PSILOGOS

Research Article

open-access

When Treatment Choice Distorts Apparent Comparability: Preference–Attrition Tipping-Point Analysis of Therapist-Supported Cognitive–Behavioral Therapy for Depression

Michael Schmidt

ABSTRACT

Comparisons of electronically delivered and in-person cognitive–behavioral therapy (CBT) become difficult when patients choose their delivery mode. Pretreatment severity may influence that choice, and incomplete follow-up may conceal clinically important differences even when conventional interaction tests are not statistically significant. This study asked how much residual preference-related distortion would be required to overturn an endpoint conclusion of comparable proportional improvement, and whether completion and persistence evidence supported the same interpretation. A reproducible aggregate cohort was constructed from reported arm-level summaries for 108 adults with major depressive disorder who selected therapist-supported electronic CBT (e-CBT; $n = 53$) or in-person CBT ($n = 55$). Preference–Attrition Tipping-Point Analysis (PATPA) combined four elements: Hedges-corrected pretreatment contrasts, Jeffreys-prior beta–binomial completion estimates, a conditional analysis of sessions completed before withdrawal, and a rounding-aware bias boundary for normalized Patient Health Questionnaire-9 (PHQ-9) change. A 10-percentage-point region was specified for completion and normalized endpoint contrasts. Pretreatment symptom differences were substantial: Hedges' g for e-CBT minus in-person CBT was -0.72 for PHQ-9 and -0.97 for the Quick Inventory of Depressive Symptomatology–Self Report; quality of life favored e-CBT ($g = 0.45$). Completion proportions were nearly identical (45.3% and 45.5%), but the posterior 95% credible interval for their difference was wide (-18.5 to 18.2 percentage points), and the probability of lying within the prespecified 10-point region was 0.710. Among participants who withdrew, e-CBT was associated with 1.65 additional sessions before withdrawal (95% interval 0.40–2.90). Normalized PHQ-9 values at week 12 were 0.78 and 0.79, producing an endpoint contrast of -0.01. The comparability conclusion held only while residual preference bias remained between -0.11 and 0.09 of the pretreatment level. Apparent endpoint similarity was compatible with modest residual preference distortion, but it was not immune to it. Completion equality was imprecisely estimated, whereas persistence among withdrawals favored e-CBT conditionally. Delivery-mode conclusions should therefore report a bias boundary and attrition uncertainty rather than rely on a non-significant treatment-by-time interaction.

Keywords: cognitive–behavioral therapy; major depressive disorder; treatment preference; attrition; sensitivity analysis; digital psychotherapy; tipping point

Department of Psychiatry and Psychotherapy, TUM University Hospital, Technical University of Munich, Munich, Germany
michael.schmidt85@gmail.com

– **Published:** 20 November 2024

© 2024 The Author(s). Published by Psilogos. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

1. INTRODUCTION

Major depressive disorder is defined not only by depressed mood and loss of interest but also by disturbances in motivation, cognition, sleep, energy, and social functioning. Treatment decisions therefore concern more than the reduction of a single questionnaire score. Canadian guidance places structured psychological care, including CBT, among the central evidence-based options for adults with depression.[1] CBT seeks to alter mutually reinforcing patterns of interpretation and action, and skill acquisition during therapy is associated with differential symptom relief.[2] Its preventive value is also clinically relevant because cognitive and behavioral modifications may reduce relapse after acute improvement.[3] These properties have encouraged structured measurement of symptom severity and quality of life across treatment. The PHQ-9 is widely used to monitor depressive change,[4] while the Quick Inventory of Depressive Symptomatology–Self Report (QIDS-SR) provides a complementary account of depressive domains.[5] Quality-of-life assessment broadens evaluation beyond symptoms; the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q) was developed for this purpose,[6] and later work supports its factor structure in clinical populations.[7]

Digital delivery changes the practical conditions under which the same therapeutic principles are encountered. Online psychotherapy can reduce travel, scheduling, and geographic constraints while permitting structured assignments and asynchronous clinician feedback.[8] Routine-care evidence suggests that internet CBT can improve depression and anxiety, although effect sizes vary across services and populations.[9] Individual-participant network evidence indicates that guided internet CBT is effective for depression and that intervention components and patient characteristics can alter response.[10] Component-level synthesis likewise suggests that personalizing support and intervention elements may matter as much as the broad digital label.[11] Direct comparisons have frequently shown similar average improvement between internet-supported and face-to-face CBT,[12] and comparative meta-analysis has found guided self-help capable of producing outcomes close to face-to-face psychotherapy under some conditions.[13] Computer assistance may also change therapist time and treatment efficiency rather than simply reproduce a conventional session on a screen.[14,15] Consequently, delivery mode should be treated as a bundle of contact pattern, effort, accessibility, feedback timing, and self-management demands rather than as a binary technological attribute.

Patient choice further complicates interpretation. Adults do not select psychotherapy delivery at random. Preferences can depend on symptom intensity, privacy, transport, perceived need for human contact, confidence with technology, and beliefs about treatment credibility. Survey evidence demonstrates meaningful variation in preference for in-person and digital psychotherapy options for depression.[16] Shared decision making can improve the fit between care and patient priorities, but it also means that treatment groups may differ before therapy begins.[17] Home-based evidence-based psychotherapy may reduce access barriers for some patients while remaining unsuitable for others.[18] A preference design therefore offers strong clinical realism and weak protection against selection bias. If participants with greater depressive severity preferentially choose in-person treatment, a raw difference after treatment combines delivery-mode response with the severity pattern that existed before the first session.

The therapeutic relationship adds another layer. Working alliance in blended and face-to-face depression treatment can differ in form without being absent online.[19] Alliance has predicted change during blended CBT[20] and long-term outcome in guided internet treatment.[21] Comparable associations have been observed when CBT is administered face to face or by telephone.[22] Therapist support and the way an online intervention is presented can also affect adherence.[23] These findings caution against attributing an outcome solely to “online” or “in-person” delivery. They also show why attrition needs separate attention: discontinuation may be related to severity, convenience, alliance, perceived improvement, dissatisfaction, or circumstances outside therapy.

Conventional repeated-measures analysis can establish whether scores change over time and whether the average time course differs statistically across arms. It cannot, by itself, demonstrate that two treatments are equivalent. Failure to reject an interaction null hypothesis is compatible with a narrow difference, a clinically important difference estimated imprecisely, or opposing biases that obscure one another. The problem is especially pronounced in a modest preference cohort with more than half of participants not reaching the endpoint. A clinically useful interpretation requires at least four distinct questions. Were the arms comparable before treatment? How precisely was completion parity estimated? Did participants who withdrew remain engaged for different lengths of time? How much unmeasured preference-related distortion would be sufficient to move the endpoint contrast outside a stated region?

This article introduces Preference–Attrition Tipping-Point Analysis (PATPA) to answer those questions from transparent arm-level quantities. The research question is: among adults who selected therapist-supported CBT delivery, how large must residual preference-related distortion be to reverse a conclusion of comparable proportional PHQ-9 improvement at 12 weeks, and do completion and conditional persistence estimates lead to the same conclusion? The question is deliberately narrower than asking whether one mode is superior. It focuses on the stability of the comparison, distinguishes completion from time retained before withdrawal, and treats pretreatment imbalance as substantive evidence about choice rather than as a nuisance to be hidden by a single

adjusted coefficient.

2. STUDY AIM AND ANALYTIC DESIGN

2.1. Research question and target quantities

PATPA separates the comparison into five target quantities. The first is standardized pretreatment imbalance for age, PHQ-9, QIDS-SR, and Q-LES-Q. The second is the arm difference in the probability of completing all 12 sessions. The third is the mean difference in sessions completed among participants who eventually withdrew. The fourth is the arm contrast in PHQ-9 expressed as a proportion of each arm's pretreatment mean at weeks 6 and 12. The fifth is the residual preference bias that would move the week-12 contrast outside a prespecified region of ± 0.10 normalized units. Together these quantities distinguish selection, completion, conditional persistence, observed change, and robustness.

The analysis does not identify a causal effect of delivery mode because assignment followed preference and participant-level covariates were unavailable. Causal effects in non-randomized research require explicit treatment versions, exchangeability assumptions, and defensible adjustment variables.[24] Matching and weighting can improve covariate comparability when individual records are available;[25,26] those procedures cannot be reconstructed credibly from arm means. PATPA instead reports what the arm-level evidence can support and calculates the point at which a conclusion would change. This is closer to sensitivity analysis for hidden bias[27] and quantitative evaluation of unmeasured confounding[28,29] than to an adjusted causal estimate.

2.2. Numerical cohort

The numerical cohort comprised 113 assessed adults, five of whom did not begin treatment, leaving 108 treatment starters. Participants selected therapist-supported e-CBT ($n = 53$) or in-person CBT ($n = 55$). The reported ages were available for 48 e-CBT and 55 in-person participants. Women accounted for 33 and 37 participants, men for 15 and 18, and five e-CBT participants did not indicate sex. Pretreatment mean (standard error) values for e-CBT and in-person CBT were 16.08 (0.70) and 19.91 (0.73) on PHQ-9, 14.61 (0.67) and 19.29 (0.63) on QIDS-SR, and 36.31 (1.19) and 32.09 (1.32) on Q-LES-Q. Twenty-four e-CBT and 25 in-person participants completed treatment. The reported normalized PHQ-9 values at weeks 0, 6, and 12 were 1.00, 0.81, and 0.78 for e-CBT and 1.00, 0.89, and 0.79 for in-person CBT. All numerical inputs originated from the published trial report.[30]

The entries in Table 1 show why a single arm-by-time test is insufficient. Symptom severity and quality of life already differed materially before treatment, whereas completion counts were almost identical. At the same time, the temporal distribution of withdrawal differed: more in-person withdrawals occurred during the first six weeks, while e-CBT withdrawals were more evenly divided between treatment halves. PATPA preserves these distinctions instead of collapsing them into one adherence indicator.

Figure 1. Asynchronous e-CBT setting

Figure 2. In-person CBT setting

2.3. Clinical delivery and measurement context

Both delivery modes used therapist-supported CBT over 12 weeks. E-CBT was asynchronous: participants reviewed structured online material, completed assigned work, and received personalized written feedback. In-person participants attended weekly therapist meetings and discussed assigned work during the session. Figures 1 and 2 provide privacy-preserving visual context for the different treatment environments. The images are staged depictions made for this manuscript; they do not portray research participants and do not constitute outcome evidence.

The home setting in Figure 1 emphasizes the practical features that may influence selection and persistence: treatment can be accessed without travel, written work can be completed at a chosen time, and therapist contact occurs through delayed feedback rather than a simultaneous conversation. These features can reduce some burdens while increasing the need for self-directed initiation and digital access.

The consultation-room setting in Figure 2 highlights immediate interpersonal exchange, scheduled attendance, and the travel and time commitments associated with face-to-face care. The contrast between settings is there-

Table 1. Arm-level analytic inputs

Quantity	e-CBT	In-person CBT
Treatment starters, n	53	55
Age, mean (SD), years	38.60 (13.61)[a]	36.36 (11.61)
Women / men / not indicated, n	33 / 15 / 5	37 / 18 / 0
PHQ-9 pretreatment, mean (SE)	16.08 (0.70)	19.91 (0.73)
QIDS-SR pretreatment, mean (SE)	14.61 (0.67)	19.29 (0.63)
Q-LES-Q pretreatment, mean (SE)	36.31 (1.19)	32.09 (1.32)
Completed 12 sessions, n (%)	24 (45.3)	25 (45.5)
Withdrew during weeks 1–6, n	18	24
Withdrew during weeks 7–12, n	11	6
Sessions before withdrawal, mean (SE)	4.65 (0.46)	3.00 (0.44)
Normalized PHQ-9, weeks 0 / 6 / 12	1.00 / 0.81 / 0.78	1.00 / 0.89 / 0.79

[a] Age was reported for 48 e-CBT participants and all 55 in-person participants. QIDS-SR, Quick Inventory of Depressive Symptomatology–Self Report; Q-LES-Q, Quality of Life Enjoyment and Satisfaction Questionnaire.

fore behavioral and logistical as well as technological. Preference may encode perceived need for direct support, capacity to travel, comfort with written communication, and confidence in self-paced work.

PHQ-9 and QIDS-SR were treated as separate depressive symptom measures rather than interchangeable repetitions. PHQ-9 is concise and suitable for monitoring, whereas QIDS-SR samples a broader set of depressive symptom domains. Q-LES-Q was retained as a positively oriented outcome, so a higher score indicates better perceived life satisfaction. This orientation matters for standardized contrasts: negative values favor e-CBT for depressive symptoms because lower scores indicate less severity, while positive values favor e-CBT for Q-LES-Q because higher scores indicate better quality of life.

2.4. Preference–Attrition Tipping-Point Analysis

2.4.1. Standardized pretreatment imbalance

Reported standard errors were converted to standard deviations by $s_j = SE_j\sqrt{n_j}$ for arm j. For each continuous quantity, the standardized difference was calculated as

$$d = \frac{\bar{x}_e - \bar{x}_p}{s_{\text{pool}}}, \qquad s_{\text{pool}} = \sqrt{\frac{(n_e-1)s_e^2 + (n_p-1)s_p^2}{n_e+n_p-2}}. \tag{1}$$

Small-sample correction used $g = Jd$, where $J = 1 - 3/[4(n_e+n_p)-9]$. This choice follows established work on standardized mean differences in pretest–posttest comparisons[31] and the small-sample distribution of effect-size estimators.[32] Values were interpreted by magnitude and direction rather than by a rigid label; standardized effects are most useful when their scale, uncertainty, and substantive meaning are stated.[33] Equation 1 does not remove confounding. It measures the extent to which treatment choice produced different arms before therapy.

2.4.2. Completion probability

For completers c_j among n_j starters, each arm's completion probability was assigned the weak Jeffreys prior

$$p_j \mid c_j, n_j \sim \text{Beta}\left(c_j + \tfrac{1}{2},\, n_j - c_j + \tfrac{1}{2}\right). \tag{2}$$

Two million seeded draws were used to obtain the posterior distribution of $p_e - p_p$. Three summaries were retained: the mean difference, its equal-tailed 95% credible interval, and $\Pr(|p_e - p_p| < 0.10)$. The 10-point interval is an operational tolerance rather than a validated clinical threshold. It asks whether completion probabilities are close enough for service planning, not whether the therapies are clinically interchangeable. Bayesian estimation is used here because it expresses the uncertainty around near-identical observed proportions directly; it does not convert the preference cohort into a randomized comparison.

2.4.3. Persistence conditional on withdrawal

Sessions completed before withdrawal were analyzed only among participants who withdrew. Let $\bar{s}_e$ and $\bar{s}_p$ denote reported means and SE_e and SE_p their standard errors. The conditional difference and its standard error were

$$\Delta_s = \bar{s}_e - \bar{s}_p, \qquad SE(\Delta_s) = \sqrt{SE_e^2 + SE_p^2}. \tag{3}$$

A normal approximation produced a 95% interval and the probability that $\Delta_s > 0$. This estimate is not an overall treatment-retention effect. Conditioning on eventual withdrawal defines a selected subgroup, and the composition of that subgroup can differ by delivery mode. Its role is narrower: it describes how long non-completers remained engaged in each arm.

2.4.4. Normalized symptom change and bias boundary

Let z_{jt} be the mean PHQ-9 score at time t divided by the pretreatment arm mean. The arm contrast at time t is

$$\widehat{\theta}_t = z_{et} - z_{pt}. \tag{4}$$

Negative values indicate a lower remaining proportion of pretreatment severity in e-CBT. Because normalized means were reported to two decimal places, each value was also treated as an interval of ± 0.005 for rounding. This avoids false precision. A comparability region of $[-0.10, 0.10]$ was specified for the endpoint contrast. Equivalence procedures require the margin to be stated separately from the observed p value,[34,35] and equivalence testing is conceptually distinct from failure to reject a difference.[36] Here the region is used for sensitivity mapping because standard errors for the normalized endpoint were not available.

Residual preference bias was denoted by B on the same normalized scale. The bias-adjusted endpoint contrast was defined as

$$\theta^*_{12} = \widehat{\theta}_{12} - B. \tag{5}$$

The values of B satisfying $|\theta^*_{12}| < 0.10$ form the robustness interval. This quantity answers the paper's central question: it gives the direction and magnitude of residual distortion that the endpoint conclusion can tolerate. It does not assert that B is known or that all unmeasured confounding acts additively. Rather, it translates an untestable assumption into a visible boundary.

2.5. Incomplete follow-up, computation, and reporting safeguards

Incomplete outcome observation can bias longitudinal comparisons when the probability of missingness depends on unobserved outcomes or arm-specific experiences. General guidance recommends preventing missing observations, documenting reasons, and examining departures from assumptions rather than relying on complete cases alone.[37] Multiple imputation by chained equations can address some missing-data structures when participant-level predictors are available;[38] controlled imputation and weighted approaches can explore missing-not-at-random departures.[39] Those procedures were not possible from aggregate summaries. PATPA therefore keeps completion and conditional persistence visible, refrains from reconstructing unavailable patient trajectories, and interprets endpoint similarity in light of approximately 55% non-completion.

All computations were implemented in Python 3 using NumPy, SciPy, and Matplotlib. The random seed was 20260814. The analytic CSV, calculation script, numerical JSON output, and individual figure files accompany the manuscript. Statistical results are presented with intervals and probabilities, avoiding the common error of equating a large p value with evidence that two arms are the same. This interpretation follows the principle that scientific conclusions should not be reduced to a bright-line significance rule.[40]

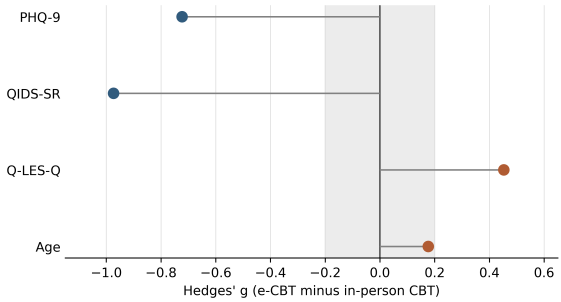

Figure 3. Pretreatment arm separation

The separation in Table 2 prevents conceptual substitution. Similar completion does not establish similar symptom improvement; a longer time to withdrawal does not establish greater endpoint effectiveness; and a small normalized endpoint contrast does not erase pretreatment selection. The five quantities must agree before a strong modality-neutral claim becomes persuasive.

3. RESULTS

3.1. Pretreatment selection pattern

The symptom arms were separated before treatment. For PHQ-9, Hedges' g for e-CBT minus in-person CBT was -0.723, indicating lower depressive severity in the e-CBT arm. The corresponding QIDS-SR contrast was larger, $g = -0.973$. Q-LES-Q pointed in the same substantive direction despite its opposite scoring orientation: $g = 0.453$ indicated better pretreatment quality of life among e-CBT participants. Age showed only a small positive contrast ($g = 0.177$), with e-CBT participants slightly older on average. The pattern is visualized in Figure 3 and summarized in Table 3.

The compact display in Figure 3 makes the selection structure visible. Both depressive symptom measures lie well outside the lightly shaded -0.20 to 0.20 zone, and Q-LES-Q also shows more than a small separation. Age lies within that zone. The coherence across PHQ-9, QIDS-SR, and Q-LES-Q suggests that delivery choice was related to clinical condition rather than to random fluctuation in one instrument. Participants entering in-person care had greater symptom burden and poorer quality of life.

The numerical contrasts in Table 3 are too large to regard the arms as exchangeable on clinical status. The result also clarifies why percentage or normalized change is attractive: it places each arm relative to its own starting level. That transformation improves interpretability but does not guarantee confounding control. Regression to the mean, different response potential, ceiling or floor behavior, and association between severity and withdrawal can all remain after normalization.

Table 2. PATPA components and interpretations

Component	Quantity	Interpretation
Pretreatment selection	Hedges' g	Direction and magnitude of arm separation before treatment
Completion	$p_e - p_p$	Difference in probability of reaching session 12
Conditional persistence	Δ_s	Session difference among eventual withdrawals
Symptom course	$z_{et} - z_{pt}$	Difference in remaining PHQ-9 proportion
Bias boundary	$\widehat{\theta}_{12} - B$	Residual distortion tolerated by the 10-point region

Table 3. Standardized pretreatment contrasts

Quantity	e-CBT mean	In-person mean	Hedges' g
PHQ-9	16.08	19.91	-0.723
QIDS-SR	14.61	19.29	-0.973
Q-LES-Q	36.31	32.09	0.453
Age, years	38.60	36.36	0.177

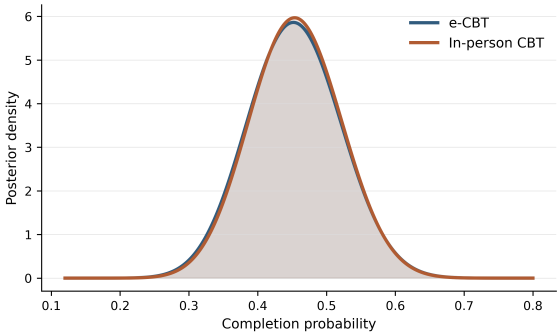

Figure 4. Completion probability distributions

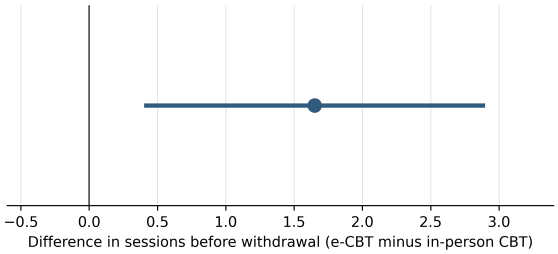

Figure 5. Persistence before withdrawal

3.2. Completion uncertainty

Observed completion proportions were 24/53 (45.28%) for e-CBT and 25/55 (45.45%) for in-person CBT, a raw difference of -0.17 percentage points. With Jeffreys priors, the posterior mean completion probabilities were 0.454 for e-CBT and 0.455 for in-person CBT. The posterior mean risk difference was -0.0016, and the equal-tailed 95% credible interval ranged from -0.185 to 0.182. The probability that e-CBT completion exceeded in-person completion was 0.493, almost exactly balanced. The probability that the difference was inside the prespecified ± 0.10 interval was 0.710.

The nearly coincident curves in Figure 4 explain both the similarity and its limitation. Their centers are almost identical, so there is no directional completion signal. Their width remains substantial because each arm contains only about 50 starters and fewer than half completed. A statement that observed completion was alike is accurate; a statement that clinically important completion differences were excluded is not. Nearly 29% of posterior draws lay outside the 10-point interval, and the 95% credible interval allowed differences approaching 18 percentage points in either direction.

This distinction matters for service decisions. If two delivery modes truly differ by 12 or 15 completion points, that difference could affect staffing, follow-up procedures, and the expected number of completed treatment courses. The observed counts do not resolve differences of that size. PATPA therefore classifies completion evidence as centered on parity but imprecise, rather than as proof of equal adherence.

3.3. Persistence among participants who withdrew

The conditional analysis produced a clearer directional result. Participants who withdrew from e-CBT completed a reported mean of 4.65 sessions, compared with 3.00 sessions among in-person withdrawals. Equation 3 gave $\Delta_s = 1.65$ sessions and $SE(\Delta_s) = 0.637$. The 95% interval was 0.40 to 2.90 sessions, and the normal-approximation probability that the difference exceeded zero was 0.995. Figure 5 shows the estimate and interval.

The interval in Figure 5 lies entirely above zero, indicating that eventual e-CBT withdrawals remained for longer on average. The estimate does not contradict the almost identical overall completion proportions. Completion

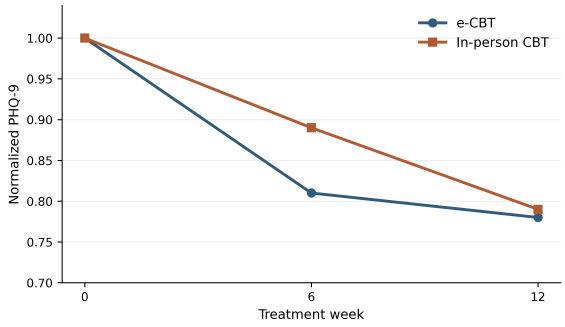

Figure 6. Normalized PHQ-9 course

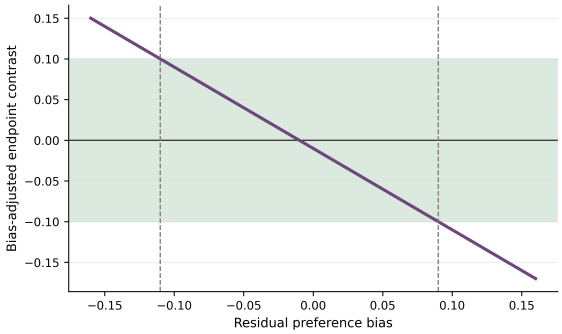

Figure 7. Preference-bias boundary

records whether session 12 was reached; persistence records how much treatment a non-completer received before leaving. These quantities can move differently. A modality can have the same number of completers but delay discontinuation among those who do not complete.

The timing counts support that interpretation. Eighteen of 29 e-CBT withdrawals occurred during weeks 1–6 and 11 during weeks 7–12. In-person care had 24 of 30 withdrawals during weeks 1–6 and only six thereafter. Thus, 37.9% of e-CBT withdrawals occurred in the later half compared with 20.0% of in-person withdrawals. The contrast may reflect easier continued access, asynchronous pacing, different expectations, or distinct reasons for stopping. Because withdrawal reasons and participant-level characteristics were unavailable, these explanations remain hypotheses rather than identified mechanisms.

3.4. Normalized PHQ-9 course

At week 6, normalized PHQ-9 was 0.81 for e-CBT and 0.89 for in-person CBT, yielding $\widehat{\theta}_6 = -0.08$. Relative to pretreatment means, this corresponds to 19% and 11% reductions. At week 12, the respective normalized values were 0.78 and 0.79, so $\widehat{\theta}_{12} = -0.01$, corresponding to 22% and 21% reductions. Figure 6 displays the arm-level course.

The trajectories in Figure 6 converge by week 12 after differing at week 6. E-CBT shows a larger early proportional decrease followed by a smaller change from week 6 to week 12. In-person CBT shows a smaller early proportional decrease followed by continued improvement during the second half. This temporal pattern is clinically more informative than an endpoint-only statement because it raises questions about pace, support intensity, and who remains under observation at each time.

Rounding-aware calculation strengthens only the descriptive endpoint observation. A reported value of 0.78 represents values from 0.775 through values just below 0.785, while 0.79 represents 0.785 through values just below 0.795. The endpoint contrast therefore lies between approximately -0.020 and 0.000 before accounting for sampling error or missing outcomes. This complete rounding interval remains inside $[-0.10, 0.10]$. The unavailable standard error of the normalized contrast prevents a formal interval-based equivalence test.

3.5. Residual-bias boundary

With the central endpoint contrast of -0.01, Equation 5 becomes $\theta^*_{12} = -0.01 - B$. Solving $|-0.01 - B| < 0.10$ gives

$$-0.11 < B < 0.09. \tag{6}$$

The conclusion of endpoint comparability therefore tolerates residual preference distortion up to 11 percentage points in one direction and 9 percentage points in the other. Figure 7 shows the complete boundary.

The green band in Figure 7 marks normalized contrasts within ± 0.10. The purple line leaves the band when residual bias passes either dashed boundary. The result is asymmetric because the observed contrast slightly favors e-CBT. A positive B shifts the adjusted contrast downward and reaches the lower edge at 0.09; a negative B shifts it upward and reaches the upper edge at -0.11. The graph turns the qualitative concern about treatment selection into a quantitative statement: preference-related distortion on the order of one tenth of the pretreatment PHQ-9 level is sufficient to change the interpretation.

The collection of estimates in Table 4 does not support a single unqualified verdict. Endpoint proportional change is close, and the full rounding interval lies inside the chosen region. Completion is centered on no difference but remains uncertain. Conditional persistence favors e-CBT. Most importantly, substantial pretreatment symptom separation shows that the endpoint comparison is vulnerable to preference-related distortion. The appropriate result is therefore conditional: descriptive endpoint comparability holds under residual bias smaller than the calculated boundary, not under every plausible selection process.

4. DISCUSSION

4.1. Direct answer to the research question

The study asked how much residual preference-related distortion would be sufficient to reverse a conclusion of comparable proportional PHQ-9 improvement at 12 weeks, and whether completion and conditional persistence supported the same interpretation. The central

Table 4. Principal PATPA estimates

Estimate	Value	Interval or probability
Completion difference, e-CBT minus in-person	-0.002	95% CrI: -0.185 to 0.182
Completion difference within ± 0.10	—	Probability: 0.710
e-CBT completion greater than in-person	—	Probability: 0.493
Sessions before withdrawal, arm difference	1.65	95% interval: 0.40 to 2.90
Positive session difference	—	Probability: 0.995
Normalized PHQ-9 difference, week 6	-0.08	Descriptive
Normalized PHQ-9 difference, week 12	-0.01	Rounding interval: -0.02 to 0.00
Residual bias preserving endpoint region	—	$-0.11 < B < 0.09$

endpoint contrast was -0.01 of the pretreatment PHQ-9 level. Under the stated ± 0.10 region, comparability remained intact only when residual bias lay between -0.11 and 0.09. Thus, the conclusion can tolerate small distortion, but preference-related distortion of roughly one tenth of the pretreatment mean is sufficient to overturn it. The answer is neither that delivery modes were proven equivalent nor that the endpoint comparison is uninterpretable. It is that endpoint similarity has a measurable and moderately narrow robustness boundary.

Completion and persistence did not produce the same evidential pattern. Completion estimates were centered almost exactly on parity, yet their uncertainty allowed meaningful differences in either direction. Conditional persistence showed a more definite e-CBT advantage of 1.65 sessions among eventual withdrawals. These results answer different parts of the question. The endpoint proportional symptom course was close at week 12; full-course completion remained uncertain despite near-identical observed proportions; and participants who did not complete stayed in e-CBT longer on average. A defensible conclusion must retain all three statements.

4.2. Treatment preference as clinical information

The strongest empirical pattern appeared before treatment. Standardized differences of -0.72 for PHQ-9 and -0.97 for QIDS-SR indicate that participants choosing in-person CBT entered with substantially greater depressive burden. Q-LES-Q reinforced this interpretation because quality of life was poorer in the in-person arm. Age showed little separation, so the principal observed selection pattern was clinical rather than demographic. This result is consistent with evidence that people hold distinct preferences for digital and in-person psychotherapy[16] and that shared decisions reflect perceived need, acceptability, and expected support.[17]

The preference pattern should not be treated solely as bias to be eliminated. It may contain useful triage information. A patient experiencing more severe symptoms or lower life satisfaction may prefer synchronous contact, immediate clarification, and the ritual of attending a clinical setting. Another patient may value privacy, flexible timing, and the ability to reflect before written disclosure. Home delivery can reduce transport and mobility barriers,[18] while digital participation can impose its own requirements for reliable connectivity, literacy, device access, and self-initiation. Selection therefore expresses both clinical need and practical capacity.

At the same time, preference prevents a direct causal reading. If severity influences both delivery choice and subsequent change, the arm contrast may reflect the natural association between initial severity and change. More severe participants may have greater numerical room to improve, but they may also experience more persistent symptoms, greater functional impairment, or higher discontinuation risk. Normalizing by each arm's pretreatment mean addresses scale difference but not these competing processes. Participant-level modeling would be required to examine nonlinear severity effects, interactions, medication, comorbidity, referral route, therapist assignment, and socioeconomic constraints.

PATPA responds by changing the question from "Which mode caused better outcomes?" to "How stable is the observed comparison to a stated amount of residual preference distortion?" That change is methodologically important. It makes the central assumption visible, leaves the non-randomized design intact, and avoids the false precision of reconstructing individual observations from published means. Sensitivity analysis has long been recommended when hidden bias cannot be excluded.[27,28] The boundary here is simple enough to communicate to clinicians: a nine-point positive bias or an eleven-point negative bias on the normalized PHQ-9 scale reaches the edge of the chosen region.

4.3. What completion and withdrawal timing reveal

The observed completion proportions differ by less than one quarter of a percentage point. In routine prose, that result is easily described as "the same completion rate." The posterior calculation shows what this shorthand omits. With 24 and 25 completers, the 95% credible interval for the arm difference extends to approximately ± 18 percentage points. A future cohort with the same underlying rates could readily produce a visibly different arm

contrast. The current counts therefore support absence of a directional signal, not exclusion of an operationally important difference.

The 0.710 probability of lying inside the 10-point completion region offers a more graded interpretation than a dichotomous test. It indicates that parity is more plausible than not under the stated prior and model, but the evidence does not reach a level that would justify certainty. A decision maker who regards a 10-point loss in completion as consequential should not ignore the remaining 0.290 probability outside that region. Larger samples or pooled service data would be needed to narrow this uncertainty.

Conditional persistence adds information that a completion proportion cannot contain. E-CBT withdrawals completed about one and two-thirds more sessions, and late withdrawal was proportionally more common in e-CBT. One plausible explanation is that asynchronous access reduces early logistical failure: a patient can continue without travel, schedule coordination, or a fixed appointment. Another is that written therapist feedback sustains participation even when a full 12-session course is not completed. Therapist support is known to affect online adherence,[23] and alliance can predict change in digital and blended treatment.[20,21] Yet the conditional estimate could also reflect subgroup composition. If patients with greater initial severity entered in-person treatment and were more likely to stop early, withdrawal timing may partly represent selection rather than a property of the setting.

The distinction between completion and persistence also has clinical consequences. A patient completing five sessions before stopping has received more therapeutic exposure than a patient stopping after one, but neither is a full completer. Whether that additional exposure yields lasting benefit cannot be determined from session counts alone. Some patients may stop because they improved, others because they deteriorated, and others because the treatment was inaccessible or unacceptable. Future data collection should record structured reasons for discontinuation, symptom status at withdrawal, and whether alternative care began. Such data would allow disengagement to be separated from planned early termination and external interruption.

4.4. Temporal course and endpoint interpretation

The normalized PHQ-9 course showed a larger early proportional decrease in e-CBT and continued later improvement in in-person CBT. By week 12, the arms converged. This shape could arise from true differences in response timing. An asynchronous program may permit rapid engagement with structured material, whereas face-to-face work may develop more gradually through therapeutic exchange. It could also arise from changing composition as participants withdraw. With more early in-person withdrawals, the week-6 and week-12 in-person means may represent a selected subset. Without individual longitudinal records, these explanations cannot be separated.

The week-12 contrast remained within the chosen region even after allowing for reporting to two decimal places. That is useful descriptive evidence, but it is not a formal equivalence result. Equivalence claims require a clinically justified margin, sufficient precision, and an analysis aligned with the estimand and missing-data assumptions.[34,35] The 10-point normalized region in this paper is a transparent operational choice used to reveal a tipping point. It should not be interpreted as an established minimal clinically important difference for PHQ-9.

The original symptom scales remain important. A normalized value of 0.78 does not identify how many participants moved below a clinical threshold, achieved a response criterion, or remitted. Mean proportional change can coexist with large heterogeneity. PHQ-9 is suitable for monitoring change,[4] and QIDS-SR provides complementary symptom information;[5] however, clinical decisions often require responder distributions, suicidality monitoring, functioning, and patient-defined goals. Q-LES-Q appropriately extends the assessment to life satisfaction,[6,7] yet only pretreatment arm summaries were available for the current sensitivity calculations.

The absence of normalized QIDS-SR and Q-LES-Q endpoints from the analytic file prevents a cross-outcome robustness comparison. If proportional convergence occurred on all three measures, confidence in a modality-neutral pattern would increase. If symptom measures converged but quality of life did not, the delivery modes might have different functional consequences. Reporting arm-specific means, standard deviations, sample sizes, and correlations at each visit would permit multivariate analysis without forcing researchers to digitize graphs or infer unavailable covariance.

4.5. Methodological contribution of PATPA

PATPA is designed for a common but under-addressed reporting problem: a preference cohort offers arm-level repeated outcomes, high attrition, and a non-significant treatment-by-time interaction, yet lacks the individual records required for complete causal adjustment. The method contributes by arranging the available evidence in a fixed logical order. Pretreatment standardized differences diagnose selection. Beta–binomial estimation quantifies completion uncertainty. Conditional session analysis describes persistence without relabeling it as completion. Normalized trajectories show the time pattern. The bias boundary states the residual distortion that changes the endpoint interpretation.

Each component can be audited. The Hedges correction follows established effect-size theory.[31,32] The completion distribution follows directly from counts and a weak reference prior. The persistence interval follows from the reported means and standard errors. The rounding interval follows from decimal precision. The bias boundary is

an algebraic consequence of the observed contrast and chosen region. No latent patient records are created, and no unavailable correlation is silently assumed.

This transparency is especially valuable because statistical non-significance is frequently overread. A non-significant arm-by-time interaction tests compatibility with zero under a particular model; it does not establish that every clinically important difference is absent. Equivalence testing reverses the logic by asking whether an interval lies entirely inside a margin.[36] PATPA cannot conduct that full test without an endpoint standard error, but it adopts the essential discipline of declaring a region and examining its boundary. This practice is aligned with recommendations to move beyond bright-line interpretations of p values.[40]

The method is also intentionally modest. A bias parameter B on the normalized scale is easier to communicate than an elaborate latent-variable model, but it compresses several mechanisms into one additive quantity. Preference may alter outcomes through severity, expectation, alliance, digital competence, travel, therapist matching, or time constraints. These processes need not be additive or constant over time. PATPA should therefore be understood as a diagnostic sensitivity analysis, not as a substitute for design-stage randomization, rich covariate measurement, or participant-level causal modeling.

4.6. Clinical and service implications

The results favor choice with safeguards rather than a universal delivery rule. Both delivery environments can support structured CBT, and comparative literature has often found similar average benefit.[12,13] The current analysis shows that patients choosing in-person care may present with greater symptom burden. Services should therefore avoid interpreting digital selection as evidence of lesser need or in-person selection as resistance to technology. Initial assessment should examine severity, safety, functional impairment, digital access, communication preference, and capacity for self-directed work.

The early withdrawal pattern suggests that the first six weeks deserve targeted monitoring, particularly in scheduled in-person care. Practical interventions could include rapid rescheduling after a missed appointment, transport support, a brief remote check-in when attendance becomes difficult, and explicit discussion of expectations during the first session. For e-CBT, later persistence among withdrawals suggests that delayed disengagement is also important. Automated reminders should not replace clinical review when assignments cease; a therapist should determine whether the patient is improving, overwhelmed, dissatisfied, or at risk.

Therapeutic alliance remains relevant across settings. Evidence from blended treatment indicates that alliance can predict depression change,[19,20] and alliance has also predicted longer-term outcome during guided internet care.[21] Digital services should therefore measure perceived collaboration, not assume that written feedback automatically produces it. In-person services, meanwhile, should not assume that physical co-presence eliminates access burdens or ensures engagement. Delivery quality depends on responsiveness, clarity, continuity, and fit in both modes.

Service evaluation should report more than the proportion who completed. At minimum, reports should include time to withdrawal, reasons for withdrawal, symptom status at the final contact, therapist contact time, and subsequent care. A survival-type analysis would use the timing of discontinuation more efficiently than a binary endpoint. Joint longitudinal and dropout modeling could examine whether symptom course and withdrawal are related, provided participant-level data and assumptions are available. These additions would determine whether e-CBT's longer conditional persistence represents useful therapeutic exposure or merely delayed disengagement.

4.7. Strengths and limitations

Several strengths improve interpretability. The research question is explicitly about robustness rather than superiority. Pretreatment clinical differences are quantified on a common scale. Completion uncertainty is estimated rather than dismissed because proportions are close. Withdrawal timing is separated from full completion. Decimal rounding is incorporated into the endpoint calculation. The complete analysis is reproducible from a compact CSV and script, and the staged context images are clearly separated from empirical evidence.

The principal limitation is the use of aggregate numerical summaries. Individual trajectories, covariances, medication status, comorbidities, therapist effects, referral pathways, and reasons for treatment choice were unavailable. These omissions prevent propensity weighting, outcome regression, doubly robust adjustment, subgroup analysis, and joint modeling. The standardized pretreatment contrasts are descriptive and do not provide an adjusted treatment effect. The normalized PHQ-9 contrast reduces scale dependence but can still be affected by regression to the mean and selection.

Attrition was substantial: only 49 of 108 starters completed all sessions. Missing outcomes are therefore a central feature, not a minor inconvenience. The analysis could not perform multiple imputation or controlled departures because the necessary individual predictors and outcome distributions were unavailable. General missing-data guidance emphasizes prevention, reason collection, and sensitivity analysis;[37] imputation methods require richer records than arm means.[38,39] The completion posterior and bias boundary reveal uncertainty, but they do not recover unobserved outcomes.

The reported materials also contained an inconsistency in sessions completed before withdrawal. The narrative specified 4.65 sessions for e-CBT and 3.00 for in-person CBT, while a tabular row presented reversed rounded val-

ues. The narrative values were used because they agree with the stated direction, standard errors, and interpretation of later withdrawal in e-CBT. This resolution is documented in the accompanying correction audit. A further limitation is that normalized endpoint standard errors were not reported, precluding a formal equivalence interval. The bias boundary therefore concerns residual selection distortion around the reported central contrast and does not include full sampling uncertainty.

The 10-point regions for completion and normalized PHQ-9 are transparent analytic choices rather than established clinical thresholds. Other decision makers may choose narrower or wider regions. This dependence is not a defect when made explicit: the script can recalculate the boundary for another value. For a margin m, the allowable bias interval is $\widehat{\theta}_{12} - m < B < \widehat{\theta}_{12} + m$ after rearranging Equation 5. What matters is that the margin is justified before interpreting the result.

Generalizability is limited to care resembling the cohort: adults with major depressive disorder who could access one of two therapist-supported CBT formats and were willing to select between them. Results may differ where digital support is synchronous, therapist contact is minimal, treatment length varies, devices are supplied, broadband access is limited, or risk and comorbidity are more complex. Comparative evidence supports internet CBT across many settings,[9,10] but intervention components and participant characteristics remain consequential.[11] The present boundary should therefore be recalculated rather than transferred unchanged to another service.

5. CONCLUSION

The paper's research question can be answered quantitatively. The observed week-12 proportional PHQ-9 contrast was -0.01, and its rounding interval remained within the prespecified ± 0.10 region. That interpretation persisted only while residual preference bias stayed between -0.11 and 0.09 of the pretreatment level. Preference-related distortion of about one tenth was therefore sufficient to reverse endpoint comparability. Completion proportions were nearly identical but imprecise, with only a 0.710 posterior probability of lying within 10 percentage points. Among eventual withdrawals, e-CBT participants remained engaged for an estimated 1.65 additional sessions.

These results support a conditional, not absolute, conclusion. Therapist-supported e-CBT and in-person CBT showed closely aligned proportional PHQ-9 values at week 12, but strong pretreatment symptom separation and extensive attrition limit causal interpretation. Delivery-mode comparisons should quantify pretreatment selection, show uncertainty around completion, distinguish persistence from completion, and report a bias boundary. Doing so replaces an unsupported claim of equivalence with an auditable statement about how much hidden distortion the observed conclusion can withstand.

BIBLIOGRAFIA / REFERENCES

1. Parikh SV, Quilty LC, Ravitz P, Rosenbluth M, Pavlova B, Grigoriadis S, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: Section 2. Psychological treatments. *Can J Psychiatry*. 2016;61:524–539. doi:10.1177/0706743716659418.
2. Hawley LL, Padesky CA, Hollon SD, Mancuso E, Laposa JM, Brozina K, et al. Cognitive-behavioral therapy for depression using Mind Over Mood: CBT skill use and differential symptom alleviation. *Behav Ther*. 2017;48:29–44. doi:10.1016/j.beth.2016.09.003.
3. Zhang Z, Zhang L, Zhang G, Jin J, Zheng Z. The effect of CBT and its modifications for relapse prevention in major depressive disorder: a systematic review and meta-analysis. *BMC Psychiatry*. 2018;18:50. doi:10.1186/s12888-018-1610-5.
4. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. *Med Care*. 2004;42:1194–1201. doi:10.1097/00005650-200412000-00006.
5. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating and self-report: a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*. 2003;54:573–583. doi:10.1016/S0006-3223(02)01866-8.
6. Endicott J, Nee J, Harrison W, Blumenthal R. Quality of Life Enjoyment and Satisfaction Questionnaire: a new measure. *Psychopharmacol Bull*. 1993;29:321–326.
7. Riendeau RP, Sullivan JL, Meterko M, Stolzmann K, Williamson AK, Miller CJ, et al. Factor structure of the Q-LES-Q Short Form in an enrolled mental health clinic population. *Qual Life Res*. 2018;27:2953–2964. doi:10.1007/s11136-018-1963-8.
8. McDonald A, Eccles JA, Fallahkhair S, Critchley HD. Online psychotherapy: trailblazing digital healthcare. *BJPsych Bull*. 2020;44:60–66. doi:10.1192/bjb.2019.66.
9. Etzelmueller A, Vis C, Karyotaki E, Baumeister H, Titov N, Berking M, et al. Effects of internet-based cognitive behavioral therapy in routine care for adults in treatment for depression and anxiety: systematic review and meta-analysis. *J Med Internet Res*. 2020;22:e18100. doi:10.2196/18100.
10. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*. 2021;78:361–371. doi:10.1001/jamapsychiatry.2020.4364.
11. Furukawa TA, Suganuma A, Ostinelli EG, Andersson G, Beevers CG, Shumake J, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic re-

view and component network meta-analysis using individual participant data. *Lancet Psychiatry*. 2021;8:500–511. doi:10.1016/S2215-0366(21)00077-8.
12. Andersson G, Topooco N, Havik O, Nordgreen T. Internet-supported versus face-to-face cognitive behavior therapy for depression. *Expert Rev Neurother*. 2016;16:55–60. doi:10.1586/14737175.2015.1125783.
13. Cuijpers P, Donker T, van Straten A, Li J, Andersson G. Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychol Med*. 2010;40:1943–1957. doi:10.1017/S0033291710000772.
14. Wright JH, Owen JJ, Richards D, Eells TD, Richardson T, Brown GK, et al. Computer-assisted cognitive-behavior therapy for depression: a systematic review and meta-analysis. *J Clin Psychiatry*. 2019;80:18r12188. doi:10.4088/JCP.18r12188.
15. Thase ME, Wright JH, Eells TD, Barrett MS, Wisniewski SR, Balasubramani GK, et al. Improving the efficiency of psychotherapy for depression: computer-assisted versus standard CBT. *Am J Psychiatry*. 2018;175:242–250. doi:10.1176/appi.ajp.2017.17010089.
16. Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the United States. *NPJ Digit Med*. 2019;2:6. doi:10.1038/s41746-019-0077-1.
17. Rodenburg-Vandenbussche S, Carlier I, van Vliet I, van Hemert A, Stiggelbout A, Zitman F. Patients' and clinicians' perspectives on shared decision-making regarding treatment decisions for depression, anxiety disorders, and obsessive-compulsive disorder in specialized psychiatric care. *J Eval Clin Pract*. 2020;26:645–658. doi:10.1111/jep.13285.
18. Borson S, Korpak A, Carbajal-Madrid P, Likar D, Brown GA, Batra R. Reducing barriers to mental health care: bringing evidence-based psychotherapy home. *J Am Geriatr Soc*. 2019;67:2174–2179. doi:10.1111/jgs.16088.
19. Kooistra LC, Ruwaard J, Wiersma JE, van Oppen P, Riper H. Working alliance in blended versus face-to-face cognitive behavioral treatment for patients with depression in specialized mental health care. *J Clin Med*. 2020;9:347. doi:10.3390/jcm9020347.
20. Vernmark K, Hesser H, Topooco N, Berger T, Riper H, Luuk L, et al. Working alliance as a predictor of change in depression during blended cognitive behaviour therapy. *Cogn Behav Ther*. 2019;48:285–299. doi:10.1080/16506073.2018.1533577.
21. Penedo JMG, Babl AM, Grosse Holtforth M, Hohagen F, Krieger T, Lutz W, et al. The association of therapeutic alliance with long-term outcome in a guided internet intervention for depression: secondary analysis from a randomized controlled trial. *J Med Internet Res*. 2020;22:e15824. doi:10.2196/15824.
22. Stiles-Shields C, Kwasny MJ, Cai X, Mohr DC. Therapeutic alliance in face-to-face and telephone-administered cognitive behavioral therapy. *J Consult Clin Psychol*. 2014;82:349–354. doi:10.1037/a0035554.
23. Alfonsson S, Olsson E, Linderman S, Winnerhed S, Hursti T. Is online treatment adherence affected by presentation and therapist support? A randomized controlled trial. *Comput Hum Behav*. 2016;60:550–558. doi:10.1016/j.chb.2016.01.035.
24. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688–701. doi:10.1037/h0037350.
25. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25:1–21. doi:10.1214/09-STS313.
26. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399–424. doi:10.1080/00273171.2011.568786.
27. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*. 1987;74:13–26. doi:10.1093/biomet/74.1.13.
28. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol*. 1996;25:1107–1116. doi:10.1093/ije/25.6.1107.
29. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017;167:268–274. doi:10.7326/M16-2607.
30. Alavi N, Moghimi E, Stephenson C, Gutierrez G, Jagayat J, Kumar A, et al. Comparison of online and in-person cognitive behavioral therapy in individuals diagnosed with major depressive disorder: a non-randomized controlled trial. *Front Psychiatry*. 2023;14:1113956. doi:10.3389/fpsyt.2023.1113956.
31. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. *Organ Res Methods*. 2008;11:364–386. doi:10.1177/1094428106291059.
32. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat*. 1981;6:107–128. doi:10.3102/10769986006002107.
33. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol*. 2013;4:863. doi:10.3389/fpsyg.2013.00863.
34. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308:2594–2604. doi:10.1001/jama.2012.87802.
35. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials*. 2011;12:106. doi:10.1186/1745-6215-12-106.
36. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Personal Sci*. 2017;8:355–362. doi:10.1177/1948550617697177.
37. Little RJA, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treat-

ment of missing data in clinical trials. *N Engl J Med*. 2012;367:1355–1360. doi:10.1056/NEJMsr1203730.
38. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30:377–399. doi:10.1002/sim.4067.
39. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res*. 2007;16:259–275. doi:10.1177/0962280206075303.
40. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129–133. doi:10.1080/00031305.2016.1154108.