employment Status, n (%)</td>
<td>Employed</td>
<td>241 (80.6)</td>
<td>144 (85.2)</td>
</tr>
<tr>
<td></td>
<td>Unemployed and actively seeking work</td>
<td>7 (2.3)</td>
<td>2 (1.2)</td>
</tr>
<tr>
<td></td>
<td>Not working and not actively seeking work</td>
<td>39 (13.0)</td>
<td>19 (11.2)</td>
</tr>
<tr>
<td></td>
<td>Prefer not to say</td>
<td>12 (4.0)</td>
<td>4 (2.4)</td>
</tr>
</tbody>
</table>

32%. Descriptive statistics of engagement with the program across modules are outlined in Table 2.

To determine if adherence across sessions differed between the groups, we compared adherence rates between the digital program, face-to-face therapy and typed therapy based on session completion. Confidence intervals for estimates of the adherence rate were overlapping indicating no difference in adherence rates across groups (Figure 3; Supplementary Table 3).
Palmer et al., (2024) – Combining AI and human support in mental health

Table 2. Engagement metrics for the digital program.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Median time since initialization (days)</th>
<th>Median time interacting in program (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-protocol sample</td>
<td>169</td>
<td>59.6</td>
<td>8.7</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intention-to-treat</td>
<td>299</td>
<td>53.1</td>
<td>6.1</td>
</tr>
<tr>
<td>sample total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engaged sample</td>
<td>232</td>
<td>14.0</td>
<td>2.0</td>
</tr>
<tr>
<td>total (up to module</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 session)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All participants by</td>
<td>Module 1</td>
<td>284</td>
<td>0.0</td>
</tr>
<tr>
<td>milestone</td>
<td>check-in</td>
<td></td>
<td>0.03</td>
</tr>
<tr>
<td>Module 2 check-in</td>
<td>240</td>
<td>13.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Module 3 check-in</td>
<td>209</td>
<td>23.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Module 4 check-in</td>
<td>180</td>
<td>35.0</td>
<td>4.1</td>
</tr>
<tr>
<td>Module 5 check-in</td>
<td>138</td>
<td>42.9</td>
<td>5.0</td>
</tr>
<tr>
<td>Module 6 check-in</td>
<td>113</td>
<td>49.5</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Median days since initialization was calculated as number of days since software download at onboarding for each sample: PP, ITT and the engaged sample (i.e. those who completed up to session 1 of module 2). Metrics are also shown for each symptom check-in at the beginning of each module.

Median days since initialization was calculated as number of days since software download at onboarding for each sample: PP, ITT and the engaged sample (i.e. those who completed up to session 1 of module 2). Metrics are also shown for each symptom check-in at the beginning of each module.

To investigate potential drivers of program adherence, demographic and study factors were associated with i) adherence, defined as number of completed sessions, and ii) non-adherence, defined as not being included in the PP sample. The number of completed sessions in the program did not show any significant associations with demographic or study factors (linear regression: F(24, 274) = 1.4, \( p = .11 \), adjusted \( R^2 = 0.03 \); Supplementary Table 4). Age was associated with non-adherence, with younger participants less likely to be in the PP sample (logistic regression: OR = 0.97, \( p = .009 \); Supplementary Table 5).

Clinical Effectiveness

Anxiety symptoms

On average, across the intervention sample, there was a large, clinically meaningful reduction in anxiety symptoms from baseline to final score (PP: mean GAD-7 change = –7.4, \( d = 1.6 \); ITT: mean GAD-7 change = –5.4, \( d = 1.1 \)). This reduction was significantly greater than that found for the waiting control (mean GAD-7 change = –1.9; PP between-subject effect: \( p < .001 \), \( d = 1.3 \); ITT between-subject effect: \( p < .001 \), \( d = 0.8 \)), and statistically non-inferior to the propensity-matched face-to-face therapy control (PP: mean GAD-7 change = –6.4, non-inferiority effect \( p < .001 \); ITT: mean GAD-7 change = –6.0, non-inferiority effect \( p = .002 \)). For the propensity matched typed-therapy control, the intervention was significantly non-inferior for the PP sample (mean GAD-7 change = –7.5; non-inferiority \( p < .001 \)), and for the ITT sample the effect was approaching significance (mean GAD-7 change = –6.6, \( p = .06 \); Figure 4, Table 3). Clinical outcomes for all groups are reported in Supplementary Table 6.

The trajectory for mean reduction in anxiety symptoms was steeper following the earlier program modules (Figure 5; Supplementary Table 7). When stratified by baseline GAD-7 severity into mild, moderate, and severe groups, the severe group showed the greatest reduction in anxiety symptoms (PP (n=48): mean change on GAD-7 = –10.7, \( d = 2.0 \); ITT (n=87): mean change on GAD-7 = –7.9, \( d = 1.3 \); Supplementary Table 8). By the end of the program, the moderate and severe groups showed a mean GAD-7 score in the mild range. The clinical effect was sustained at one-month follow-up (Figure 5). Between final score and one month follow-up, there was no change in GAD-7 mean score for the PP (n=166) and ITT (n=210) samples with follow-up data (both sample mean difference = 0.0; Supplementary Table 7).
Palmer et al., (2024) – Combining AI and human support in mental health

Figure 4. Change in anxiety symptoms from baseline to final score for all groups. A) Mean change (final score – baseline) in GAD-7 scores for the PP sample (n=169), propensity-matched waiting control group, face-to-face CBT group, and typed CBT group. B) Mean change in GAD-7 scores for the ITT sample (n=299) and a propensity-matched waiting control group, face-to-face CBT group, and typed CBT group. C) Mean GAD-7 scores at baseline and final score with 95% confidence intervals for the PP sample (n=169) and propensity-matched waiting control group, face-to-face CBT group, and typed CBT group. D) Mean GAD-7 scores at baseline and final score with 95% confidence intervals for the ITT sample (n=299) and propensity-matched waiting control group, face-to-face CBT group, and typed CBT group. *** = p < .001, ** = p < .005

The associations between participant demographics, study factors and change in GAD-7 score were explored with a linear regression: F(24, 274) = 3.45, p < .001, adjusted R² of 0.16. Greater reductions in GAD-7 scores were associated with higher baseline GAD-7 scores (β = 0.70, SE = 0.09, t = 7.6, p < .001), and higher baseline age (β = 0.07, SE = 0.02, t = 3.0, p = .003) (Supplementary Table 9), such that more severe, older participants saw a larger change in GAD-7 score.

Mood symptoms

As intended, given the specificity of the program for targeting symptoms of generalized anxiety, there was a statistically significant yet smaller effect for low mood symptoms as measured with the PHQ-9 (PP: mean PHQ-9 change = −3.1, d = 0.7; ITT: mean PHQ-9 change = −1.6, d = 0.3) (Table 4). This mean change was significantly greater than the mean change in the waiting control group for the PP sample (mean PHQ-9 change = −1.0, between-subject effect, p < .001, d = 0.5), but not for the ITT sample (p = .11 d = 0.1). Despite this, PHQ-9 remission rate (based on n=80 above the clinical cut-off at baseline) was 78.8% (Supplementary Table 6). Participants with severe and moderate baseline GAD-7 scores experienced the largest improvement in PHQ-9 scores (Supplementary Table 8). There was minimal mean change in scores between post intervention and follow-up for both PP and ITT samples (PP mean difference = 0.5; ITT mean difference = 0.4) (Supplementary Table 10).
Palmer et al., (2024) – Combining AI and human support in mental health

Table 3. Change in GAD-7 score from baseline to final score for all groups.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Comparator</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Mean change</th>
<th>SD</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>Within-subject effect size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-protocol</td>
<td>Waiting control</td>
<td>169</td>
<td>12.5</td>
<td>3.3</td>
<td>−1.9</td>
<td>4.0</td>
<td>−1.3</td>
<td>−2.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>ieso Digital Program</td>
<td>169</td>
<td>12.4</td>
<td>3.4</td>
<td>−7.4</td>
<td>4.6</td>
<td>−6.7</td>
<td>−8.1</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>Face-to-face CBT</td>
<td>253</td>
<td>13.0</td>
<td>3.1</td>
<td>−6.4</td>
<td>4.8</td>
<td>−5.8</td>
<td>−7.0</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>Typed CBT</td>
<td>229</td>
<td>12.5</td>
<td>3.4</td>
<td>−7.5</td>
<td>4.1</td>
<td>−7.0</td>
<td>−8.0</td>
<td>1.8</td>
</tr>
<tr>
<td>Intention-to-treat</td>
<td>ieso Digital Program</td>
<td>299</td>
<td>12.5</td>
<td>3.3</td>
<td>−5.4</td>
<td>5.1</td>
<td>−4.8</td>
<td>−6.0</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>Face-to-face CBT</td>
<td>299</td>
<td>12.9</td>
<td>3.1</td>
<td>−6.0</td>
<td>4.9</td>
<td>−5.5</td>
<td>−6.6</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Typed CBT</td>
<td>299</td>
<td>12.6</td>
<td>3.5</td>
<td>−6.6</td>
<td>4.6</td>
<td>−6.1</td>
<td>−7.1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Mean difference in GAD-7 score was calculated between baseline and final score for the intervention group (“ieso Digital Program”) and all propensity-matched comparator arms: waiting control; face-to-face CBT; and typed-CBT. A negative mean difference denotes a reduction in GAD-7 total scores. Within-subject effect sizes (Cohen’s d) were estimated for the mean change in GAD-7 scores for each group. Change scores were calculated for PP and ITT samples.

Figure 5. Mean reduction in anxiety symptoms across digital program. Mean GAD-7 score for each time point for all participants that completed the questionnaires at each time point. Trajectories split by GAD-7 baseline severity: mild, moderate and severe (Supplementary Table 7).

Safety

The digital program was well tolerated, and no serious adverse events were identified during the study. There was one report of migraine and two reports of insomnia. There were 10 software deficiencies that occurred (for 7 participants; 90% prior to the software update) for reasons such as technical issues or difficulties with the conversational agent understanding users. In all instances participants were offered an appointment to discuss any potential impact of this on their mental health and reminded of their right to withdraw. These instances resulted in one active participant withdrawal. Across the study, 10 participants were withdrawn by a study clinician following a conversation between the participant and a study clinician. These withdrawals were linked to the study exclusion criteria and suitability for the program rather than the safety of the intervention.

Work and social functioning

There was a significant improvement in work and social functioning measured using the WSAS from baseline to final score for the intervention group (PP: mean WSAS change = −5.3, d = 0.9; ITT n=295): mean WSAS change = −4.7, d = 0.7) (Table 4). This mean change was significantly greater than the mean change in the waiting control group (mean WSAS change = −0.1; PP between-subject effect p< .001, d= 1.2; ITT between-subject effect, p< .001, d= 0.8). The largest changes in functioning were for severe and moderate anxiety groups (Supplementary Table 8).

Discussion

This study demonstrates that an evidence-based, human-supported digital intervention for adults with mild, moderate and severe anxiety produced a large clinically meaningful reduction in anxiety symptoms significantly greater than a propensity-matched waiting control and non-inferior to real-world human delivered therapy. Engagement with the program was high and participants adhered to the intervention at a similar rate to the external therapy control groups. The intervention achieved comparable outcomes to human-delivered care with significantly reduced clinician time. By integrating technology and human support, this study demonstrates...
Palmer et al., (2024) – Combining AI and human support in mental health

<table>
<thead>
<tr>
<th>Sample</th>
<th>Comparator</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>Within-subjects effect size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Baseline score</td>
<td>Change in score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9</td>
<td>Waiting control</td>
<td>169</td>
<td>8.4</td>
<td>3.4</td>
<td>−1.0</td>
<td>3.6</td>
<td>−0.4</td>
<td>−1.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Per-protocol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−3.1</td>
<td>4.5</td>
<td>−2.4</td>
<td>−3.8</td>
<td>0.7</td>
</tr>
<tr>
<td>Face-to-face CBT</td>
<td></td>
<td>253</td>
<td>8.5</td>
<td>3.7</td>
<td>−3.0</td>
<td>4.8</td>
<td>−2.4</td>
<td>−3.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Typed CBT</td>
<td></td>
<td>229</td>
<td>8.1</td>
<td>3.5</td>
<td>−4.1</td>
<td>3.9</td>
<td>−3.6</td>
<td>−4.6</td>
<td>1.1</td>
</tr>
<tr>
<td>Intention-to-treat</td>
<td></td>
<td>299</td>
<td>8.0</td>
<td>3.7</td>
<td>−1.6</td>
<td>4.8</td>
<td>−1.1</td>
<td>−2.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Face-to-face CBT</td>
<td></td>
<td>299</td>
<td>8.4</td>
<td>3.6</td>
<td>−2.7</td>
<td>4.8</td>
<td>−2.2</td>
<td>−3.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Typed CBT</td>
<td></td>
<td>299</td>
<td>8.1</td>
<td>3.6</td>
<td>−3.3</td>
<td>4.2</td>
<td>−2.9</td>
<td>−3.8</td>
<td>0.8</td>
</tr>
<tr>
<td>WSAS</td>
<td>Waiting control</td>
<td>153</td>
<td>10.6</td>
<td>6.1</td>
<td>−0.1</td>
<td>1.3</td>
<td>0.1</td>
<td>−0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Per-protocol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−5.3</td>
<td>6.2</td>
<td>−4.4</td>
<td>−6.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Face-to-face CBT</td>
<td></td>
<td>253</td>
<td>14.1</td>
<td>7.6</td>
<td>−4.3</td>
<td>8.6</td>
<td>−3.3</td>
<td>−5.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Typed CBT</td>
<td></td>
<td>223</td>
<td>10.8</td>
<td>6.4</td>
<td>−4.6</td>
<td>5.5</td>
<td>−3.8</td>
<td>−5.3</td>
<td>0.8</td>
</tr>
<tr>
<td>Intention-to-treat</td>
<td></td>
<td>295</td>
<td>14.9</td>
<td>6.6</td>
<td>−4.7</td>
<td>6.5</td>
<td>−3.8</td>
<td>−5.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Face-to-face CBT</td>
<td></td>
<td>299</td>
<td>14.1</td>
<td>7.6</td>
<td>−3.9</td>
<td>8.3</td>
<td>−2.9</td>
<td>−4.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Typed CBT</td>
<td></td>
<td>291</td>
<td>10.8</td>
<td>6.3</td>
<td>−3.9</td>
<td>5.7</td>
<td>−3.2</td>
<td>−4.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Mean differences in PHQ-9 and WSAS scores were calculated between baseline and final score for the intervention group ("ieso Digital program") and all propensity-matched comparator arms: waiting control; face-to-face CBT; and, typed-CBT. A negative mean difference denotes a reduction in scores. Within-subject effect sizes (Cohen’s d) were estimated for the mean change for each group. Change scores were calculated for PP and ITT samples.

The potential to expand global access to high-quality, effective mental healthcare.

The large clinical effect of the digital intervention across participants with moderate or severe symptoms highlights the clinical value of the combined program content and human support. Here, the PP (d=1.3) and ITT (d=0.8) effect sizes relative to waitlist are larger than the pooled effect size reported in a recent meta-analysis (n comparisons = 96, g = 0.26) (14). Unlike the PP sample which is designed to demonstrate the clinical effectiveness of an intervention when the intervention is adhered to, the ITT sample provides an estimate of effectiveness more reflective of the real-world context by accounting for disengagement. The large ITT effect was significantly non-inferior to face-to-face therapy, and approaching significance for non-inferiority to typed-therapy (p = 0.06). Human-delivered care enables greater flexibility to respond to patient concerns and adapt content compared to a digital program, therefore the comparable clinical effects and adherence rates across groups indicates the potential of this digital intervention to significantly impact real-world patient outcomes.

It is important to note that high relapse and recurrence rates have implications for both patient quality of life and economic healthcare costs, therefore ensuring effects are durable is imperative (46–48). Incorporating cognitive and behavioral principles into daily life through practical exercises can enable meaningful behavioral change that persists beyond treatment end (16). Here, both the persistent clinical effect at one month follow-up, and the significant improvement in the impact of anxiety on participants’ day-to-day functioning (as measured with the WSAS) highlights the potential of the intervention to instigate long-lasting behavioral change. Retrospective analysis of recurrence data from electronic health records is needed to accurately measure the persistence of the clinical effect in the real world over a longer follow-up time-period.

The engagement rate of the digital program (78%) and time to reach “engaged” (~2 hours of program interaction over 2 weeks) is comparable to engagement rates and time in therapy observed in NHS TT services for treatment of GAD (70%; 2022-2023) (49). Adherence rates across groups in the study were also similar. Average program interaction time (median 6.1 hours) across the ITT sample was greater than that reported for similar app-based interventions (e.g. median 3.4 hours) (50), indicating high engagement with the program. Study
attrition (32%) is higher than previous reports from studies of conversational agent-delivered mental health interventions (22%) (51), yet similar to real-world global treatment drop-out rates (~20-40%) (52,53). This may be due to the pragmatic design of the study: 30% of the sample recruited through ieso’s therapy referrals could choose to withdraw at any time and immediately access 1:1 human-delivered therapy; and participants had the option to discuss their progress or any issues with the clinical team at any point. These factors could have increased withdrawal rates more than previous studies, but more readily reflect real-world patient choice and clinical decision-making.

To our knowledge, this study is the first to compare the effectiveness of a digital intervention to standard of care using external propensity-matched comparator groups from real-world patient data. There is increasing acceptability for the use of externally controlled clinical trials (54–57) made possible by the availability of large-scale, standardized datasets. Generating external comparator groups reduces patient burden, study costs, and avoids delaying treatment for the comparator group receiving no intervention (58). However, creating standard of care control arms that are directly comparable to a novel intervention remains difficult due to differences in how to define comparable doses, treatment completion and account for study-specific assessments. Moreover, the lack of randomization means selection bias and study effects are not controlled for in the current study. Nevertheless, this is more reflective of real-world care where treatment outcomes are biased by a patient’s preference and choice over their treatment.

The clinical effect and engagement rate reported in the current study could have been driven by a combination of the three key features of the digital intervention: i) a curated and structured evidence-based program, ii) a conversational agent to deliver the program content, and iii) a human user and clinical support model akin to standard healthcare delivery. First, the structured evidence-based program was curated by a team of accredited cognitive behavioral therapists with an average of 14 years direct clinical experience. The program used principles from traditional CBT (22) including third wave approaches, such as ACT. This approach encourages individuals to accept their thoughts and feelings while committing to actions aligned with their values. There is a growing body of evidence indicating that ACT demonstrates comparable effectiveness to other forms of CBT for anxiety disorders (59–61), and has been shown to be acceptable and engaging within a digital program for GAD (62,63).

Second, a conversational agent was used to personalize the content delivery and enhance engagement. Despite rapid growth in the development of AI conversational agents, use of this technology remains rare in digital mental health interventions, with only ~5% using this technology (14). The majority of these systems employ a tree-based dialogue approach, where natural language processing analyzes user input, and responses are selected from a predefined set of pre-written answers. However, previous research has shown users find this frustrating, particularly when it feels the agent does not understand them (64,65). Recent advances in the development of large language models now make it possible to flexibly generate personalized language for a more engaging user experience. In the current study, the digital program primarily used a tree-based dialogue system with controlled use of natural language generation in specific instances to enhance engagement. Increased use of generative technology and reduced reliance on tree-based approaches will continually improve the capability of conversational agents to create a personalized and engaging experience. However, allowing fully autonomous language generation within the context of mental health, where patient problems can be nuanced, complex and require the consideration of social and cultural contexts, poses a high risk for patient harm and misuse (66). Stringent validation of these new AI technologies with a phased roll out alongside human oversight will be essential to ensure patient safety (67).

Finally, a ‘blended’ design of human support and conversational technology has been suggested to be key for maximizing real-world engagement (51). Previous research has highlighted lack of trust, lack of user-centric design, privacy concerns, poor usability, and being unhelpful in emergencies as key drivers of poor engagement with digital interventions (12). To address these concerns, we mirrored a real-world treatment model including user support services, clinician referral to the program, proactive symptom monitoring and clinician availability for collaborative decision-making with each participant. This service created a credible and trustworthy patient experience that we believe positively impacted patient outcomes. Although this study was not designed to demonstrate the economic value of the intervention, the average clinician time spent per participant was <2 hours, which is significantly lower than current standards of care globally: approximately 4 times less than an average episode of treatment in the UK for GAD (~8 appointments between 45–60 mins; NHS Digital 2021–2022) (49) and ~approximately 8 times less globally (~15 appointments; mean across reported naturalistic studies in (68)). This new model, combining an AI-driven program with clinical support, allows the current, limited supply of trained therapists to help more people than current standards of care.
Limitations of the current study include the use of compensation for time for those who volunteered to participate, and the selection of a sample with limited low mood symptoms. In particular, in line with the study exclusion criteria, individuals with severe depression symptoms were not included. Nevertheless, the propensity-matching across groups controls for this, i.e. all groups included patients with similar baseline anxiety and depression symptoms. Differences in PP sample sizes across the control groups were likely driven by the definition of PP in each context rather than engagement, given similar adherence rates across the groups. Defining a comparable PP sample across groups is challenging due to differences in dose intensity, delivery mechanism and data collected, as well as significant variation across patients in both clinical presentation of generalized anxiety and response to treatment. The PP samples for the therapy control groups were based on completed episodes of care, therefore were agnostic of therapy dose and would have included those who received a low number of sessions and recovered quickly. Those individuals would not have been included in the intervention PP sample which was conservatively defined based on minimum program interaction.

There were also limitations in terms of the diversity of the intervention sample, with enrolled participants predominantly white, highly educated, and female. This sample is reflective of the typical profile of GAD patients in the UK and US (49,69). Although we attempted to increase diversity in this sample through focused marketing campaigns, these efforts were not successful. Needs differ across individuals, conditions and contexts, and a greater understanding of the barriers to research participation is required to fully understand these needs, particularly where groups have been systematically excluded from research, and where there is stigma around mental health. Increasing access to mental health support could play a substantial role in addressing unmet need in underserved groups, therefore future work will aim to evidence the inclusivity of this digital intervention and its potential to counter existing health inequalities.

In conclusion, this study demonstrates that a digital intervention, designed for adults with symptoms of generalized anxiety, produces comparable outcomes to human-delivered CBT while significantly reducing the required clinician time. This result indicates the potential for digital interventions to provide high quality, evidence-based care at scale to address unmet need worldwide. As AI technologies rapidly progress, it is evident that generative dialogue systems that emulate creative and flexible human language will soon be widely accessible. This accessibility will radically change how individuals seek mental health support. Our responsibility lies in leveraging these advances, addressing the ethical and social challenges inherent with AI, and combining the best of technology with the best of clinical care to increase access to effective, safe and engaging mental health support for everyone. Rigorous evidence, particularly to understand the optimal blend of human and computer support for different individuals, will be key to accelerate precision treatment, maintain scalability, maximize uptake and adherence, and integrate digital interventions into health systems.

Acknowledgements

We extend our gratitude to the patients who participated in the study and to the dedicated clinicians and support staff involved. We would like to thank Gerald Chan, Stephen Bruso, Andy Richards, Ann Hayes, David Icke, Michael Black, Clare Hurley, Florian Erber, Richard Marsh, Sam Williams, Jo Parfrey for their support and encouragement. We are grateful to Prof Thalia Eley for introducing us to NIHR BioResource. We thank NIHR BioResource volunteers for their participation, and gratefully acknowledge NIHR BioResource centers, NHS Trusts and staff for their contribution. We thank the National Institute for Health and Care Research, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Program. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. We thank Dorset Healthcare University NHS Foundation Trust (DHC) for providing external data for comparison.

Funding

This research was funded by ieso Digital Health Ltd.

Competing Interests

Chief Investigator (EMa) and other investigators (CEP, EMI, GW, MPE, EC, SL, AS, CH, JY, MB, LM, SM, RC, VT, AC, AW, AB) are employees of ieso Digital Health Limited (the company funding this research) or its subsidiaries. None of these authors had a direct financial incentive related to the results of this study or the publication of the manuscript.

Author contributions

CEP, EMa, AW & AB conceptualized the study. CEP and EMI drafted the paper. GW, EMI, MPE, EC, MB, AS & AC contributed to data analyses and interpretation. SL, JY, CH, EMa & CEP conducted the study. All authors contributed to the interpretation of results and paper revision, and approved the final version.

Data availability

Owing to the potential risk of patient identification, and following data privacy policies at ieso and DHC, individual-level data are not available. Aggregated data are available upon request, subject to a data-sharing agreement with ieso and DHC. Data requests should be sent to the corresponding author and will be responded to within 30 days.
Palmer et al., (2024) – Combining AI and human support in mental health

References


Palmer et al., (2024) – Combining AI and human support in mental health


Palmer et al., (2024) – Combining AI and human support in mental health


60. Han A, Kim TH. Efficacy of Internet-Based Acceptance and Commitment Therapy for Depressive Symptoms, Anxiety, Stress, Psychological Distress, and Quality of Life: Systematic Review and Meta-analysis. J Med Internet Res. 2022 Dec 1;24(12).


65. Huang YS (Sandy), Dootson P. Chatbots and service failure: When does it lead to customer aggression. Journal of Retailing and Consumer Services. 2022 Sep 168.


