PSILOGOS

Research Article

open-access

When Average Benefit Is Not the Decision: Predictive Certainty and Delivery Concentration in Online Psychotherapy for Pandemic Distress

Jorge Chuaqui Kettlun

ABSTRACT

Random-effects summaries can establish an average association while leaving clinicians uncertain about the effect expected in another service or population. Online psychotherapy for coronavirus disease 2019 (COVID-19)-related distress is a pertinent example because trials differed markedly in delivery channel, therapist involvement, treatment dose, population, and measured outcome. To determine whether pooled improvements in depression, anxiety, stress, and sleep remain persuasive when between-study dispersion is translated into probabilities for a comparable new trial, and to identify which delivery contrasts retain decision value under propagated uncertainty. The analytic collection comprised 13 randomized trials (1,897 participants), four outcome-level random-effects summaries, and prespecified delivery contrasts. Standard errors were recovered from published 95% confidence intervals. For each outcome, a t-based 95% prediction interval and predictive probabilities of any benefit, a benefit of at least 0.20 standardized mean difference, a benefit of at least 0.50, and harm were calculated from the reported mean effect and between-study variance. A 200,000-draw Monte Carlo check evaluated numerical agreement. Normalized entropy described design concentration, and uncertainty in two delivery contrasts was propagated from subgroup confidence intervals. Mean effects favored online psychotherapy for depression ($\mathrm{SMD} = -0.45$), anxiety (-0.67), stress (-0.73), and sleep (-0.53), but all 95% prediction intervals crossed no effect: -1.22 to 0.32, -1.72 to 0.38, -2.06 to 0.60, and -3.74 to 2.68, respectively. Probabilities that a comparable new trial would achieve at least a 0.20-SMD benefit were 0.779, 0.856, 0.826, and 0.671. The corresponding harm probabilities were 0.083, 0.065, 0.098, and 0.239. Therapist guidance had a 0.991 probability of outperforming self-help for anxiety, whereas short daily treatment had a 0.997 probability of outperforming seven- to eight-week weekly treatment for depression; the latter contrast depended on a single short-daily comparison. The evidence supports probable benefit for emotional distress, yet does not justify assuming benefit in every comparable implementation. Anxiety has the strongest combination of predictive benefit and low harm probability. Sleep remains too uncertain for a directional recommendation. Guidance for anxiety is more transportable than the apparent depression advantage of an intensive short dose, which requires direct replication.

Keywords: online psychotherapy; telepsychology; COVID-19; prediction interval; heterogeneity; treatment guidance; decision uncertainty

Escuela de Sociología, Universidad de Valparaíso, Valparaíso, Chile
jorge.chuaqui@uv.cl

– **Published:** 5 November 2024

1. INTRODUCTION

The rapid transfer of psychological care from consulting rooms to telephones, web portals, and videoconferencing systems was one of the most consequential service changes during the COVID-19 pandemic. This transition answered an immediate access problem: physical distancing reduced conventional contact at the same time that bereavement, fear of infection, occupational disruption, isolation, and financial uncertainty increased psychological strain. Quarantine has been consistently associated with anxiety, depressive symptoms, anger, and post-traumatic stress, with longer restriction, inadequate information, and financial loss contributing to worse experience.[7] Reviews conducted during the early pandemic also documented elevated distress across infected patients, health-care workers, and community populations.[32,40] The service response was therefore not merely a change of medium. It was an attempt to preserve clinical continuity during a population-level shock while conventional capacity was constrained.[18,41]

Remote psychological care did not begin with the pandemic. Videoconferencing psychotherapy had already been used across clinical populations, treatment formats, and health systems, with generally acceptable feasibility and satisfaction.[3] Later reviews found meaningful improvement in depression and anxiety with synchronous telepsychology, although study quality and clinical diversity limited broad equivalence claims.[5,6,39] Internet-delivered psychological treatment likewise developed from simple self-help pages into structured programs with therapist messaging, automated exercises, symptom monitoring, and stepped-care pathways.[1,2] Direct comparisons suggest that internet cognitive behavioral therapy can approach face-to-face outcomes for several psychiatric and somatic conditions;[8] nevertheless, effectiveness depends on who enters treatment, how support is delivered, how adherence is maintained, and which clinical endpoint is prioritized.

Delivery distinctions matter because the phrase "online psychotherapy" combines interventions with different causal ingredients. Synchronous sessions can preserve conversational responsiveness and allow risk assessment, whereas asynchronous programs can increase reach and scheduling flexibility. Therapist contact may enhance accountability, clarify exercises, and respond to deterioration, but it also constrains scale. Self-guided programs reduce professional time and may be sufficient for people with mild symptoms, yet adherence can decline when expectations, prompts, and human support are weak. A systematic review of guidance in internet interventions found that professional support generally improved efficacy, although the magnitude depended on the comparison and study design.[4] The supportive-accountability model explains this pattern through a credible, benevolent coach who establishes expectations and monitors progress.[24] Digital design itself also influences engagement: persuasive elements, tailoring, reminders, and dialogue support are associated with adherence, although their effects are not reducible to interface features alone.[23]

The relevant clinical question is consequently more demanding than whether the pooled mean differs from zero. A statistically precise average can coexist with large between-study dispersion. Under a random-effects model, the confidence interval around the mean estimates uncertainty about the average effect in the sampled distribution; it does not delimit the true effect in another comparable setting. A prediction interval addresses the latter quantity by combining uncertainty in the mean with estimated between-study variance.[20] This distinction becomes decisive when I^2 is high, trial numbers are small, interventions differ in dose and channel, and most controls are inactive. In such conditions, the average may be beneficial while a plausible future effect includes negligible benefit or harm. Presenting only the pooled mean compresses clinically relevant variation into a single number.

Three further problems complicate interpretation. First, outcomes measured in the same trials are not independent. Depression, anxiety, stress, and sleep share symptoms and determinants, yet each has a different number of contributing comparisons. Multivariate synthesis can use cross-outcome covariance when the required matrices are available;[28,29] absent those matrices, claims of joint precision would be unwarranted. Second, subgroup contrasts can appear compelling even when one level contains very few studies. A large difference based on one comparison is mathematically estimable but weakly transportable. Small-sample random-effects work emphasizes cautious uncertainty estimation and the limitations of asymptotic intervals.[19,31] Third, significance is threshold-dependent. A clinician may regard any improvement as relevant, whereas a commissioner may require at least a small standardized effect to justify implementation cost, and a patient may prioritize the probability of avoiding deterioration. A continuous threshold analysis displays these choices rather than fixing a single dichotomy.

The present research therefore asks: when reported between-study variance is carried into the expected effect of a comparable new trial, how certain is benefit for each psychological outcome, and which observed delivery contrasts remain persuasive? The analysis retains the published random-effects means and heterogeneity estimates, then adds prediction intervals, probability-of-benefit curves, Monte Carlo verification, and design-concentration summaries. The purpose is not to substitute a new efficacy estimate for the published mean. It is to translate the same aggregate evidence into quantities aligned with a future implementation decision. This research question differs from a conventional efficacy review because the primary endpoint is predictive certainty rather than statistical significance of the average.

2. METHODS

This study used an aggregate evidence-synthesis design. The analytic collection consisted of 13 randomized controlled trials conducted in Canada, China, Germany, Iran, Italy, Sweden, Switzerland, Tunisia, and the United States. Together they randomized 1,897 adults. The study-level table recorded country, population, randomized group sizes, female percentage, mean age, intervention family, therapist involvement, delivery channel, treatment frequency and duration, control condition, and measured outcomes. Four outcome records contained the number of comparisons, pooled SMD, 95% confidence interval, between-study variance τ^2, Cochran's Q, and I^2. Eighteen subgroup records represented two intervention types, two guidance conditions, and two dose patterns across depression, anxiety, and stress. These values were transcribed from Tables 2–4 of Chi et al.;[9] that article is cited once to identify provenance, while the enclosed CSV files provide the values used by every calculation.

No participant-level record was created, reconstructed, or imputed. Post-treatment SMDs were treated as oriented so that negative values favored online psychotherapy. The analytic unit for predictive calculations was an outcome-level pooled effect. Study descriptors were used only for composition and coverage analyses. Four trials had follow-up assessments, but the aggregate tables did not supply pooled follow-up effects; long-term benefit was therefore not estimated. All controls were inactive, comprising waiting-list or routine-care conditions. The resulting estimand is the difference from an inactive comparator at the post-treatment assessment, not noninferiority to face-to-face care or superiority over an attention-matched digital control.

Reporting was organized with the transparency principles of PRISMA 2020 where they applied to a synthesis of an already delimited trial collection.[26] Risk-of-bias judgments were not recalculated because trial reports and domain-level supporting material were not all available in the analytic collection. This limitation prevents an independent RoB 2 assessment[36] and is addressed in the certainty interpretation rather than represented as a new quality score. The computational work used Python 3 with NumPy, pandas, SciPy, and Matplotlib; all tables and figures are regenerated by a single script with a fixed random seed.

2.1. Descriptive composition and design concentration

Trial size was calculated as the sum of intervention and control participants. Intervention duration was expressed in days, using one day for a single session, five days for a five-day course, 14 days for two weeks, 21 days for three weeks, 28 days for four weeks, 49 days for seven weeks, and 56 days for eight weeks. Outcome coverage was coded as a binary study-by-domain matrix. Therapist guidance and self-help were retained as reported categories. Delivery comprised website, online platform, mobile application, or hospital call system.

Normalized Shannon entropy described how evenly studies were distributed across the observed levels of each design dimension. For dimension j with m_j observed levels and level proportions p_{jl}, entropy was

$$H_j = -\frac{\sum_{l=1}^{m_j} p_{jl} \log(p_{jl})}{\log(m_j)}. \tag{1}$$

The quantity ranges from zero, when all studies occupy one level, to one, when studies are evenly distributed across observed levels. High entropy does not imply adequate evidence. A dimension can be evenly populated while individual cells remain small, and the index does not account for trial precision. It is used here as a concise description of design spread, interpreted jointly with cell counts and the study landscape.

2.2. Outcome uncertainty and prediction intervals

For each pooled outcome, the standard error of the mean effect was recovered from its 95% confidence limits (L, U) as

$$SE_{\hat{\mu}} = \frac{U - L}{2 \times 1.96}. \tag{2}$$

The reported τ^2 was retained rather than re-estimated because study-level effects and within-study variances were not available. A two-sided 95% prediction interval for the true effect θ_{new} in a comparable new trial was calculated as

$$\hat{\mu} \pm t_{0.975,k-2} \sqrt{\tau^2 + SE_{\hat{\mu}}^2}, \tag{3}$$

where k is the number of comparisons. The t critical value reflects the small number of studies, particularly for sleep. Prediction intervals can be unstable when τ^2 is imprecisely estimated; they are presented as translations of the reported model rather than fixed population limits. This emphasis follows recommendations to report prediction intervals alongside random-effects means.[17,20]

The predictive distribution was approximated as normal with mean $\hat{\mu}$ and variance $\tau^2 + SE_{\hat{\mu}}^2$. Four probabilities were calculated: any benefit, $\Pr(\theta_{\text{new}} < 0)$; at least a small benefit, $\Pr(\theta_{\text{new}} < -0.20)$; at least a moderate benefit, $\Pr(\theta_{\text{new}} < -0.50)$; and harm, $\Pr(\theta_{\text{new}} > 0)$. These thresholds are interpretive anchors on the standardized scale, not universal clinical cutoffs. To show sensitivity to the chosen requirement, $\Pr(\theta_{\text{new}} < -\delta)$ was evaluated continuously for δ from 0 to 1.25. The threshold curve prevents a favorable conclusion from depending on one selected boundary.

A Monte Carlo check drew 200,000 values of the pooled mean from $N(\hat{\mu}, SE_{\hat{\mu}}^2)$ and then drew one true effect from $N(\mu^{(b)}, \tau^2)$ for each iteration b. The seed was 20260814. Agreement within 0.002 between simulation and analytic small-benefit probabilities was required. The simulation is a numerical verification of uncertainty propagation; it does not simulate people, symptoms, attrition, or treatment delivery.

2.3. Delivery contrasts

Two contrasts were selected because their subgroup tests had been emphasized clinically and because they addressed distinct service decisions: therapist-guided versus self-help delivery for anxiety, and daily treatment of at most two weeks versus weekly treatment of seven to eight weeks for depression. For each subgroup mean, the standard error was recovered from its reported confidence interval. Under an independence approximation, the contrast $\Delta = \hat{\mu}_A - \hat{\mu}_B$ had standard error

$$SE_\Delta = \sqrt{SE_A^2 + SE_B^2}. \tag{4}$$

Negative values favored the first level named. We calculated the 95% confidence interval, $\Pr(\Delta < 0)$, and $\Pr(\Delta < -0.20)$. The independence assumption is conservative only under some covariance structures and cannot replace a participant-level interaction model. The dose contrast was explicitly down-weighted in interpretation because the short-daily depression level contained one comparison. No causal ranking across all delivery combinations was attempted.

2.4. Multiplicity, missing structure, and certainty rules

The analysis did not retest all 18 subgroup rows as separate discoveries. It propagated uncertainty for two clinically defined contrasts and treated the remaining subgroup summaries descriptively. No multiplicity-adjusted declaration was attached to the probability statements. Cross-outcome covariance was unavailable, so the four outcomes were not combined into a single score. Robust variance estimation can address dependent effects when study-level effect estimates and a defensible covariance structure are available;[14] those prerequisites were absent. Likewise, multivariate pooling was not approximated by assigning an arbitrary correlation.

Interpretation used three rules. First, a mean confidence interval excluding zero established precision of the average, but did not establish transportability. Second, a prediction interval crossing zero indicated that no-effect or adverse true effects remained compatible with the reported heterogeneity. Third, a delivery contrast required both a high directional probability and more than one contributing comparison in each level before it could support a practice preference. These rules separate numerical strength from evidence architecture. Conventional heterogeneity quantities were retained because I^2 describes the proportion of observed variation attributable to between-study differences,[16] but I^2 alone was not treated as the magnitude of clinical variation.

3. RESULTS

3.1. Trial composition and outcome coverage

The 13 trials randomized a median of 51 participants (range 27–599). Six interventions involved a therapist and seven were self-guided. Five used websites, four online platforms, three mobile applications, and one a hospital call system. Duration ranged from a single session to 56 days. The participant-weighted female proportion was 77.5%, indicating that the pooled evidence describes predominantly female samples. General-community samples contributed five trials, university students three, people with COVID-19 three, and health-care workers two. Sample size and duration were not aligned: the largest trials were self-guided programs lasting three to eight weeks, while several therapist-guided trials randomized fewer than 60 participants.

The design landscape in Fig. 1 makes this imbalance visible. Self-guided studies occupy much of the high-sample region, whereas therapist-supported studies cluster at smaller sample sizes across both brief and eight-week treatments. Because guidance is partly entangled with trial size, delivery channel, and duration, the observed guidance contrast cannot be interpreted as though guidance had been randomized across otherwise identical programs. The plot also shows that the available evidence is broad in format but thin within many specific combinations.

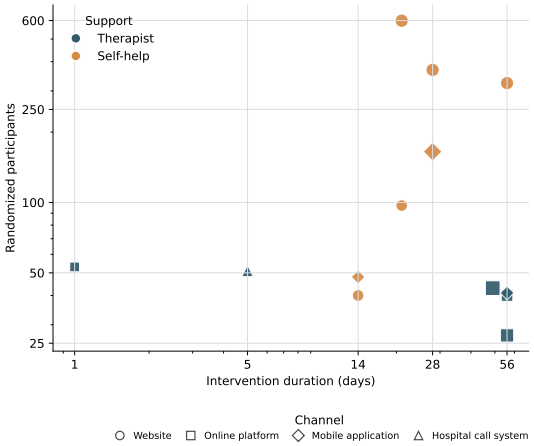

Figure 1. Trial duration and sample-size landscape.

Normalized entropy was 0.996 for guidance, 0.913 for delivery channel, 0.961 for population, and 0.906 for intervention family. These values confirm that the collection spans several forms of care rather than being dominated by one observed category. Their interpretation is limited by small cell counts: eight intervention labels across 13 studies create nominal diversity without enough replication for stable high-order interactions. In practical terms, breadth of design is a strength for detecting variation but a weakness for explaining it.

Outcome availability was asymmetric (Fig. 2). Depression, anxiety, and stress each had nine comparisons, whereas sleep had four. Only one trial supplied all four domains. Several large trials omitted sleep, and two sleep comparisons came from small therapist-supported studies. Repeated measurement of emotional outcomes within the same trials means the three nine-comparison

summaries share evidence, while the sleep estimate rests on a smaller and partly different subset. The matrix therefore argues against reading outcome ranks as though they were obtained from identical trial collections.

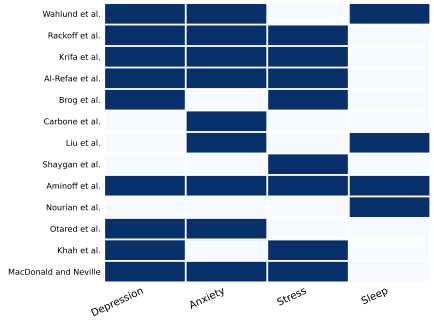

Figure 2. Outcome coverage across the 13 trials.

3.2. Mean effects and predictive uncertainty

All four pooled SMDs were negative, favoring online psychotherapy (Table 1). Confidence intervals for depression, anxiety, and stress excluded zero; the interval for sleep did not. Between-study variance increased from depression ($\tau^2 = 0.09$) to anxiety (0.17), stress (0.28), and sleep (0.43). The corresponding I^2 values were 78%, 87%, 88%, and 88%, showing that dispersion was not a peripheral feature of any outcome.

The prediction intervals crossed zero for every domain (Fig. 3). Their widths materially changed the clinical reading. Depression had a precise beneficial mean, yet its prediction interval allowed a true effect as high as 0.32 in the unfavorable direction. Anxiety combined the second-largest mean effect with the lowest harm probability, 0.065. Stress had the largest mean benefit but also greater between-study variance; its harm probability, 0.098, exceeded that for anxiety. The sleep interval was exceptionally wide because only four comparisons contributed and τ^2 was largest. A conclusion that remote psychotherapy improves sleep would therefore require stronger assumptions than a conclusion about emotional distress.

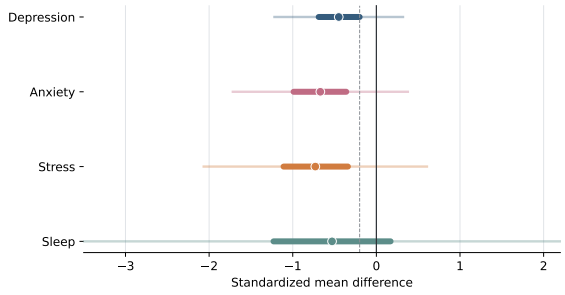

Figure 3. Confidence and prediction intervals by outcome.

The probability of any benefit was 0.917 for depression, 0.935 for anxiety, 0.902 for stress, and 0.761 for sleep. Requiring at least a 0.20-SMD benefit reduced these probabilities to 0.779, 0.856, 0.826, and 0.671. Requiring at least 0.50 produced probabilities of 0.439, 0.650, 0.658, and 0.516. Thus, anxiety rather than stress was most attractive at the small-benefit threshold, but stress slightly exceeded anxiety at the moderate-benefit threshold because its mean was larger and its predictive distribution wider. This reversal is not contradictory. Greater dispersion places more mass in both highly favorable and unfavorable regions.

Continuous threshold profiles in Fig. 4 show that no outcome dominates over the full range of requirements. Anxiety maintains the highest probability through much of the small-effect region; stress becomes comparable or higher as the required benefit increases; depression declines more rapidly because its mean is nearer zero and its between-study variance is smaller. Sleep begins with the weakest directional confidence, then declines more slowly at stringent thresholds because its uncertainty is so large. That tail behavior should not be mistaken for reliable large benefit: the same dispersion produces the 0.239 harm probability.

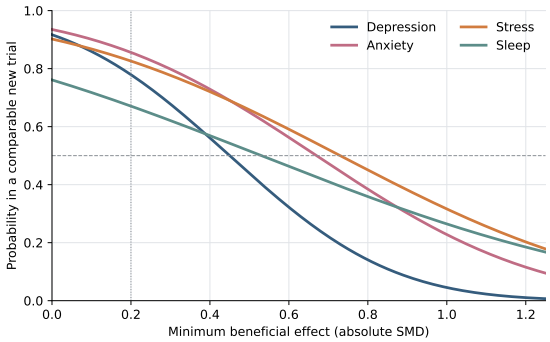

Figure 4. Predictive benefit across effect thresholds.

The Monte Carlo estimates for a benefit of at least 0.20 were 0.777, 0.856, 0.825, and 0.669. Each differed from the analytic result by less than 0.002, satisfying the numerical agreement requirement. This check confirms that the analytic probabilities were implemented as specified; it does not reduce uncertainty in τ^2 or compensate for missing study-level effects.

3.3. Guidance and treatment-dose contrasts

Therapist-guided interventions had an anxiety SMD of -1.16 (95% CI -1.75 to -0.57), compared with -0.35 (-0.67 to -0.03) for self-help. The propagated between-subgroup difference was -0.81 (95% CI -1.48 to -0.14). Its probability of favoring therapist guidance was 0.991, and the probability that the advantage exceeded 0.20 SMD was 0.963 (Table 2). The density in Fig. 5 places little mass above zero, supporting a directional preference for guidance when anxiety reduction is the service priority. Even so, this remains a between-study contrast: therapist involvement co-varied with delivery and sample size, and the calculation does not isolate which component of professional contact produced the difference.

The dose contrast was numerically stronger. Daily treatment lasting at most two weeks had a depression SMD

Table 1. Outcome-level uncertainty translation.

Outcome	k	Mean SMD	95% CI	95% PI	$P_{0.20}$	P_{harm}
Depression	9	-0.45	$[-0.69, -0.20]$	$[-1.22, 0.32]$	0.779	0.083
Anxiety	9	-0.67	$[-0.99, -0.36]$	$[-1.72, 0.38]$	0.856	0.065
Stress	9	-0.73	$[-1.11, -0.34]$	$[-2.06, 0.60]$	0.826	0.098
Sleep	4	-0.53	$[-1.23, 0.17]$	$[-3.74, 2.68]$	0.671	0.239

$P_{0.20} = \Pr(\theta_{\text{new}} < -0.20)$; $P_{\text{harm}} = \Pr(\theta_{\text{new}} > 0)$; PI, prediction interval.

Table 2. Uncertainty in two delivery contrasts.

Contrast	Difference	95% CI	$\Pr(\Delta < 0)$	$\Pr(\Delta < -0.20)$
Therapist minus self-help, anxiety	-0.81	$[-1.48, -0.14]$	0.991	0.963
Short daily minus long weekly, depression	-1.10	$[-1.87, -0.33]$	0.997	0.989

of -1.78 (95% CI -2.53 to -1.04), while weekly treatment lasting seven to eight weeks had an SMD of -0.68 (-0.87 to -0.48). The propagated difference was -1.10 (-1.87 to -0.33), with 0.997 probability of favoring the short daily schedule and 0.989 probability of an advantage exceeding 0.20. The distribution in Fig. 6 is almost entirely favorable. Its evidence architecture is nevertheless fragile because the short-daily level contains one depression comparison. The numerical probability conditions on that estimate and its confidence interval; it does not express uncertainty about whether the single study is representative of short daily programs.

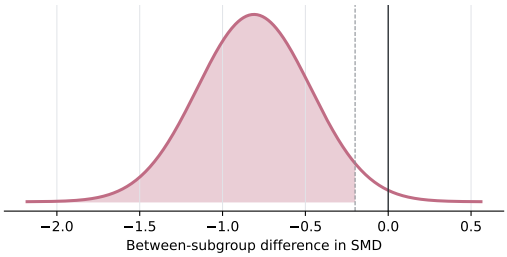

Figure 5. Guidance contrast for anxiety.

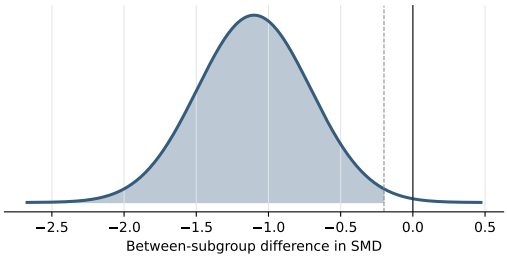

Figure 6. Dose contrast for depression.

Reading the two contrasts together illustrates why probability and replication count must remain separate. The depression dose contrast is farther from zero than the anxiety guidance contrast, but it is less ready for implementation because one short-daily comparison defines the favored level. Guidance has five therapist-supported and four self-help anxiety comparisons, providing a broader empirical basis despite residual confounding. The appropriate action is therefore to prioritize guidance when feasible and to test short daily dosing directly against a longer weekly schedule rather than adopting the latter contrast as settled.

4. DISCUSSION

4.1. Principal interpretation

The central finding is a divergence between average efficacy and predictive certainty. Mean post-treatment effects support online psychotherapy for depression, anxiety, and stress, but the reported heterogeneity leaves room for a null or adverse true effect in every outcome. This does not invalidate the beneficial means. It changes the question they can answer. The mean estimates the center of a distribution of trial effects; the prediction interval and threshold probabilities describe the variation that a future service must confront. For implementation, the second question is often more consequential.

Anxiety provided the most favorable balance at the small-benefit threshold. Its mean effect was substantial, its probability of achieving at least 0.20 SMD in a comparable new trial was 0.856, and its harm probability was 0.065. Stress had a slightly larger mean and the highest probability of exceeding 0.50, but its broader dispersion also increased the possibility of no benefit. Depression showed a more modest but relatively concentrated benefit. Sleep remained qualitatively different: the mean leaned favorable, yet the confidence interval, prediction interval, and harm probability did not support a directional clinical recommendation. These distinctions would be obscured by a single statement that online psychotherapy "works" or "does not work."

The interpretation agrees with the broader digital-treatment literature while adding a transportability qualification. Internet-delivered psychological treatment

has demonstrated efficacy across depression and anxiety, and implementation studies show that structured programs can operate in routine care.[11,37] Individual-participant meta-analysis found that self-guided internet CBT can reduce depressive symptoms and that adherence predicts outcome.[21] A later network meta-analysis suggested that guided treatment is especially useful for people with more severe depression, while unguided treatment may perform similarly for mild symptoms.[22] These findings make it plausible that population severity, support intensity, and adherence contribute to the wide trial distribution observed here.

Mindfulness-based digital interventions also show small-to-moderate benefits for depression, anxiety, and stress across broader populations.[33,34] The pandemic trial collection combined mindfulness, cognitive behavioral therapy, acceptance and commitment therapy, positive psychology, relaxation, and hybrid programs. Such diversity may increase relevance across services, but it also means that the pooled distribution is not the expected performance of one standardized treatment. A service adopting a therapist-guided videoconference protocol should not assume the mean from self-guided applications, websites, call systems, and synchronous platforms applies without adjustment.

4.2. Why heterogeneity changes a clinical decision

I^2 was at least 78% for every outcome, yet percentages alone do not reveal the plausible scale of future effects. The prediction intervals express heterogeneity in SMD units and show its practical consequence. For stress, the interval from -2.06 to 0.60 encompasses a very large benefit and a nontrivial unfavorable effect. A high threshold probability can therefore coexist with meaningful harm probability. This duality is intrinsic to a wide predictive distribution. It should motivate local measurement and adaptive service review, not deterministic optimism or rejection.

The sleep interval illustrates the additional effect of small k. With four comparisons, the t critical value is large, the mean is imprecise, and between-study variance is substantial. Even if the point estimate appears comparable with depression, the evidence carries much less directional information. Sleep was also measured by a partly different set of trials. One intervention may improve distress without changing insomnia, another may target relaxation and sleep directly, and a third may alter device exposure or daily routine in ways not represented by an emotional-symptom score. Treating sleep as interchangeable with depression or anxiety would disregard both clinical mechanism and evidence coverage.

Prediction intervals remain conditional on the random-effects assumptions. The normal distribution of true effects may be inaccurate with a small and heterogeneous trial collection. The reported τ^2 values are themselves uncertain, and recovering standard errors from rounded confidence limits introduces minor approximation. Small-sample procedures such as Hartung–Knapp adjustments often improve interval coverage compared with conventional normal approximations;[19] modified approaches have been recommended when study precision varies sharply.[31] Because study-level estimates were unavailable, the present analysis could not refit those models. The wide intervals should therefore be read as a warning against overgeneralization, not as exact bounds on all future services.

4.3. Meaning of therapist guidance

The anxiety contrast favors therapist involvement with high directional probability. Several mechanisms could account for the difference. A therapist can correct misunderstandings, tailor exposure or cognitive exercises, reinforce completion, and identify worsening symptoms. Scheduled contact creates accountability and may reduce silent disengagement. Support may also signal treatment credibility, which can influence expectancy and persistence. The supportive-accountability account predicts that contact is most effective when the coach is viewed as trustworthy and when expectations are clear.[24] Reviews of internet intervention guidance similarly find a general advantage, but caution that study quality and comparison type affect the estimated magnitude.[4]

The contrast should not be converted directly into a fixed staffing rule. Therapist-supported trials in this collection were generally smaller, and support co-occurred with particular channels and schedules. Guidance also varies in qualification, amount, timing, and purpose. Ten minutes of asynchronous encouragement is not equivalent to a synchronous psychotherapy session. Routine-care evidence indicates that guided internet CBT can be effective at scale,[37] but professional time remains a limiting resource. A rational service may reserve intensive guidance for severe symptoms, risk, poor early response, or low adherence while offering structured self-help to people with milder distress. The present estimates support that tiered logic more strongly than a universal claim that one delivery mode is best.

4.4. Meaning of dose and duration

The short-daily depression contrast is large but rests on one comparison in the favored level. Dose in psychotherapy is not simply elapsed calendar time. It includes session number, practice frequency, therapist contact, content volume, and participant engagement. A brief daily program may concentrate attention during an acute stressor, whereas a weekly eight-week program may allow skills to consolidate and be tested across contexts. Post-treatment SMD cannot determine whether the brief course has comparable durability. Four trials collected follow-up information, but the aggregate records did not provide a follow-up synthesis.

The appropriate next trial is a direct randomized comparison that holds therapeutic content and total contact

approximately constant while varying temporal distribution. Repeated symptom assessments would permit examination of onset, peak response, and decay. Adherence should be defined before recruitment and reported in accordance with digital-intervention guidance, because completion denominators and exposure definitions vary widely. CONSORT-EHEALTH emphasizes transparent reporting of access, use, prompts, and human involvement.[12] Without these details, a label such as "daily" cannot be translated into a reproducible clinical dose.

4.5. Control conditions and practical meaning

All included trials used waiting-list or routine-care controls. These comparisons estimate added benefit over little or no structured psychological contact. They do not establish superiority to credible psychoeducation, supportive attention, face-to-face therapy, or another digital treatment. In psychotherapy research, inactive controls can amplify between-group effects through expectancy, attention, and differential attrition. The present probabilities inherit that comparator. A commissioner choosing between two active services should not treat $P_{0.20}$ as the probability that online psychotherapy is superior to usual in-person care.

This limitation also affects safety interpretation. The calculated harm probability represents a true SMD above zero relative to an inactive control, not adverse events, suicidality, privacy breaches, or therapeutic alliance failures. Digital mental health implementation during the pandemic raised questions about data governance, access inequality, crisis management, and clinician training.[25,38] Those outcomes were not part of the aggregate table. Clinical adoption therefore requires procedures for identity verification, emergency location, safeguarding, confidentiality, technology failure, and referral. Effectiveness evidence is necessary but not sufficient for a safe service.

Implementation outcomes deserve measurement alongside symptoms. Acceptability, adoption, appropriateness, feasibility, fidelity, cost, penetration, and sustainability describe distinct properties of a health service.[27] None can be inferred from a favorable SMD. A self-guided program may produce a smaller average effect but reach many more people; a therapist-led program may produce greater symptom improvement but face capacity constraints. Economic and equity decisions require absolute symptom change, resource use, uptake, completion, and distributional effects, not only standardized differences.

4.6. Outcome dependence and selective availability

Depression, anxiety, and stress effects were obtained from overlapping trials, making naïve cross-outcome ranking overly confident. A participant with reduced generalized distress may improve on several scales, and shared measurement error induces correlation. Multivariate meta-analysis can improve precision and reduce selective-outcome distortion when within-study covariance is available.[28,30] That covariance was not reported in the aggregate tables. The analysis therefore preserved outcome separation and avoided a composite decision score.

Selective outcome availability remains possible. Outcome reporting bias can occur when measured endpoints are omitted or reported incompletely, and empirical reviews show that publication and selective reporting can alter the apparent evidence base.[10] Funnel-plot asymmetry is difficult to interpret with fewer than ten studies and can reflect heterogeneity rather than publication bias.[35] Because each outcome had at most nine comparisons, the absence of visible asymmetry would not resolve this concern. Confidence intervals robust to some publication-bias processes have been developed,[15] but their use requires study-level estimates. The coverage matrix is therefore a descriptive warning: sleep is sparse, and emotional outcomes are repeatedly drawn from the same trials.

5. CONCLUSION

This study asked whether the average benefits of online psychotherapy for pandemic-related distress remain convincing when reported heterogeneity is carried into a comparable new trial, and which delivery contrasts are sufficiently stable to guide care. The answer is outcome-specific. Depression, anxiety, and stress have beneficial pooled means and a high probability of at least some benefit, but none has a prediction interval confined to benefit. Anxiety provides the strongest decision profile at a 0.20-SMD requirement, with an estimated 0.856 probability of reaching that threshold and a 0.065 probability of an unfavorable true effect. Stress may yield larger improvement, but its greater dispersion makes performance less predictable. Sleep cannot support a directional recommendation because its evidence is sparse and its predictive uncertainty is wide.

Therapist guidance for anxiety is the most credible delivery preference: its advantage is large, highly probable, and represented by several comparisons in both levels. The apparent superiority of short daily treatment for depression is numerically stronger but empirically weaker because a single comparison defines the short-daily level. Clinical services should therefore prioritize guided anxiety care when resources allow, monitor depression and stress outcomes locally, and avoid claiming sleep benefit without dedicated evidence. The decisive finding is not that remote psychotherapy has one universal effect, but that probable emotional benefit coexists with substantial implementation-to-implementation variation.

BIBLIOGRAFIA / REFERENCES

1. Andersson G. Internet interventions: past, present and future. *Internet Interventions*. 2018;12:181–188. doi:10.1016/j.invent.2018.03.008.
2. Andersson G, Titov N, Dear BF, Rozental A, Carlbring P. Internet-delivered psychological treatments: from innovation to implementation. *World Psychiatry*. 2019;18(1):20–28. doi:10.1002/wps.20610.
3. Backhaus A, Agha Z, Maglione ML, Repp A, Ross B, Zuest D, Rice-Thorp NM, Lohr J, Thorp SR. Videoconferencing psychotherapy: a systematic review. *Psychological Services*. 2012;9(2):111–131. doi:10.1037/a0027924.
4. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions: a systematic review. *Internet Interventions*. 2014;1(4):205–215. doi:10.1016/j.invent.2014.08.003.
5. Berryhill MB, Culmer N, Williams N, Halli-Tierney A, Betancourt A, Roberts H, King M. Videoconferencing psychotherapy and depression: a systematic review. *Telemedicine and e-Health*. 2019;25(6):435–446. doi:10.1089/tmj.2018.0058.
6. Berryhill MB, Halli-Tierney A, Culmer N, Williams N, Betancourt A, King M, Ruggles H. Videoconferencing psychological therapy and anxiety: a systematic review. *Family Practice*. 2019;36(1):53–63. doi:10.1093/fampra/cmy072.
7. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, Rubin GJ. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet*. 2020;395(10227):912–920. doi:10.1016/S0140-6736(20)30460-8.
8. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based versus face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*. 2018;47(1):1–18. doi:10.1080/16506073.2017.1401115.
9. Chi D, Zhang Y, Zhou D, Xu G, Bian G. The effectiveness and associated factors of online psychotherapy on COVID-19 related distress: a systematic review and meta-analysis. *Frontiers in Psychology*. 2022;13:1045400. doi:10.3389/fpsyg.2022.1045400.
10. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PLoS ONE*. 2013;8(7):e66844. doi:10.1371/journal.pone.0066844.
11. Etzelmueller A, Vis C, Karyotaki E, Baumeister H, Titov N, Berking M, Cuijpers P, Riper H, Ebert DD. Effects of Internet-based cognitive behavioral therapy in routine care for adults in treatment for depression and anxiety: systematic review and meta-analysis. *Journal of Medical Internet Research*. 2020;22(8):e18100. doi:10.2196/18100.
12. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. *Journal of Medical Internet Research*. 2011;13(4):e126. doi:10.2196/jmir.1923.
13. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *Journal of Clinical Epidemiology*. 2011;64(4):407–415. doi:10.1016/j.jclinepi.2010.07.017.
14. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*. 2010;1(1):39–65. doi:10.1002/jrsm.5.
15. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*. 2010;29(29):2969–2983. doi:10.1002/sim.4029.
16. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557–560. doi:10.1136/bmj.327.7414.557.
17. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A*. 2009;172(1):137–159. doi:10.1111/j.1467-985X.2008.00552.x.
18. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, Ballard C, Christensen H, Silver RC, Everall I, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *The Lancet Psychiatry*. 2020;7(6):547–560. doi:10.1016/S2215-0366(20)30168-1.
19. IntHout J, Ioannidis JPA, Borm GF. The Hartung–Knapp–Sidik–Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian–Laird method. *BMC Medical Research Methodology*. 2014;14:25. doi:10.1186/1471-2288-14-25.
20. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*. 2016;6:e010247. doi:10.1136/bmjopen-2015-010247.
21. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, Mackinnon A, Meyer B, Botella C, Littlewood E, et al. Efficacy of self-guided Internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. *JAMA Psychiatry*. 2017;74(4):351–359. doi:10.1001/jamapsychiatry.2017.0044.
22. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Riper H, Patel V, Mira A, Gemmil AW, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*. 2021;78(4):361–371. doi:10.1001/jamapsychiatry.2020.4364.
23. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JEWC. Persuasive system design does mat-

ter: a systematic review of adherence to Web-based interventions. *Journal of Medical Internet Research*. 2012;14(6):e152. doi:10.2196/jmir.2104.
24. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. *Journal of Medical Internet Research*. 2011;13(1):e30. doi:10.2196/jmir.1602.
25. Moreno C, Wykes T, Galderisi S, Nordentoft M, Crossley N, Jones N, Cannon M, Correll CU, Byrne L, Carr S, et al. How mental health care should change as a consequence of the COVID-19 pandemic. *The Lancet Psychiatry*. 2020;7(9):813–824. doi:10.1016/S2215-0366(20)30307-2.
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71.
27. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, Griffey R, Hensley M. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health*. 2011;38(2):65–76. doi:10.1007/s10488-010-0319-7.
28. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*. 2008;9(1):172–186. doi:10.1093/biostatistics/kxm023.
29. Riley RD, Jackson D, Salanti G, Burke DL, Price M, Kirkham J, White IR. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods*. 2015;6(2):157–174. doi:10.1002/jrsm.1129.
30. Riley RD, Jackson D, Salanti G, Burke DL, Price M, Kirkham J, White IR. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ*. 2017;358:j3932. doi:10.1136/bmj.j3932.
31. Röver C, Knapp G, Friede T. Hartung–Knapp–Sidik–Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*. 2015;15:99. doi:10.1186/s12874-015-0091-1.
32. Salari N, Hosseinian-Far A, Jalali R, Vaisi-Raygani A, Rasoulpoor S, Mohammadi M, Rasoulpoor S, Khaledi-Paveh B. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. *Globalization and Health*. 2020;16:57. doi:10.1186/s12992-020-00589-w.
33. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. *JMIR Mental Health*. 2021;8(7):e28168. doi:10.2196/28168.
34. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. *Clinical Psychology Review*. 2016;45:102–114. doi:10.1016/j.cpr.2016.03.009.
35. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, Carpenter J, Rücker G, Harbord RM, Schmid CH, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*. 2011;343:d4002. doi:10.1136/bmj.d4002.
36. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng H-Y, Corbett MS, Eldridge SM, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi:10.1136/bmj.l4898.
37. Titov N, Dear BF, Nielssen O, Staples L, Hadjistavropoulos HD, Nugent M, Adlam K, Nordgreen T, Bruvik KH, Hovland A, et al. ICBT in routine care: a descriptive analysis of successful clinics in five countries. *Internet Interventions*. 2018;13:108–115. doi:10.1016/j.invent.2018.07.006.
38. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR Mental Health*. 2020;7(3):e18848. doi:10.2196/18848.
39. Varker T, Brand RM, Ward J, Terhaag S, Phelps A. Efficacy of synchronous telepsychology interventions for people with anxiety, depression, posttraumatic stress disorder, and adjustment disorder: a rapid evidence assessment. *Psychological Services*. 2019;16(4):621–635. doi:10.1037/ser0000239.
40. Vindegaard N, Benros ME. COVID-19 pandemic and mental health consequences: systematic review of the current evidence. *Brain, Behavior, and Immunity*. 2020;89:531–542. doi:10.1016/j.bbi.2020.05.048.
41. Wind TR, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: the "black swan" for mental health care and a turning point for e-health. *Internet Interventions*. 2020;20:100317. doi:10.1016/j.invent.2020.100317.