AI-Enabled Continuous Care Features in Real-World Psychotherapy: Treatment Engagement and Clinical Outcomes

Scott Graupensperger, PhD [1]

Ondrej Kopecky [1]

Millard Brown, M.D.[1]

Adam Chekroud, PhD [1, 2]

Matt Hawrilenko, PhD [1]

1. Spring Health; New York, NY, United States of America
2. Yale School of Medicine, Psychiatry; New Haven, CT, United States of America

Word Count:

Correspondence: Please address correspondence regarding this article to Dr. Scott Graupensperger (scott.graupensperger@springhealth.com).

Spring Health
60 Madison Ave.
New York, NY 10010.
United States.

Key Points

Question. Is access to AI-enabled continuous care features embedded within routine psychotherapy delivery associated with improved treatment engagement and clinical outcomes under real-world conditions?

Findings. In this cluster-level, matched, quasi-experimental study of adults receiving psychotherapy within an employer-sponsored behavioral health program, access to AI-enabled continuous care features was associated with significantly greater early treatment engagement and faster improvement in depressive and anxiety symptoms compared with psychotherapy alone.

Meaning. AI-enabled support features may incrementally enhance the delivery and effectiveness of established psychotherapies when implemented as complements to routine care at scale.

Abstract

Importance. AI-enabled features have the potential to incrementally improve the effectiveness of routine mental health care, yet evidence from large-scale, real-world evaluations is limited.
Objective. To evaluate whether access to AI-enabled continuous care features embedded within routine psychotherapy delivery is associated with improved treatment engagement and clinical outcomes under real-world conditions.
Design. Preregistered cluster-level, matched, quasi-experimental study using a real-world rollout of AI-enabled continuous care features compared with psychotherapy alone, analyzed under an intention-to-treat framework.
Setting. An employer-sponsored behavioral health program providing outpatient psychotherapy for employees and dependents.
Participants. Adults initiating a new episode of psychotherapy from 25 employers with access to continuous care features and 75 matched employers without access. Treatment engagement was assessed over 7 weeks (n=26,208), and clinical outcomes were evaluated for up to 180 days (n=5,518).
Exposure. Employer-level access to AI-enabled continuous care features designed to support engagement and continuity before and between psychotherapy sessions, compared with psychotherapy alone.
Main Outcomes. Early treatment engagement (number of psychotherapy sessions attended in 7-week study period and time to second session) and changes in depressive and anxiety symptom severity measured using the Patient Health Questionnaire–9 (PHQ-9) and Generalized Anxiety Disorder–7 (GAD-7).
Results. Compared with matched controls receiving psychotherapy alone, the intervention group attended more psychotherapy sessions during the first 7 weeks (rate ratio, 1.05 [1.01, 1.10]) and completed their second session sooner (mean difference, −0.62 days [−1.05, −0.18]). Both groups demonstrated substantial symptom improvement over time; however, access to continuous care features was associated with additional improvement in depressive symptoms (d=0.16) and anxiety symptoms (d=0.15) at the median duration of care (day 44) . These effects translated into clinically meaningful differences in reliable improvement by the median duration of care (NNT=25 for both depressive and anxiety symptoms).
Conclusions and Relevance. In this real-world evaluation, access to AI-enabled continuous care features embedded within routine psychotherapy delivery was associated with incremental improvements in treatment engagement and clinical outcomes beyond psychotherapy alone. While effects were modest, the continuous care features enhanced an already effective base of care and may have a clinically relevant system-level impact when deployed at scale.

AI-Enabled Continuous Care Features in Real-World Psychotherapy: Treatment Engagement and Clinical Outcomes

Evidence-based psychotherapies improve outcomes for common mental health conditions, including depression and anxiety.[1–5] Yet the population burden of mental health concerns remains substantial, indicating that the benefits of effective treatments have not been fully realized at scale.[6–8] This gap does not stem from inherent limitations of existing therapeutic approaches, as major evidence syntheses indicate that evidence-based psychotherapies are consistently effective across modalities in carefully controlled trials.[9] Instead, shortfalls in population-level impact are largely attributable to challenges in real-world delivery. In routine care, a substantial proportion of individuals fail to achieve meaningful improvement following an initial course of care due to a combination of early treatment discontinuation, limited engagement, and insufficient continuity.[10] Accordingly, the central opportunity for advancing mental health outcomes lies in improving engagement and delivery of existing psychotherapies in ways that can produce incremental but meaningful gains in real-world effectiveness.

Employer-sponsored behavioral health benefits are becoming a pragmatic pathway to mental health care, including psychotherapy, offering scalable delivery of evidence-based care.[11–13] Prior evaluations suggest that such programs can improve care accessibility, clinical outcomes, and workplace-relevant endpoints across diverse populations.[14–16] As these platforms mature, the focus has shifted from access alone to optimizing care through features that incrementally improve engagement and outcomes beyond the effective base care. When applied at scale, even modest gains in engagement or symptom improvement may yield meaningful population-level impact across large, covered populations.

Efforts to optimize mental health care have increasingly focused on integrating AI–enabled tools into routine care.[17–19] Unlike standalone digital therapeutics or chatbot-based interventions, these tools are designed to augment clinician-delivered care by supporting the patient and provider before and between sessions.[20] Opportunities to enhance the therapy

experience begin with onboarding (i.e., prior to the first therapy session). Gathering routine information around presenting concerns, treatment history, and the patient's goals prior to therapy can create the bandwidth to form deeper therapeutic connections earlier in an episode of care. AI features can also supplement care between sessions by supporting recall, reflection, and follow-through on clinician-directed strategies, thereby promoting continuity and treatment momentum between sessions. However, empirical evidence for the real-world effectiveness of implementing such tools remains limited, leaving a critical gap regarding whether AI-enabled features improve engagement or clinical outcomes when deployed at scale.

Current Study

The current study leveraged a large-scale, real-world rollout of a first phase of AI-enabled continuous care features across 25 employers, with comparisons to participants from 75 matched employers that did not receive access during the study period. Continuous care refers to AI-enabled tools designed to support engagement and continuity before and between psychotherapy sessions. We hypothesized that access to continuous care features would be associated with greater treatment engagement within the program and greater symptom improvement over time compared to access to psychotherapy alone. See preregistration at AsPredicted #237134: https://aspredicted.org/s7vr-6m7k.pdf).

Method

Intervention

The intervention was delivered within an employer-sponsored behavioral health program (Spring Health), which has been described previously and has demonstrated improvements in depression, anxiety, and workplace outcomes.[13,16,21,22] The program provides access to a predefined number of prepaid psychotherapy and/or medication management sessions with licensed clinicians (6–18 sessions, depending on employer plan level), with additional care available through in-network health plan coverage. Other program components include dedicated care navigation, 24/7 crisis support,[22] data-driven patient–provider matching,[23] and

financial incentives for clinicians to use measurement-based care.

Within this program, continuous care consisted of three optional AI-enabled features designed to augment psychotherapy delivery and support engagement between sessions: guided intake, session summaries, and session takeaways. Guided intake supported structured onboarding prior to the initial psychotherapy session, while session summaries and takeaways provided post-session summaries and action-oriented follow-up content derived directly from provider-delivered content in recorded sessions. Detailed descriptions and implementation considerations are provided in the Supplement.

Study Design

This preregistered study used a cluster-level, matched, quasi-experimental design, with access to AI-enabled continuous care features implemented at the employer level rather than the individual level. Twenty-five employers that received access to continuous care features were matched to 75 employers without access using a 1:3 matching ratio. Matching was conducted at the employer level to improve comparability between intervention and comparison groups on key organizational characteristics, including size, industry, benefit design, and prior-year treatment-seeking rates; detailed matching procedures are provided in the Supplement.

Analyses were conducted using two prespecified analytic subsamples aligned with distinct study objectives. Treatment engagement was evaluated during a 7-week early engagement window following psychotherapy initiation, whereas clinical outcomes were evaluated over a longer follow-up period of up to 6 months. Observation windows and censoring rules were defined to ensure equal opportunity for follow-up and to minimize bias.

The study was preregistered prior to launch, and substantive deviations from the preregistered analysis plan and their implications for interpretation are described in the Supplement. This study involved secondary analysis of deidentified data generated from routine clinical care and program implementation, independent of research activities. Use of these data for research purposes is governed by an ongoing determination by the Yale Institutional Review

Board that such analyses do not constitute human subjects research; accordingly, institutional review board approval and informed consent were not required for this study.

Participants and Eligibility

Participants included adults actively enrolled in the behavioral health program who completed a baseline assessment prior to initiating psychotherapy. Enrollment occurred on a rolling basis as participants initiated a new episode of program-based care, defined as either a first encounter or the first encounter following a 6-month washout period. Treatment engagement analyses were restricted to participants with a complete 7-week observation window following psychotherapy initiation (n=26,208). Clinical outcome analyses focused on participants with at least one follow-up symptom assessment after baseline (n=5,676), who contribute within-person symptom change information; however, all participants were retained in mixed-effects models under an intention-to-treat framework, including those without follow-up assessments. Participant inclusion across analytic subsamples is summarized in the study flow diagram (Figure 1). Baseline characteristics and post-matching balance for each analytic subsample are presented in Table S1. Additional analytic details, including eligibility refinements and deviations from the preregistered plan implemented to preserve analytic validity and statistical power, are described in the Supplement.

Outcome Measures

Treatment engagement outcomes focused on early psychotherapy utilization and treatment momentum. Early utilization was defined as the number of attended psychotherapy sessions within 7 weeks of the initial session, and treatment momentum was assessed using time to second psychotherapy session among participants with at least two attended sessions (n=21,116; 81% of total sample).

Therapeutic alliance was assessed using an optional, session-level alliance survey administered via e-mail following psychotherapy visits. For participants with available data, therapeutic alliance was operationalized as the participant-level mean of all alliance scores

collected within the 7-week study period. Response rates were low (10.3% of participants had at least one report, n=2,705).

Clinical outcomes included depressive and anxiety symptom severity, assessed using the Patient Health Questionnaire–9 (PHQ-9[24]) and the Generalized Anxiety Disorder–7 (GAD-7[25]), respectively. Symptom measures were collected at baseline and during routine measurement-based care follow-up. Additional details on outcome measures is provided in the Supplement.

Statistical Analyses

Analyses were conducted under an intention-to-treat framework based on employer-level intervention assignment. Treatment engagement and clinical outcomes were compared between intervention and matched comparison groups using regression models appropriate for each outcome, with clustering accounted for at the employer level. Longitudinal clinical outcomes were analyzed using mixed-effects models including treatment, time, and their interaction. To contextualize clinical significance, we estimated the population-average probability of clinically reliable improvement at the median follow-up time (44 days),[26,27] defined as a reduction of ≥5 points on the PHQ-9 or ≥4 points on the GAD-7 from baseline. Numbers needed to treat were calculated as the inverse of the absolute difference in improvement probabilities between groups.

Sensitivity analyses included difference-in-differences and temporal falsification analyses by leveraging a large, independent comparison sample (n=36,294) drawn from the same 100 employers and meeting identical inclusion criteria from the year prior to implementing the continuous care features. Full analytic details are provided in the Supplement.

Results

Baseline demographic and clinical characteristics were balanced between intervention and matched control groups across both analytic subsamples, with standardized mean differences below conventional thresholds for imbalance (Table S1). Baseline depressive and

anxiety symptom severity was highly comparable between groups, with mean baseline PHQ-9 and GAD-7 scores differing by less than 0.1 points between intervention and comparison participants, supporting the comparability of cohorts prior to outcome assessment.

Among participants in intervention employers, utilization of continuous care features varied across individuals and features (Table S2). In the treatment engagement analytic sample, 38.2% of participants used at least one continuous care feature during the early engagement period, whereas uptake was higher in the clinical outcomes analytic sample (47.0%). Lower utilization was observed for features requiring additional provider and participant consent (Table S3). Exploratory analyses indicated that feature utilization was not strongly associated with baseline symptom severity or demographic characteristics (Table S4). Additional descriptive details, including provider- and participant-level opt-in rates for session recordings, are provided in the Supplement.

Treatment Engagement Outcomes

In intent-to-treat analyses, intervention assignment was associated with statistically significant improvements in early therapy engagement (Model A in Table 1). Participants in the intervention group completed a higher number of therapy sessions during the first 7 weeks of care compared with controls (rate ratio = 1.05 [95% CI: 1.01, 1.10], p=.017), corresponding to a 5% relative increase in expected session counts after adjustment for baseline symptom severity, age, gender, and employer-level clustering. Model-implied mean session counts were 3.30 sessions for the intervention group and 3.13 sessions for the control group over the 7-week study period. The intervention group also had faster follow-up after treatment initiation. Among participants with at least two completed therapy sessions during the 7-week study period, those in the intervention group completed their second session an average of 0.62 days sooner than controls (β = −0.62 days [−1.05, −0.18], p=.006). Model-implied estimates indicated a mean time to second session of 12.39 days in the intervention group compared with 13.01 days in the control group.

Therapeutic alliance was examined among the ~10% of participants with available alliance survey data. Average alliance scores were high and similar between intervention and comparison groups (means 9.13 and 9.12 out of 10, respectively), and the intervention effect was non-significant (Table 1).

Sensitivity analyses leveraging a temporally shifted falsification cohort and difference-in-differences specifications yielded results consistent with the primary findings across all outcomes (Models B and C in Table 1). Together, these analyses provide convergent evidence that observed effects were attributable to the intervention rather than to secular trends or unobserved employer-level differences.

Clinical Outcomes

Across both intervention and comparison groups, participants experienced substantial reductions in depressive and anxiety symptoms over the course of care, consistent with the expected benefits of psychotherapy. For reference, PHQ-9 scores in the matched control group declined by 3.5 points, on average, by the median duration of care (i.e., day 44), with a similar reduction of 3.2 points on GAD-7 scores.

Within this context of effective base care, assignment to the continuous care features was associated with additional, incremental improvements in symptom trajectories over time. In primary intent-to-treat analyses among participants with elevated baseline symptoms (PHQ-9 ≥10 or GAD-7 ≥10) the group×time interactions indicated faster symptom reduction among intervention participants relative to matched-controls for both outcomes (Table 2; Figure 2).

For depressive symptoms, the intervention was associated with an additional reduction of 0.021 PHQ-9 points per day over controls. At the median duration of care (44 days), this corresponded to a marginal effect of −0.71 PHQ-9 points, reflecting a small standardized effect size (Cohen's d=0.16). For anxiety symptoms, intervention was associated with an additional reduction of 0.018 GAD-7 points per day. At the median duration of care (44 days), the marginal effect was −0.50 GAD-7 points (d=0.15). To contextualize the clinical significance of these

effects, we estimated the probability of reliable improvement at the median duration of care (44 days). The probability of reliable improvement on the PHQ-9 (≥5-point reduction) was 53.2% in the intervention group compared with 49.2% in matched controls (NNT=25). For GAD-7 (≥4-point reduction), the corresponding probabilities were 53.1% and 49.1% (NNT=25).

Falsification analyses applying identical trajectory models to a pre-intervention period showed no evidence of differential symptom improvement by intervention assignment for either outcome. In these placebo models, the group×time interaction was small and not statistically significant for both outcomes. Difference-in-differences sensitivity analyses further supported the primary findings, as intervention assignment was associated with a significantly greater rate of symptom improvement in the post-intervention period relative to the pre-intervention period for both depressive and anxiety symptoms. The group×time×phase interaction indicated an additional reduction of 0.019 PHQ-9 points per day ($p<.001$) and 0.018 GAD-7 points per day ($p<.001$) attributable to the intervention (See Figure S1).

Baseline Severity as a Moderator of Intervention Effects. In models including three-way interactions between intervention assignment, time, and baseline symptom severity, higher baseline PHQ-9 and GAD-7 scores were associated with significantly stronger effects (both interaction estimates $p<.001$). Probing model-estimated trajectories indicated that intervention effects were driven by patients with higher baseline severity (Figure 3). For example, participants with a baseline score of 10 experienced comparable effects to the control group at the median length of care, whereas participants with a baseline score of 20 had 1.69 [-1.99, -1.40] point larger decreases in PHQ-9 (Cohen's d=0.29) and 1.48 [-1.84, -1.13] point larger decreases in GAD-7 (d=0.30) relative to matched-controls. Together, these patterns indicate greater clinical benefit among participants with more severe baseline symptoms.

Discussion

In this large-scale evaluation of AI-enabled continuous care features embedded within

an employer-sponsored behavioral health program, access to these features integrated into psychotherapy was associated with improved treatment engagement and greater symptom improvement compared with psychotherapy alone. Under routine implementation conditions, these findings demonstrate that lightweight, AI-enabled supports can enhance psychotherapy delivery at scale, even when layered onto an already effective base of care. Although the incremental effects were modest in magnitude (d=0.15-0.16), they translated into clinically meaningful differences at the population level, with numbers needed to treat being ~25 to achieve one additional reliable improvement over a relatively short treatment window of 44 days, corresponding to the median duration of care.

Improvements in clinical outcomes were accompanied by meaningful differences in early treatment engagement, suggesting a plausible mechanism through which the continuous care features may exert their effects. The intervention group completed 5% more therapy sessions during the 7-week follow-up period and progressed more quickly from initial to follow-up visits, indicating denser and more rapid engagement during a critical treatment window, despite no significant impact on therapeutic alliance. It follows that these features may operate on a causal pathway outside of therapeutic alliance. For example, increasing memory of treatment and between-session skill rehearsal have been shown to predict symptom change without necessarily strengthening patients' perceived bond with the therapist.[28–30] It is also possible that, given alliance scores were uniformly near the ceiling across both groups, there was little opportunity to detect incremental differences.

Intervention effects varied across participants, with larger benefits observed among individuals with higher baseline severity. This common pattern[31] suggests that incremental support features integrated into psychotherapy may be most beneficial for individuals with greater initial clinical need. One potential explanation is that individuals with higher symptom severity may experience greater emotional activation during sessions and, as a result, have more difficulty retaining session content and benefit more from structured memory supportive

interventions.[32,33] Structured summaries and postsession takeaways may support consolidation and application of therapeutic material after the session has ended.

Taken together, these findings suggest that modest gains from lightweight, AI-enabled support features can yield meaningful clinical impact when deployed at scale. In contrast to psychotherapy enhancements that require substantial clinician training or sustained workflow changes, the continuous care features evaluated here can be implemented with minimal burden, such that even small incremental improvements may translate into meaningful population-level benefits.

Limitations

Although the study deviated from the preregistered analysis plan in several respects, these changes were made to address data availability and modeling considerations and ultimately strengthened the rigor of the analyses (Table S5). While the employer-level rollout and matched comparison design strengthen causal inference, the study was not randomized, and residual confounding cannot be fully excluded. Despite rigorous matching, it is possible that employers invited to participate in, and electing to opt into, beta rollout waves differed in unmeasured ways from those that did not opt in; however, falsification and difference-in-differences sensitivity analyses support the assumption that outcomes were equivalent prior to intervention implementation. Uptake of the AI-enabled features varied across participants, and the intent-to-treat framework does not permit isolation of effects attributable to specific features or levels of use when participants had access to all or no features. Therapeutic alliance was assessed via an optional, session-level survey with limited coverage, constraining inference regarding effects on relational processes. Finally, findings may not generalize beyond employees and dependents with access to employer-sponsored behavioral health benefits or to populations with severe mental illness or acute suicidal risk, who were not the focus of this evaluation.

Conclusions

The present findings demonstrate that AI-enabled support features embedded within routine psychotherapy delivery can yield meaningful improvements in engagement and clinical outcomes under real-world conditions. Although these effects are incremental at the individual level, they are a lightweight addition and accrue on top of an already effective base of mental health care, translating into clinically meaningful population-level impact when deployed at scale. More broadly, these results showcase the value of evaluating digital augmentations as complements to clinician-delivered psychotherapy rather than as standalone interventions, and highlight the importance of rigorous, pragmatic evaluation as such tools continue to be integrated into mental health care delivery.

Disclosures:
Drs. Graupensperger, Kopecky, Brown, Hawrilenko, and Chekroud report being employed by and holding equity in Spring Health, Inc. outside the submitted work. In addition, Dr. Chekroud reported being the lead inventor on 3 patent submissions relating to treatment for major depressive disorder (US Patent and Trademark Office number Y0087.70116US00 and provisional application numbers 62/491 660 and 62/629 041) outside the submitted work. Finally, Dr. Chekroud reported holding equity in Carbon Health Technologies Inc, Wheel Health Inc, Parallel Technologies Inc, Healthie Inc, and UnitedHealthcare; and providing unpaid advisory services to health care technology startups outside the submitted work.

Tables and Figures

Table 1. Early Treatment Engagement Outcomes During the First 7 Weeks of Care: Primary and Sensitivity Analyses.

Model	Total Therapy Sessions (7-week Study Period)			Time to Second Therapy Session (Among those with 2+)			Therapeutic Alliance Mean Score (7-week Study Period)		
	Rate Ratio	(95% CI)	*p*-value	β (days)	(95% CI)	*p*-value	β	(95% CI)	*p*-value
A. Primary Intervention Models:									
Treatment (vs Control)	1.05	(1.01, 1.10)	.017	-0.62	(-1.05, -0.18)	.006	-0.03	(-0.22, 0.15)	.722
Sensitivity Analyses									
B. Falsification Models:									
Treatment (vs Control)	0.99	(0.96, 1.03)	.688	0.16	(-0.26, 0.57)	.466	0.03	(-0.05, 0.11)	.430
C. Difference-in-Differences Models:									
Treatment × Phase Interaction (DiD estimate)	1.05	(1.01, 1.08)	.005	-0.53	(-0.97, -0.08)	.020	-0.04	(-0.17, 0.10)	.574

Note: Panel A presents primary intent-to-treat estimates of the association between treatment assignment and early therapy engagement outcomes. Panel B presents falsification (placebo) analyses applying the same model specifications to a pre-intervention period in which no treatment effect is expected. Panel C presents difference-in-differences sensitivity analyses estimating the treatment-by-period interaction. Total therapy sessions were modeled using Poisson mixed-effects regression, with rate ratios representing relative differences in the expected number of sessions during the 7-week study period (e.g., a rate ratio of 1.05 corresponds to a 5% relative increase). Time to second therapy session was modeled using linear mixed-effects regression among participants who completed at least two sessions during the study period. Therapeutic alliance was operationalized as the participant-level mean of all alliance survey scores collected within 49 days (approximately 7 weeks) following the first therapy session and was modeled using linear mixed-effects regression; coefficients represent differences in alliance score units. All models adjusted for baseline PHQ-9, baseline GAD-7, age, and gender, and included a random intercept for employer.

Table 2. Clinical Symptom Trajectories Among Participants With Elevated Baseline Symptoms: Primary and Sensitivity Analyses.

Model	Depressive Symptoms (Baseline PHQ-9 ≥10)			Anxiety Symptoms (Baseline GAD-7 ≥10)		
	β (days)	**(95% CI)**	***p*-value**	**β (days)**	**(95% CI)**	***p*-value**
A. Primary Intervention Models:						
Treatment × Day Interaction	−0.021	(−0.026, −0.014)	<.001	−0.018	(−0.022, −0.013)	<.001
Marginal Effect at Median Length of Care (Day 44)	−0.71	(−0.92, −0.49)	<.001	−0.50	(−0.71, −0.29)	<.001
Sensitivity Analyses						
B. Falsification Models:						
Treatment × Day Interaction	−0.001	(−0.003, 0.001)	.099	0.000	(−0.001, 0.001)	.910
C. Difference-in-Differences Models:						
Treatment × Day × Phase Interaction (DiD estimate)	−0.019	(−0.025, −0.014)	<.001	−0.018	(−0.023, −0.013)	<.001

Note: Panel A presents primary intent-to-treat mixed-effects models estimating treatment-by-time interactions, with marginal treatment effects evaluated at the observed median length of care for these samples (44 days). Panel B presents falsification analyses applying identical model specifications to a pre-intervention period in which no treatment effect was expected. Panel C presents difference-in-differences sensitivity analyses estimating the treatment-by-time-by-period interaction. All models adjusted for baseline symptom severity, age, and gender, and included random intercepts for employer and participant.

Participant Flow Diagram

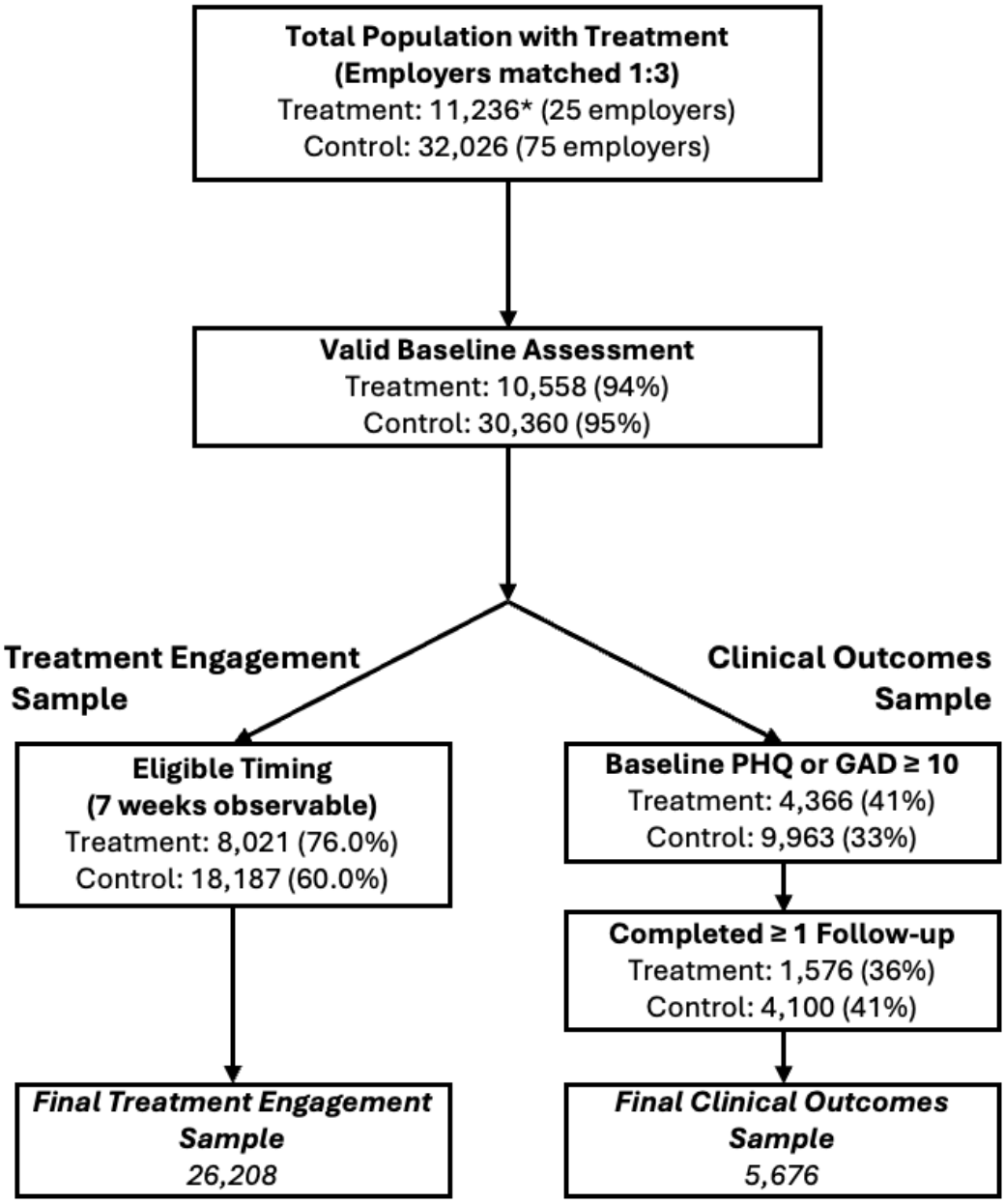

Figure 1. Eligible timing for treatment engagement models required 7 weeks of observable follow-up to avoid right-censoring. *39 treatment participants were removed from the sample because they were randomly selected to use an additional test-feature that was not part of the current study.

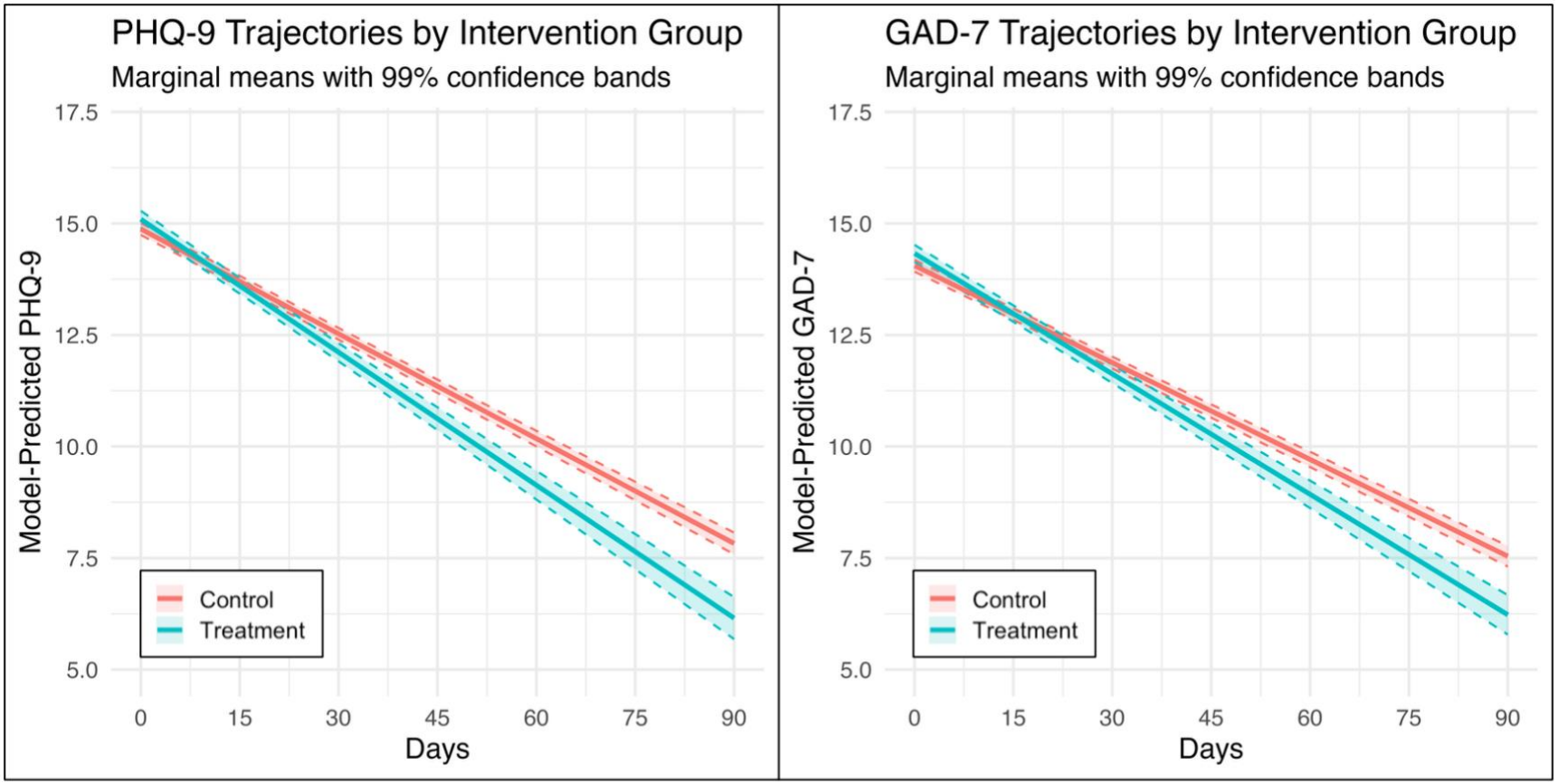

Figure 2. Model-predicted trajectories of depressive (PHQ-9) and anxiety (GAD-7) symptoms by intervention group. Lines represent marginal mean symptom scores estimated from mixed-effects linear models among participants with elevated baseline symptoms (PHQ-9 ≥10 or GAD-7 ≥10), adjusted for baseline severity, age, gender, and employer-level clustering. Shaded bands denote 99% confidence intervals. Differences in the rate of symptom improvement over time reflect the treatment × time interaction.

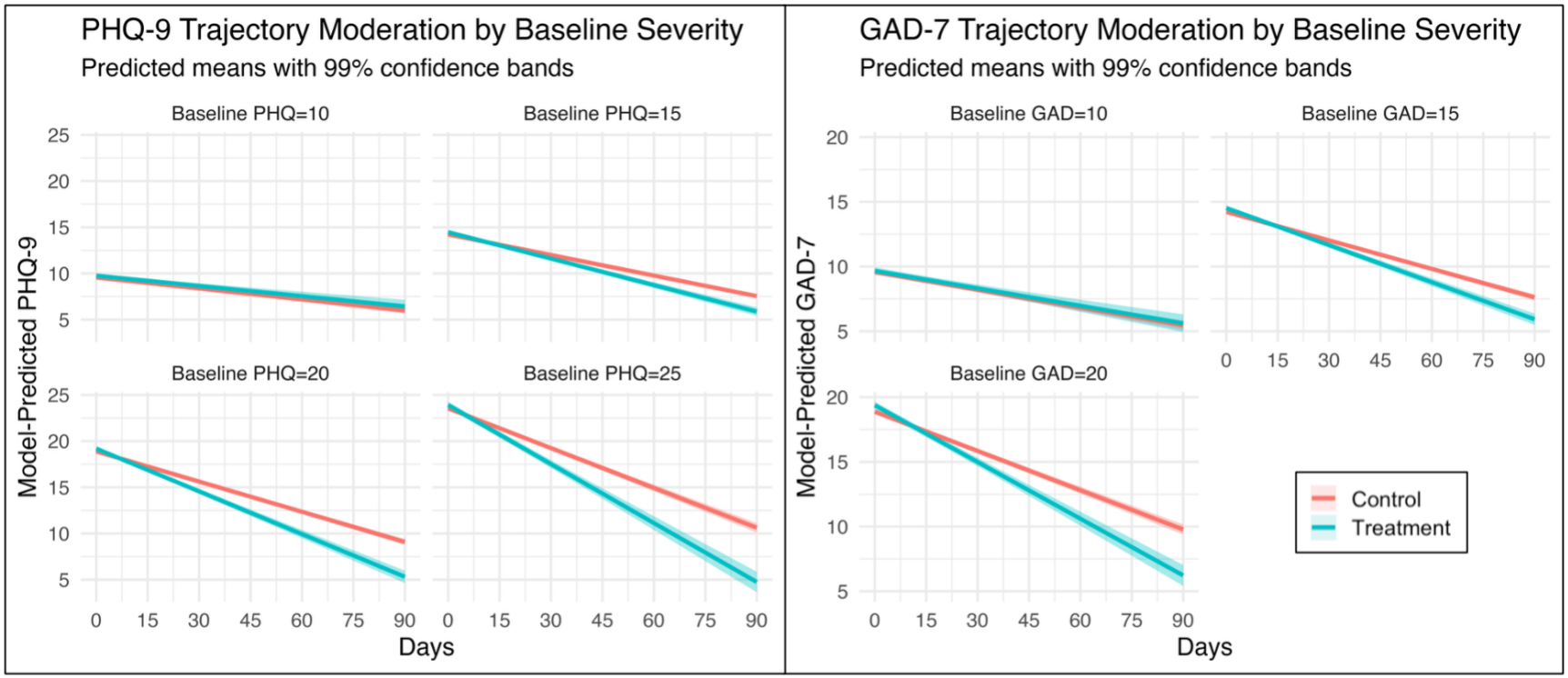

Figure 3. Model-predicted trajectories of depressive (PHQ-9) and anxiety (GAD-7) symptoms by baseline symptom severity. Lines represent marginal mean symptom scores estimated from mixed-effects linear models among participants with elevated baseline symptoms (PHQ-9 ≥10 or GAD-7 ≥10), adjusted for baseline severity, age, gender, and employer-level clustering. Shaded bands denote 99% confidence intervals. Facets display predicted trajectories at representative baseline severity levels, illustrating greater intervention-associated improvement over time among participants with higher baseline symptom burden.

Supplemental Materials

Employer-Level Matching Procedures

Employers receiving access to continuous care features ("intervention employers") were matched to employers without access ("comparison employers") using a 1:3 matching ratio. Matching was conducted at the employer level to reduce baseline differences between intervention and comparison groups and to improve comparability consistent with a cluster-randomized design, given that access to continuous care features was implemented at the employer rather than individual level.

Matching variables were prespecified and selected to capture key organizational characteristics plausibly associated with treatment access and engagement. These included number of covered lives (as a measure of employer size), industry, plan level (defined by the number of prepaid psychotherapy sessions included in the benefit), and prior treatment-seeking rate, operationalized as the proportion of covered lives with at least one behavioral health treatment encounter during the year preceding the study period. Matching was performed prior to participant-level outcome analyses and without reference to post-intervention outcomes.

Rollout Timing and Censoring

Access to AI-enabled continuous care features was implemented in a staged employer-level rollout. Six intervention employers received access on July 8, 2025, followed by an additional 19 intervention employers on August 13, 2025. General availability of continuous care features to all employer customers, including comparison employers, occurred on December 3, 2025. To avoid contamination of the comparison condition, all analyses were censored at the date of general availability. That is, to ensure comparability between intervention and matched control groups, data collected after December 3rd, 2025 were not used in the present study.

Outcome Definitions and Measurement

Early treatment engagement. Early treatment engagement outcomes were designed to capture initial uptake and momentum following psychotherapy initiation. Utilization during the early engagement window was operationalized as the number of attended psychotherapy sessions within 7 weeks of the initial session. Treatment momentum was additionally assessed using time to second psychotherapy session among participants with at least two attended sessions, as shorter time to follow-up reflects more rapid progression beyond initial access and is commonly used as a behavioral indicator of early engagement.

Therapeutic alliance. Therapeutic alliance was measured using an optional, session-level survey administered via email following psychotherapy visits. The survey was adapted from the Working Alliance Inventory and included three items reflecting core alliance domains: (1) bond (e.g., "I have a good working relationship with my provider"), (2) agreement on goals (e.g., "I understand and agree with the goals my provider and I have set"), and (3) agreement on tasks (e.g., "My provider and I agree on the best approach for addressing my problems"). Each item was rated on a 1–10 scale, with anchors of 1 ("not at all true"), 5 ("undecided"), and 10 ("completely true"). For participants with available data, alliance was operationalized as the participant-level mean of all alliance scores collected within the 7-week study period.

Clinical symptoms and reliable improvement. Depressive and anxiety symptom severity were assessed using the Patient Health Questionnaire–9 (PHQ-9) and the Generalized Anxiety Disorder–7 (GAD-7), respectively, collected at baseline and during routine measurement-based care follow-up. In addition to modeling continuous symptom trajectories, clinically reliable improvement was defined as a reduction of ≥5 points on the PHQ-9 or ≥4 points on the GAD-7 from baseline, consistent with prior validation studies. These thresholds were used to estimate the probability of reliable improvement at the median follow-up time in model-based analyses.

Assessment of Post-Matching Balance

Following employer-level matching, balance between intervention and comparison groups was assessed at the participant level within each analytic subsample using standardized mean differences (SMDs). Balance was evaluated separately for the treatment engagement analytic sample (participants with a complete ≥49-day observation window) and the clinical outcomes analytic sample (participants with PHQ-9 or GAD-7 score ≥10 and at least one follow-up assessment), as these samples differed slightly in eligibility criteria and follow-up structure.

Table S1 presents baseline demographic and clinical characteristics for treatment and control participants within each analytic subsample. Across both subsamples, standardized mean differences for all assessed covariates were small in magnitude and below the conventional threshold of |SMD| < 0.10, indicating good balance between groups after matching.

Across both analytic subsamples, baseline clinical symptom severity was highly comparable between treatment and comparison groups. SMDs for baseline PHQ-9 and GAD-7 scores were near zero in both the treatment engagement and clinical outcomes samples, indicating nearly identical baseline symptom severity across groups. Given that depressive and anxiety symptom severity represent the primary study outcomes and are most plausibly related to treatment engagement and improvement, this close balance provides particularly strong evidence of comparability between groups following employer-level matching.

Small residual differences were observed for select demographic characteristics, most notably gender composition, with standardized mean differences for women approaching but remaining below conventional thresholds for imbalance. All other covariates demonstrated standardized mean differences well below 0.10.

To further mitigate the potential influence of any residual baseline differences, all primary regression models were additionally adjusted for baseline covariates, including age, baseline symptom severity, and gender. This analytic approach combines employer-level matching with covariate adjustment at the participant level and is consistent with recommended best practices for observational and quasi-experimental studies. As a result, study findings reflect differences between groups that are robust to minor residual imbalances observed after matching.

Table S1. Baseline balance by analytic subsample after employer-level matching.

A. Treatment Engagement Analytic Sample (n = 26,208)

Characteristic	Treatment	Control	SMD
Baseline PHQ-9, mean (SD)	9.74 (6.3)	9.75 (6.2)	−0.002
Baseline GAD-7, mean (SD)	9.31 (5.6)	9.30 (5.5)	0.003
Age, mean (SD)	37.40 (11.7)	36.6 (11.7)	0.066
Man, %	28.3%	30.0%	−0.037
Woman, %	55.1%	50.8%	0.086
Other / unknown gender, %	16.6%	19.2%	−0.067
Dependent enrollee, %	21.9%	21.0%	0.021

B. Clinical Outcomes Analytic Sample (n = 5,676)

Characteristic	Treatment	Control	SMD
Baseline PHQ-9, mean (SD)	13.85 (5.6)	13.90 (5.7)	−0.008
Baseline GAD-7, mean (SD)	12.90 (4.7)	13.01 (4.7)	−0.022
Age, mean (SD)	37.1 (12.0)	36.5 (11.8)	0.052
Man, %	27.4%	27.8%	−0.009
Woman, %	56.2%	53.8%	0.049
Other / unknown gender, %	16.3%	18.4%	−0.054
Dependent enrollee, %	21.4%	22.0%	−0.014

Note: Values shown are standardized mean differences (SMDs), calculated as the difference in means or proportions between groups divided by the pooled standard deviation. Absolute SMD values less than 0.10 are commonly considered indicative of adequate balance between groups. To further account for any residual imbalance, all primary regression models are additionally adjusted for baseline covariates, including age, baseline symptom severity, and gender.

Detailed Description of Continuous Care Features and Session Recordings

Continuous care comprised three optional AI-enabled features designed to support engagement and continuity of care between psychotherapy sessions: guided intake, session summaries, and session takeaways. Each feature leveraged large language models to structure unstructured information, personalize content to the individual member, and reduce cognitive and logistical burden outside of live sessions.

Guided intake provided an AI-assisted conversational experience prior to the first psychotherapy session. Using a natural-language dialogue, the system prompted participants to describe their primary concerns, relevant history, symptoms, and treatment goals in their own words. AI models synthesized these free-text responses into a structured intake summary, organizing information into clinically relevant categories (e.g., presenting problems, goals, contextual factors). With participant consent, the intake summary was made available to the assigned provider ahead of the first session. The intended purpose of guided intake was to help participants reflect on and articulate their needs before initiating care, while also enabling providers to begin treatment with clearer context, potentially reducing time spent on administrative intake and supporting earlier therapeutic alignment.

Guided intake was introduced approximately one month into the study period (mid-August), such that participants initiating care early in the rollout period did not have the opportunity to use this feature prior to their first psychotherapy session, though it remained available thereafter. This staggered availability reflects real-world implementation conditions for digital care tools and, under the intent-to-treat analytic framework used here, would be expected to attenuate rather than inflate observed treatment effects.

Session summaries used AI to generate structured summaries of psychotherapy sessions that were recorded and transcribed with explicit consent from both participant and provider. Large language models processed session transcripts to identify key discussion themes, therapeutic focus areas, and high-level progress indicators, producing a concise, non-verbatim summary of each session. Summaries were delivered to participants after sessions to support recall, reflection, and continuity across appointments. By externalizing session content into a persistent artifact, the feature aimed to reinforce therapeutic insights, reduce reliance on memory alone, and help participants maintain momentum between sessions without requiring additional effort from providers.

Session takeaways complemented session summaries by generating brief, action-oriented follow-up points drawn directly from guidance, exercises, or recommendations articulated by the provider during the session. AI models identified these clinician-directed elements within the session transcript and translated them into clear, participant-facing takeaways, without introducing new clinical content. Takeaways were intended to support between-session engagement by clarifying provider-recommended next steps and encouraging application of therapeutic strategies discussed during the session. Whereas summaries emphasized reflection and continuity, takeaways emphasized behavioral follow-through and practical application of provider guidance.

All features were available to participants in intervention employers; however, feature use was optional and participant-driven. Uptake of continuous care features varied across participants and was lower than anticipated at the time of preregistration, particularly for session summaries and takeaways, which required additional consent and workflow integration. As a result, all primary analyses were conducted using an intent-to-treat framework based on employer-level access to continuous care features. Observed uptake rates by feature are summarized in Table S2.

Access to session summaries and takeaways required that both the provider and the participant consent to having the psychotherapy session recorded and transcribed, with all identifying information removed prior to processing. Lower-than-anticipated opt-in rates—

particularly among providers—limited the number of sessions eligible for these features and contributed to the observed utilization rates. Provider and participant opt-in rates are shown in Table S3.

Table S2. Utilization rates of the continuous care features.

Continuous Care Feature	Treatment Engagement Sample	Clinical Outcomes Sample
Any continuous care feature use	*38.2%*	*47.0%*
Used takeaways	25.0%	36.8%
Used session summaries	15.6%	22.7%
Used guided intake	16.4%	14.1%

Table S3. Provider and participant opt-in rates to session recordings (needed for transcript generation).

Analytic Sample	Opt-In Metric	Provider Opt-In	Participant Opt-In
Treatment Engagement Sample	Session-level opt-in rate	60.3%	71.6%
Treatment Engagement Sample	Ever opt-in (≥1 session)	67.3%	85.6%
Treatment Engagement Sample	First-session opt-in rate	56.9%	74.7%
Clinical Outcomes Sample	Session-level opt-in rate	59.9%	73.8%
Clinical Outcomes Sample	Ever opt-in (≥1 session)	67.2%	89.8%
Clinical Outcomes Sample	First-session opt-in rate	55.6%	76.2%

Statistical Analyses

All analyses were conducted at the participant level under an intention-to-treat framework based on employer-level assignment to AI-enabled continuous care features. All regression models adjusted for baseline symptom severity, age, and self-reported gender. Treatment engagement outcomes were analyzed using regression models appropriate for the distribution of each outcome (e.g., negative binomial models for count outcomes). Employer-level clustering was accounted for using random intercepts.

Clinical outcomes were analyzed using mixed-effects regression models to account for repeated symptom assessments within participants and clustering within employers. Models included fixed effects for treatment group, time since psychotherapy initiation, and their interaction, with random intercepts for both participants and employers. Participants without follow-up assessments were retained in all models to preserve the intention-to-treat estimand, but only those with follow-up data contributed to the symptom trajectory coefficients of interest.

To estimate clinically interpretable effects, population-average predicted symptom trajectories were used to calculate the probability of clinically reliable improvement at the median follow-up time (44 days). Reliable improvement was defined as a reduction of ≥5 points on the PHQ-9 or ≥4 points on the GAD-7 from baseline. Individual-level probabilities were derived from model-implied outcome distributions using population-average predictions and averaged within treatment and comparison groups. Numbers needed to treat were calculated as the inverse of the absolute difference in improvement probabilities.

Sensitivity Analyses. To evaluate the robustness of estimated treatment effects and assess the potential influence of residual confounding or secular trends, we conducted a series of prespecified sensitivity analyses designed to test key identifying assumptions of the primary analytic approach.

First, to further assess whether observed treatment effects could be attributed to time-varying confounding or cohort effects unrelated to the intervention, temporal falsification analyses were conducted using data from the same 100 employers during the year preceding intervention rollout. Identical model specifications—including covariates, random effects structure, and outcome definitions—were applied to this pre-implementation cohort (n=36,294), with a pseudo-implementation date assigned to mirror the timing of the actual rollout. Under this falsification test, detection of a treatment-associated effect in the pre-implementation period would suggest residual confounding, whereas null effects would provide evidence that post-implementation differences are unlikely to reflect secular trends or unobserved employer-level differences.

Difference-in-differences (DiD) models were estimated to compare changes in symptom trajectories between intervention and comparison groups before and after implementation of AI-enabled continuous care features. These models contrasted the effect found among the present intervention sample to a pre-implementation sample (i.e., the year prior to launch), where the DiD treatment effect was defined as the differential change in slopes following rollout. Under this framework, a null or comparable change in slopes between groups during the pre-implementation period would support the parallel trends assumption, whereas emergence of differential improvement only after implementation would be consistent with a causal intervention effect.

Together, these sensitivity analyses provide complementary tests of the core identifying assumptions underlying the primary analyses and support the interpretation that observed post-implementation differences in clinical outcomes reflect the effect of access to AI-enabled continuous care features rather than artifacts of timing, regression to the mean, or unmeasured baseline differences. All analyses were conducted using R (version 4.4.3).

Predictors of Continuous Care Feature Utilization

To examine whether baseline clinical or demographic characteristics were associated with use of continuous care features among participants with access, we conducted an exploratory analysis restricted to participants in intervention employers. Any continuous care feature use was defined as use of at least one AI-enabled feature (guided intake, session summaries, or session takeaways) during the 7-week observation window. Associations between baseline characteristics and feature utilization were estimated using a mixed-effects logistic regression model with a random intercept for employer to account for clustering. Fixed-

effect covariates included baseline PHQ-9 and GAD-7 scores, age, gender, and dependent enrollee status.

In this analysis of 8,021 participants nested within 25 employers, no strong associations were observed between baseline clinical severity or demographic characteristics and use of continuous care features (Table S4). Baseline PHQ-9 and GAD-7 scores were not associated with feature uptake, and effect estimates for age, gender, and dependent enrollee status were small in magnitude. Overall, these findings indicate that utilization of continuous care features among participants with access was not strongly patterned by baseline symptom severity or demographic factors.

Table S4. Predictors of any continuous care feature use among participants in intervention employers (n=8,021)

Covariate	Odds Ratio (95% CI)	*p*-value
Baseline PHQ-9 (per point)	1.00 (0.99–1.01)	.468
Baseline GAD-7 (per point)	1.00 (0.99–1.02)	.493
Age (per year)	1.00 (0.99–1.00)	.124
Woman (vs man)	1.08 (0.97–1.20)	.185
Other/unknown gender (vs man)	1.09 (0.88–1.35)	.427
Dependent enrollee (vs employee)	0.94 (0.79–1.12)	.495

Deviations from Preregistration

This study was preregistered prior to the launch of the AI-enabled continuous care features, with the goal of prospectively specifying key design and analytic decisions (https://aspredicted.org/s7vr-6m7k.pdf). As is common in evaluations embedded within real-world program rollouts, several aspects of implementation evolved in ways that could not be fully anticipated at the time of preregistration. In particular, program development priorities, staggered employer launches, variation in provider and patient opt-in to session recordings and specific features, and the transition to general availability necessitated targeted analytic refinements to preserve internal validity, statistical power, and interpretability of treatment effects. Importantly, the core study design, comparison strategy, outcome measures, and causal estimands remained closely aligned with the preregistered plan. Where deviations occurred, they reflected practical constraints inherent to naturalistic implementation and were made to strengthen inference, reduce bias, or avoid underpowered or selectively defined analytic samples. All substantive deviations from the preregistered analysis plan are transparently documented below, along with their rationale and implications for interpretation.

Table S5. Summary of Deviations from the Preregistered Analysis Plan

Preregistered Plan	Deviation	Rationale and Impact
Employer-level design: Six employers assigned to the intervention condition and 12 matched employers assigned to control (2:1 matching).	**Expanded employer sample:** Twenty-five intervention employers matched to 75 comparison employers using a 1:3 matching ratio (total n=100 employers).	The scale of the real-world rollout exceeded initial projections, allowing inclusion of substantially more employers without altering the cluster-level assignment mechanism. Increasing the number of clusters improved statistical power, balance, and precision of employer-level estimates while preserving the original quasi-experimental design.
Primary analytic focus on opt-in participants: Primary analyses restricted to treatment participants who opted into session recordings and received the full AI feature set, with individual-level matching to controls.	**Intent-to-treat (ITT) primary analyses:** All participants in intervention employers were included regardless of feature opt-in, with exposure defined at the employer level.	Restricting analyses to opt-in participants would have introduced post-assignment selection bias due to partial uptake and consent-based feature access. The ITT approach preserves the original cluster-level assignment, avoids post-assignment selection bias, and yields more conservative estimates of program impact. Because ITT estimates average effects across all eligible participants, including those who did not engage with intervention features, estimated effects may be attenuated relative to treatment-on-the-treated contrasts; however, the substantially larger analytic sample increases statistical power to detect even modest but policy-relevant effects. This change strengthens causal interpretability relative to the preregistered plan.
Clinical outcomes timing: Assess changes in anxiety and depressive symptoms from baseline to a fixed 7-week post-treatment follow-up.	Clinical outcomes were modeled as continuous symptom trajectories up to 6 months following psychotherapy initiation, with follow-up restricted to observations occurring before the general availability launch date to avoid contamination.	A fixed 7-week endpoint would have discarded substantial valid follow-up data and reduced power. Modeling continuous trajectories better reflects real-world care patterns, improves estimation efficiency, and allows treatment effects to accumulate over clinically meaningful time horizons while preserving the original estimand of differential symptom improvement.
Clinical improvement assessment: Estimate change from baseline to a single post-treatment follow-up at 7 weeks, with treatment effects evaluated at that endpoint.	Clinical outcomes were analyzed as longitudinal symptom trajectories over time using group × time interactions. Robustness of estimated intervention effects were evaluated using large-scale falsification and difference-in-differences analyses applied to an independent pre-intervention cohort from the same employers in the year prior to launch.	Focusing inference on a single post-treatment endpoint would not have permitted direct tests distinguishing true intervention effects from secular trends or cohort differences. Modeling symptom trajectories enabled explicit tests of whether differential improvement emerged only following intervention rollout. Incorporating a full pre-launch cohort from the same employers allowed direct falsification of the treatment effect under identical model specifications, providing strong evidence that observed post-launch differences reflect the causal impact of the intervention rather than artifacts of timing, regression to the mean, or unobserved employer-level differences. This approach strengthens causal inference relative to the preregistered endpoint-based analysis.
Small-sample cluster corrections were planned due to a limited number of employer clusters (n=18).	Large-cluster mixed-effects models supplemented with falsification and difference-in-differences sensitivity analyses.	The final sample included a substantially larger number of clusters than anticipated (i.e., 100 employers), reducing reliance on small-sample corrections. The use of falsification tests and DiD models provides stronger evidence against residual confounding than cluster-adjustment alone, improving inferential robustness.

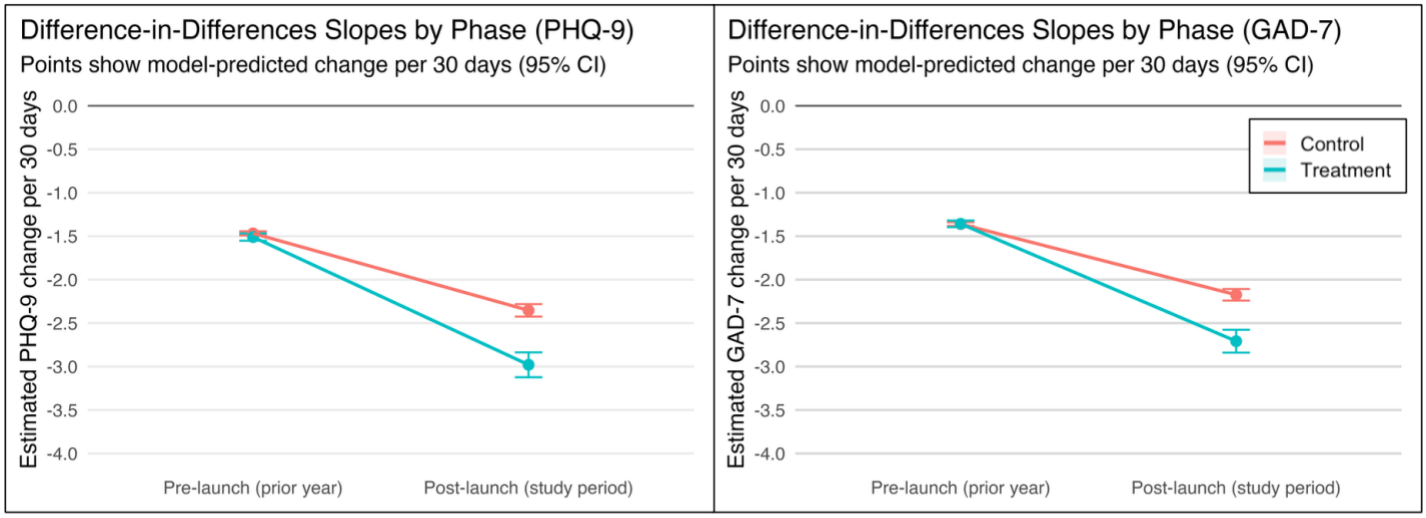

Figure S1. Difference-in-differences estimates of symptom improvement slopes. Slopes are shown for intervention and comparison groups in the pre-launch (prior year) and post-launch (current intervention period) phases. The difference-in-differences estimand corresponds to the differential change in improvement slopes from the pre- to post-launch period between groups.

References

1. Cuijpers P, Karyotaki E, Eckshtain D, et al. Psychotherapy for Depression Across Different Age Groups: A Systematic Review and Meta-analysis. *JAMA Psychiatry*. 2020;77(7):694-702. doi:10.1001/jamapsychiatry.2020.0164

2. Cuijpers P, Sijbrandij M, Koole SL, Andersson G, Beekman AT, Reynolds III CF. The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: a meta-analysis of direct comparisons. *World Psychiatry*. 2013;12(2):137-148. doi:10.1002/wps.20038

3. Cuijpers P, Miguel C, Ciharova M, Ebert D, Harrer M, Karyotaki E. Transdiagnostic treatment of depression and anxiety: a meta-analysis. *Psychol Med*. 2023;53(14):6535-6546. doi:10.1017/S0033291722003841

4. Munder T, Flückiger C, Leichsenring F, et al. Is psychotherapy effective? A re-analysis of treatments for depression. *Epidemiol Psychiatr Sci*. 2018;28(3):268-274. doi:10.1017/S2045796018000355

5. Wampold BE, Budge SL, Laska KM, et al. Evidence-based treatments for depression and anxiety versus treatment-as-usual: A meta-analysis of direct comparisons. *Clin Psychol Rev*. 2011;31(8):1304-1312. doi:10.1016/j.cpr.2011.07.012

6. Ringeisen H, Edlund M, Guyer H, et al. Prevalence of Past-Year Mental and Substance Use Disorders, 2021–2022. *Psychiatr Serv*. 2025;76:720-728. doi:10.1176/appi.ps.20240329

7. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*. 2013;382(9904):1575-1586. doi:10.1016/S0140-6736(13)61611-6

8. Baxter AJ, Scott KM, Vos T, Whiteford HA. Global prevalence of anxiety disorders: A systematic review and meta-regression. *Psychol Med*. 2013;43:897-910. doi:10.1017/S003329171200147X

9. Cuijpers, Pim, Karyotaki, Eirini. A Meta-Analytic Database Of Randomized Trials On Psychotherapies For Depression. *Open Science Foundation*. Preprint posted online 2021. doi:https://doi.org/ 10.17605/OSF.IO/825C6.

10. Swift JK, Greenberg RP, Tompkins KA, Parkin SR. Treatment refusal and premature termination in psychotherapy, pharmacotherapy, and their combination: A meta-analysis of head-to-head comparisons. *Psychotherapy*. 2017;54(1):47-57. doi:10.1037/pst0000104

11. Langlieb AM, Langlieb ME, Xiong W. EAP 2.0: reimagining the role of the employee assistance program in the new workplace. *Int Rev Psychiatry*. 2021;33(8):699-710. doi:10.1080/09540261.2021.2013172

12. Attridge M. A global perspective on promoting workplace mental health and the role of employee assistance programs. *Am J Health Promot*. 2019;33(4):620-633. doi:10.1177/0890117119838101

13. Ward EJ, Fragala MS, Birse CE, et al. Assessing the impact of a comprehensive mental health program on frontline health service workers. *PLoS ONE*. 2023;18(11 November):1-15. doi:10.1371/journal.pone.0294414

14. Baum, Graham, Hawrilenko, Matt, Cascalheira, Cory, et al. Mental Health Service Use and Equity in a Comprehensive Employer-Sponsored Benefit Program: A Retrospective Cohort Study. *Popul Health Manag*. Published online 2026. doi:10.1177/19427891261420041

15. Hawrilenko M, Smolka C, Ward E, Kavelaars R, Brown M, Chekroud AM. The Impact of Enhanced Behavioral Health Services on Total Healthcare Costs Among US Employers: A Site-Level Analysis of 19 Cohort Studies. *J Health Econ Outcomes Res*. 2025;12(1):246-251. doi:10.36469/001c.138634

16. Bondar J, Babich Morrow C, Gueorguieva R, et al. Clinical and financial outcomes associated with a workplace mental health program before and during the COVID-19 pandemic. *JAMA Netw Open*. 2022;5(6):e2216349. doi:10.1001/jamanetworkopen.2022.16349

17. Graham S, Depp C, Lee EE, et al. Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Curr Psychiatry Rep*. 2019;21(11):116. doi:10.1007/s11920-019-1094-0

18. Ali M, Ali S, Abbas Q, Abbas Z, Lee SW. Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health. *Digit Health*. 2025;11:20552076251395548. doi:10.1177/20552076251395548

19. D'Alfonso S. AI in mental health. *Curr Opin Psychol*. 2020;36:112-117. doi:10.1016/j.copsyc.2020.04.005

20. Shimada K. The Role of Artificial Intelligence in Mental Health: A Review. *Sci Insights*. 2023;43(5):1119-1127. doi:10.15354/si.23.re820

21. Hawrilenko M, Smolka C, Ward E, et al. Return on Investment of Enhanced Behavioral Health Services. *JAMA Netw Open*. 2025;8(2):1-14. doi:10.1001/jamanetworkopen.2024.57834

22. Graupensperger S, Hawrilenko M, Brown M, Baum G, Ward EJ, Chekroud A. Crisis Outreach, Treatment Engagement, and Outcomes After Suicide Risk Screening in a Comprehensive Mental Health Platform. *Psychiatr Serv*. Published online November 14, 2025. Accessed November 14, 2025. https://psychiatryonline.org/doi/10.1176/appi.ps.20250319

23. Ward, E. J., Hawrilenko, M., Wu, M., Novak, J., Brown, Millard, Chekroud, Adam. Large-Scale Evaluation of Provider-Patient Matching in an Employer-Sponsored Mental Health Program. In Revision.

24. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity scale. *J Gen Intern Med*. 2001;16:606-613.

25. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch Intern Med*. 2006;166:1092-1097. doi:10.1001/archinte.166.10.1092

26. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *J Affect Disord*. 2010;127(1-3):122-129. doi:10.1016/j.jad.2010.04.030

27. Toussaint A, Hüsing P, Gumz A, et al. Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7). *J Affect Disord*. 2020;265:395-401. doi:10.1016/j.jad.2020.01.032

28. Kazantzis N, Whittington C, Zelencich L, Kyrios M, Norton P, Hofmann S. Quantity and Quality of Homework Compliance: A Meta-Analysis of Relations With Outcome in Cognitive Behavior Therapy. *Behav Ther*. 2016;47. doi:10.1016/j.beth.2016.05.002

29. Dong L, Zhao X, Ong SL, Harvey AG. Patient recall of specific cognitive therapy contents predicts adherence and outcome in adults with major depressive disorder. *Behav Res Ther*. 2017;97:189-199. doi:10.1016/j.brat.2017.08.006

30. Ryum T, Bennion M, Kazantzis N. Homework as a driver of change in psychotherapy. *J Clin Psychol*. 2024;80(4):733-743. doi:10.1002/jclp.23627

31. Cuijpers P, Harrer M, Miguel C, Karyotaki E, Papola D. Does baseline severity interact with the effects of psychotherapy for depression? A meta-analytic review. *J Affect Disord*. 2026;399:121106. doi:10.1016/j.jad.2025.121106

32. Bruijniks S, Harvey AG, Hollon SD, et al. Use and Effects of Therapist Memory Support Strategies in Cognitive Behavioral Therapy and Interpersonal Psychotherapy for Depression. *Cogn Ther Res*. 2025;49(5):1017-1030. doi:10.1007/s10608-024-10569-8

33. Harvey AG. Maximizing benefits from evidence-based psychological treatments: Memory support and habit formation as key strategies. *Behav Res Ther*. 2025;191:104767. doi:10.1016/j.brat.2025.104767