Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A systematic review and Bayesian network meta-analysis of digital interventions for depression and anxiety

Chengzhen Liu
Southwest University
Jiang Qiu
Southwest University
Dongtao Wei
Southwest University
Zheng Zhang
South China Normal University
Peng Yao
Southwest University
Zihao Zeng
Hunan Normal University
Xiaobing Cui
Chinese Academy of Sciences
Yang Liu
Jishou University
Antao Chen
Shanghai Jiao Tong University
Geng Li
ligeng@sus.edu.cn

Shanghai Jiao Tong University

Article

Keywords:

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11032447/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A systematic review and Bayesian network meta-analysis of digital interventions for depression and anxiety

Chengzhen Liu[1,2†], Jiang Qiu[1,2†,*], Dongtao Wei[1], Zheng Zhang[3], Peng Yao[1,2], Zihao Zeng[4], Xiaobing Cui[5], Yang Liu[6], Antao Chen[7], Geng Li[7,8*]

Correspondence should be addressed to Jiang Qiu (qiuj318@swu.edu.cn) or Geng Li (ligeng@sus.edu.cn).

Chengzhen Liu and Jiang Qiu contributed equally as co-first authors. Jiang Qiu and Geng Li serve as co-corresponding authors.

Author Affiliations

[1]Faculty of Psychology, Southwest University, Chongqing 400715, China

[2]Key Laboratory of Cognition and Personality (SWU), Ministry of Education, Chongqing 400715, China

[3]School of Psychology, Center for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou 510631, China

[4]Department of Psychology, Cognition and Human Behavior Key Laboratory of Hunan Province, Hunan Normal University, Changsha 410081, China

[5]Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

[6]School of Physical Education, Jishou University, Jishou 416000, China

[7]Brain Health Institute, National Center for Mental Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine and School of Psychology, Shanghai 200030, China

[8]School of Psychology, Research Center for Exercise and Brain Science, Shanghai University of Sport, Shanghai 200082, China

Abstract

Digital mental health interventions differ in how therapeutic content is organized or generated and delivered. We searched five databases through 31 May 2026 and included 404 randomized studies involving 110,037 participants. Twelve Bayesian random-effects network meta-analyses compared intervention architecture, delivery modality and their combinations for depression and anxiety at post-intervention and follow-up. The pooled direct effect across outcomes and stages favored digital interventions over primary controls (Hedges' g = 0.338, 95% credible interval 0.306–0.371). Web ranked highest at post-intervention and App at follow-up, although web–app contrasts were uncertain. Large language model-based interventions ranked favorably, with limited follow-up evidence. Comparative rankings should be interpreted cautiously because most networks showed evidence of inconsistency and confidence in comparisons with Standard care was limited. Digital interventions showed modest symptom benefits, but the comparative evidence does not establish a preferred architecture or delivery modality.

Introduction

Depression and anxiety remain major global public health challenges[1,2]. Recent estimates suggest that depressive and anxiety disorders affect approximately 332 million and 359 million people, respectively[3]. Together, they impose substantial burdens on individuals, families and societies and are associated with elevated risks of self-harm and suicide[2,4]. Despite the availability of effective psychological and pharmacological treatments, access remains limited and highly unequal[3,5]. Costs, workforce shortages and uneven service provision restrict coverage, particularly in low- and middle-income countries and other resource-limited settings, while stigma and concerns about diagnostic labels can delay help-seeking[5–7]. As population need continues to outpace service capacity, closing this treatment gap will require effective support that is accessible, affordable and scalable.

Digital mental health interventions use apps, web-based platforms and other technologies to provide psychological information, support and treatment[8–12]. They address needs ranging from self-management and early support to clinical care. Digital delivery can extend access beyond conventional service settings and reduce reliance on sustained synchronous contact[8,9]. In clinical settings, these interventions may complement diagnostic assessment, pharmacological treatment and clinician-led support.

Comparative evidence synthesis has not fully kept pace with the growing diversity of digital mental health interventions. Meta-reviews and umbrella reviews have primarily characterized broad patterns of effectiveness across heterogeneous intervention types[13,14], whereas more recent comparative reviews and network meta-analyses have distinguished interventions by therapeutic components[15], human guidance[16] and technological form[17], with some adopting multidimensional classification frameworks[18]. Yet delivery modality and intervention architecture do not map neatly onto one another. A single delivery channel, such as an app, may deliver fixed modules, rule-based interactions or content generated by large language models (LLMs), whereas the same intervention architecture may be implemented through different channels[19–21]. This overlap makes it difficult to determine whether observed differences in effects reflect delivery modality, intervention architecture or their specific combinations.

An affordance-based perspective provides a basis for distinguishing these properties[22,23]. Accessibility, interactivity and adaptability can shape both how users receive support and how therapeutic content is organized[21,24]. Intervention architecture describes how content is selected, structured or generated, whereas delivery modality describes the primary channel through which it reaches users[19–21]. Cross-classifying these dimensions retains the combinations

evaluated in trials and allows their comparative effects to be examined within a common evidence base.

The architecture–delivery framework characterizes intervention configuration, but effects may also vary according to the populations targeted, how interventions are implemented and when outcomes are assessed[16,25]. Baseline symptom severity and whether samples were selected using either a clinician-established diagnosis or a validated symptom threshold have both been associated with intervention benefit, although the direction and magnitude of these associations have varied across studies[16,25,26]. Human guidance is among the most extensively studied implementation features, and some meta-analyses suggest that guided interventions may produce greater symptom improvement, although this pattern has not been consistent across intervention types[16,18,27]. Mean age, female proportion, intervention duration, use of a cognitive behavioral therapy (CBT) framework and country income level have also been examined as potential effect modifiers, but findings remain mixed and are often confined to particular populations, symptom domains or forms of digital intervention[25,28,29]. Symptom improvement at the end of treatment and maintenance at follow-up may not show the same comparative patterns; follow-up interval defines the time horizon over which maintenance is evaluated and is therefore important for interpreting longer-term effects[16,30]. At a finer-grained level, autonomy and behavior change techniques (BCTs) offer additional dimensions for characterizing intervention design and content[31–33], but their associations with symptom outcomes have not been systematically compared. Taken together, previous syntheses have generally examined these factors in relation to overall effects or within a single intervention classification[16,25]. Whether their associations are consistent across intervention architecture,

delivery modality and their joint configurations, as well as across depression and anxiety and post-intervention and follow-up outcomes, remains unclear.

We conducted a prospectively registered systematic review and Bayesian network meta-analysis of digital interventions for depression and anxiety. We estimated overall direct effects and compared post-intervention and follow-up outcomes by intervention architecture, delivery modality and their observed combinations (Fig. 1 and Table 1). We also examined prespecified study-level moderators and explored associations with autonomy and behavior change techniques. Sensitivity analyses assessed the dependence of the findings on effect-size construction and statistical assumptions.

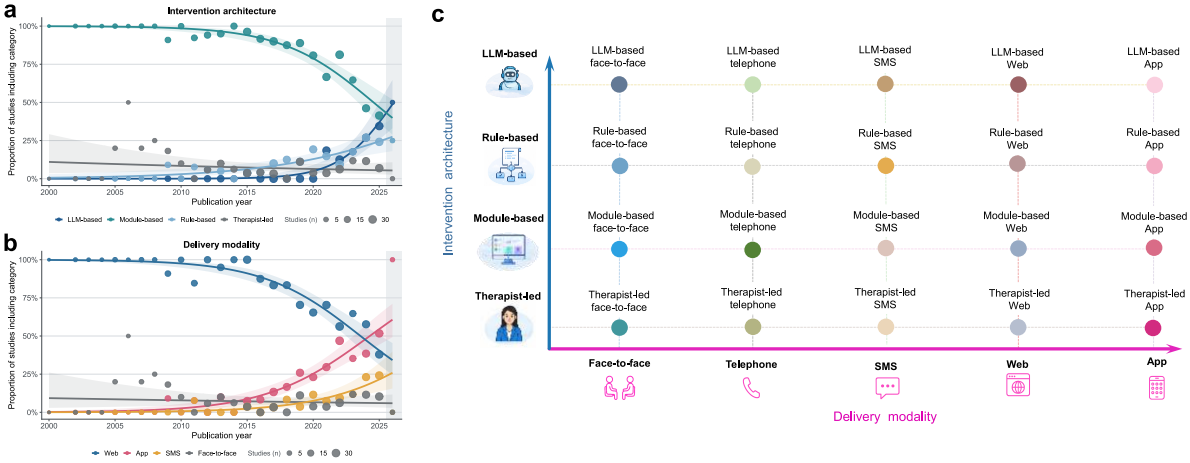


Fig. 1 | Temporal evolution and conceptual classification framework of digital mental health interventions. a, b, Annual proportions of included studies evaluating each intervention architecture (a) and delivery modality (b) by publication year. Points show observed annual proportions, with point size representing the number of studies contributing to each category-year estimate. Solid curves show smoothed temporal trends, with shaded bands indicating 95% confidence intervals. Categories were not mutually exclusive because studies could include multiple intervention arms. Gray vertical shading denotes 2026, for which the literature search covered publications through 31 May. c, Conceptual cross-classification of intervention architecture (Therapist-led, Module-based, Rule-based or LLM-based) and delivery modality (Face-to-face, Telephone, SMS, Web or App). Only combinations observed among eligible intervention arms were included as nodes in the outcome-specific networks. LLM, large language model; SMS, short message service.

Results

Study selection and characteristics

Study selection. Fig. 2 summarizes study selection. The searches identified 174,164 database records and 62 records through other methods. Of 1,980 reports assessed in full text, 377 new studies met the eligibility criteria. Together with 27 eligible studies identified in an earlier version of this review and reassessed against the final eligibility criteria, the final sample comprised 404 randomized studies. Of these, 274 reported both depression and anxiety outcomes, 117 reported depression only and 13 reported anxiety only. Overall, 391 studies contributed depression outcomes and 287 contributed anxiety outcomes.

Study and participant characteristics. The studies included 110,037 participants, with reported mean ages across studies ranging from 8.2 to 84.5 years. The participant-weighted female proportion was 70.7% among 107,974 participants with available data. Publication years ranged from 2000 to 2026; among studies with month-resolved publication dates, dates ranged from September 2002 to May 2026. Eligibility required either a clinician-established diagnosis or a validated symptom threshold in 162 studies (40.1%); neither criterion was required in 215 studies (53.2%), and eligibility varied across participants or study arms in 27 studies (6.7%). Ninety-six studies (23.8%) explicitly targeted participants with a medical condition other than depression or anxiety (Supplementary Table 1).

Intervention and outcome characteristics. The median intervention duration was 8.0 weeks (IQR, 6.0–10.0). Module-based interventions predominated, whereas Rule-based and large language model (LLM)-based interventions were represented in 45 and 32 studies, respectively.

Web delivery was represented in 310 studies, App delivery in 87 and short message service (SMS) delivery in 29; 23 studies included more than one digital delivery modality. At least one active intervention arm used a cognitive behavioral therapy framework in 300 studies (74.3%). Planned human guidance was present across active digital arms in 202 studies (50.0%), varied across active digital arms in 29 studies (7.2%) and was absent in 173 studies (42.8%). Standard care was the most common comparator, represented by 308 randomized arms (Table 1). Because studies could evaluate multiple interventions or comparators, the study-level categories were not mutually exclusive (Supplementary Table 2).

Across assessment stages, 391 studies contributed depression outcomes and 287 contributed anxiety outcomes. Of these, 390 and 286 studies, respectively, reported post-intervention outcomes, whereas 99 and 124 reported follow-up outcomes. The Patient Health Questionnaire-9 was used in 124 of the 391 studies contributing depression outcomes (31.7%), and the Generalized Anxiety Disorder-7 in 105 of the 287 studies contributing anxiety outcomes (36.6%) (Supplementary Table 3).

Process measures and assessment timing. Post-intervention retention, used as a proxy for acceptability, was reported in 88 studies comprising 102 active intervention arms and 11,921 participants; 9,454 participants were retained (79.3%). Compliance data were available for 247 studies comprising 281 active arms and 37,616 participants, with a median arm-level compliance of 72.1% (IQR, 55.0–84.6%). Only 26 studies (6.4%) explicitly reported the interval between intervention completion and the designated post-intervention assessment. Of these, 24 assessed outcomes at treatment completion, one within 1 week and one within 4 weeks after completion (Supplementary Table 4).

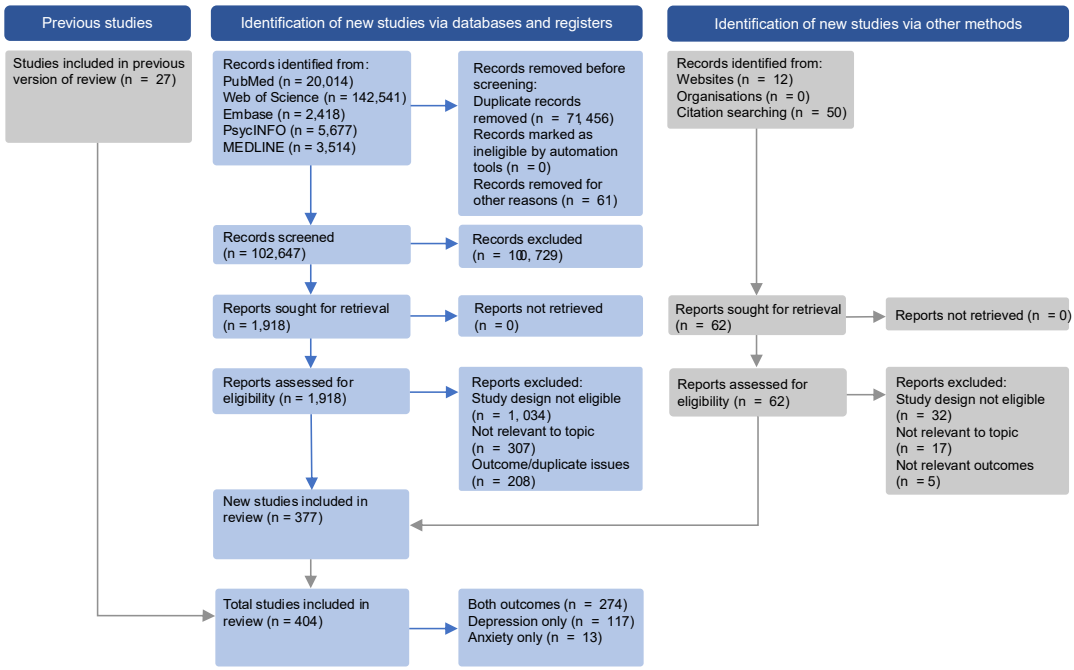


Fig. 2 | PRISMA flow diagram of study selection. The review included 404 randomized studies, of which 274 reported both depression and anxiety outcomes, 117 reported depression only and 13 reported anxiety only. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Overall effects of digital mental health interventions

Across 858 direct effect-size estimates from 391 independent analysis units in 379 studies, 739 (86.1%) favored digital interventions over primary controls and 119 (13.9%) did not (Fig. 3). Primary controls comprised Standard care, Education control and Placebo control. Twenty-five studies evaluated only comparisons with other active interventions and therefore did not contribute to this pooled estimate. Both multilevel models yielded a pooled point estimate of Hedges' g = 0.338, with a Bayesian 95% credible interval of 0.306 to 0.371 and a frequentist 95% confidence interval of 0.305 to 0.370. This estimate summarizes direct contrasts across symptom domains and assessment stages. The additive multilevel meta-regression identified differences by assessment stage (P = 0.006) and symptom domain (P = 0.014).

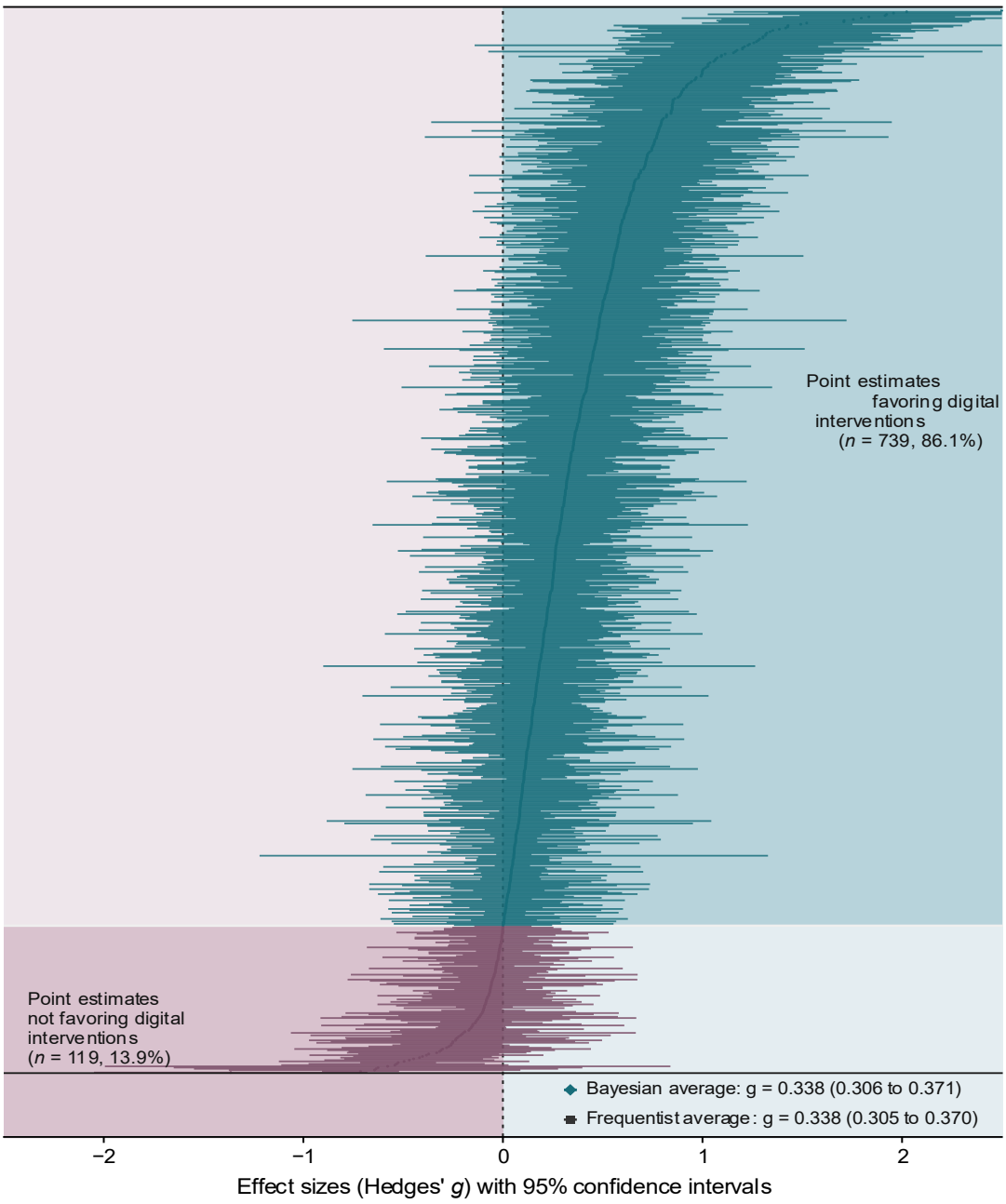


Fig. 3 | Direct effects of digital mental health interventions versus primary controls. Horizontal lines show 858 Hedges' g estimates with 95% confidence intervals; positive values indicate greater symptom improvement with digital interventions. Estimates were derived from 391 independent analysis units in 379 studies, and n denotes the number of effect-size estimates. The Bayesian and frequentist pooled estimates were both g = 0.338 (95% CrI, 0.306 to 0.371; 95% CI, 0.305 to 0.370, respectively). Counts describe the direction of point estimates, not the number of statistically significant effects. CI, confidence interval; CrI, credible interval.

Network meta-analysis

All 12 primary Bayesian network models converged satisfactorily, and model-fit diagnostics were acceptable (Supplementary Table 5). Complete Bayesian league tables and rank-probability summaries are provided in Supplementary Table 6.

Comparative effects by delivery modality. Relative to Standard care, Web was associated with greater improvement in post-intervention depression (g = 0.390, 95% CrI, 0.346 to 0.434; N = 33,949; k = 302) and anxiety (g = 0.352, 95% CrI, 0.299 to 0.405; N = 20,929; k = 212), ranking highest in both post-intervention networks (SUCRA = 0.860 and 0.902, respectively; Fig. 4). App was associated with greater improvement in follow-up depression (g = 0.275, 95% CrI, 0.146 to 0.411; N = 6,895; k = 20) and anxiety (g = 0.286, 95% CrI, 0.178 to 0.394; N = 4,636; k = 28), ranking highest in both follow-up networks (SUCRA = 0.881 and 0.950, respectively). Relative to Face-to-face interventions, neither Web at post-intervention nor App at follow-up showed clear evidence of a difference in either symptom domain (Supplementary Table 7a). The Web–App contrasts were uncertain in all four outcome-stage combinations, with 95% credible intervals spanning zero (Supplementary Tables 6e–6h).

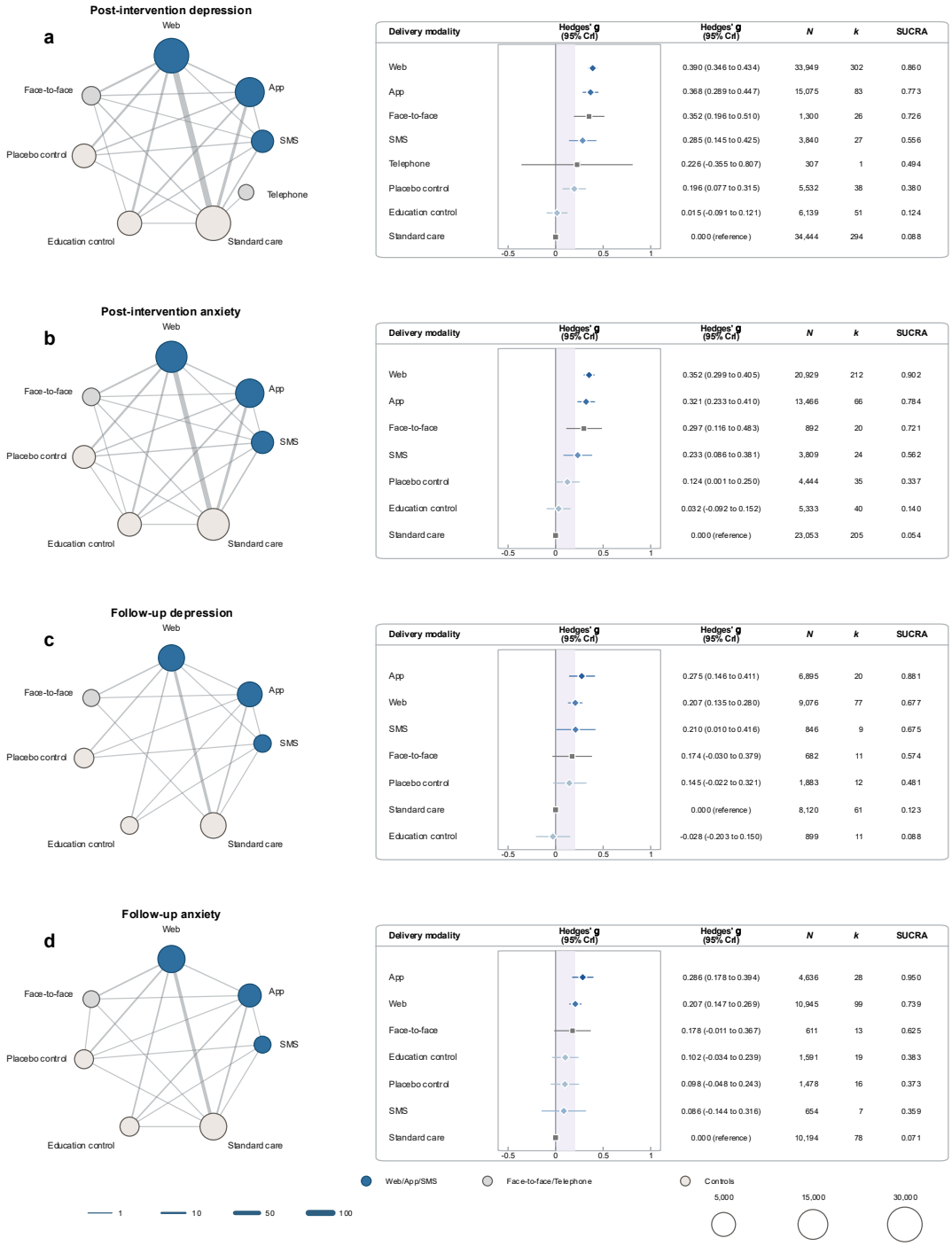


Fig. 4 | Delivery modality network geometry and Bayesian network meta-analysis estimates relative to Standard care. a–d, Evidence networks and corresponding forest plots for post-intervention depression (a), post-intervention anxiety (b), follow-up depression (c) and follow-up anxiety (d). Node area is proportional to the number of participants contributing to each node, and edge width is proportional to the number of direct comparisons. Node colors distinguish Web, App and SMS delivery, Face-to-face and Telephone delivery, and control conditions. Forest plots show Bayesian random-effects network meta-analysis estimates of Hedges' g with 95% credible intervals relative to Standard care, which comprised waitlist and usual or routine care and was set to g = 0. Positive values favor the listed intervention. Higher SUCRA values indicate a more favorable estimated ranking. N and k describe evidence contributing to each node, not exclusively to its direct comparison with Standard care. CrI, credible interval; N, participants contributing to the node; k, network study units; SMS, short message service; SUCRA, surface under the cumulative ranking curve.

Comparative effects by intervention architecture. Relative to Standard care, LLM-based interventions were associated with greater improvement in post-intervention depression (g = 0.425, 95% CrI, 0.285 to 0.564; N = *2,620;* k = 33) and anxiety (g = 0.404, 95% CrI, 0.264 to 0.546; N = 2,736; k = 33), ranking highest in both post-intervention networks (SUCRA = 0.896 and 0.927, respectively; Fig. 5). LLM-based interventions also ranked highest for follow-up depression (g = 0.584, 95% CrI, 0.258 to 0.906; N = *298;* k = 6; SUCRA = 0.987) and follow-up anxiety (g = 0.386, 95% CrI, 0.164 to 0.606; N = *542;* k = 7; SUCRA = 0.964). Relative to Therapist-led interventions, the 95% credible intervals included zero for both post-intervention outcomes and follow-up anxiety, whereas the follow-up depression estimate favored LLM-based interventions (g = 0.407, 95% CrI, 0.036 to 0.776; Supplementary Table 7b).

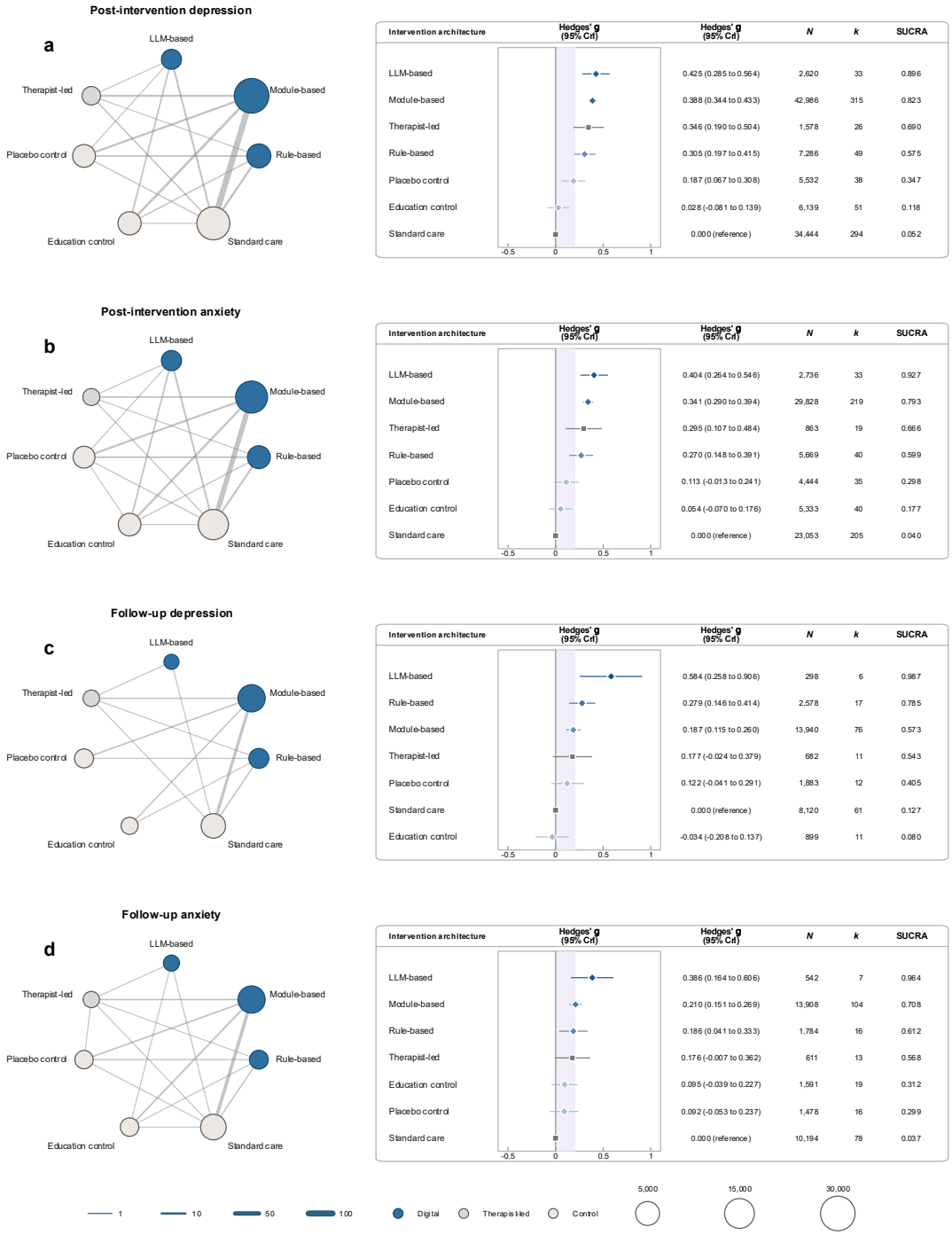


Fig. 5 | Intervention architecture network geometry and Bayesian network meta-analysis estimates relative to Standard care. a–d, Evidence networks and corresponding forest plots for post-intervention depression (a), post-intervention anxiety (b), follow-up depression (c) and follow-up anxiety (d). Node area is proportional to the number of participants contributing to each node, and edge width is proportional to the number of direct comparisons. Node colors distinguish digital intervention architectures, Therapist-led interventions and control conditions. Forest plots show Bayesian random-effects network meta-analysis estimates of Hedges' g with 95% credible intervals relative to Standard care, which comprised waitlist and usual or routine care and was set to g = 0. Positive values favor the listed intervention. Higher SUCRA values indicate a more favorable estimated ranking. N and k describe evidence contributing to each node, not exclusively to its direct comparison with Standard care. CrI, credible interval; LLM, large language model; N, participants contributing to the node; k, network study units; SUCRA, surface under the cumulative ranking curve.

Comparative effects by architecture–delivery modality. In the joint framework, LLM-based Web ranked highest for post-intervention depression (g = 0.441, 95% CrI, 0.198 to 0.680; N = 617; k = 10; SUCRA = 0.755) and post-intervention anxiety (g = 0.460, 95% CrI, 0.227 to 0.690; N = 763; k = 13; SUCRA = 0.845; Fig. 6). LLM-based App ranked highest for follow-up depression (g = 0.585, 95% CrI, 0.261 to 0.901; N = 257; k = 6; SUCRA = 0.942) and follow-up anxiety (g = 0.440, 95% CrI, 0.213 to 0.663; N = 415; k = 7; SUCRA = 0.941). Relative to Therapist-led face-to-face interventions, the 95% credible intervals for the highest-ranked combinations included zero for both post-intervention outcomes and follow-up anxiety, whereas the follow-up depression estimate favored LLM-based App (g = 0.405, 95% CrI, 0.038 to 0.766; Supplementary Table 7c).



Fig. 6 | Architecture–delivery modality networks and Bayesian network meta-analysis estimates. a–d, Networks and forest plots for post-intervention depression (a), post-intervention anxiety (b), follow-up depression (c) and follow-up anxiety (d). Nodes represent observed architecture–delivery combinations. Node area reflects participant numbers; edge width reflects direct-comparison counts. Colors distinguish LLM-based, Module-based, Rule-based, Therapist-led and control interventions. Forest plots show Hedges' g with 95% credible intervals relative to Standard care (waitlist or usual/routine care; g = 0). Positive values favor the listed intervention. Higher SUCRA values indicate more favorable estimated rankings. N and k describe evidence contributing to a node, not solely its direct comparison with Standard care. CrI, credible interval; LLM, large language model; N, participants; k, network study units; SMS, short message service; SUCRA, surface under the cumulative ranking curve.

Moderator analyses

Bayesian moderator estimates were generally consistent in direction and magnitude across the three classification frameworks (Fig. 7 and Supplementary Table 8). Greater mean baseline symptom severity was associated with larger effects across all four outcome-stage combinations: post-intervention depression (β = 0.080–0.085), post-intervention anxiety (β = 0.070–0.080), follow-up depression (β = 0.115–0.120) and follow-up anxiety (β = 0.060–0.063). For post-intervention depression, symptom-threshold samples (β = 0.188–0.190), human guidance (β = 0.150–0.157) and use of a CBT framework (β = 0.144–0.153) were associated with larger effects, whereas longer intervention duration was associated with smaller effects (β = −0.047 to −0.045). For post-intervention anxiety, symptom-threshold samples (β = 0.153–0.159) and human guidance (β = 0.112–0.126) were associated with larger effects. At follow-up, longer follow-up intervals were associated with smaller depression effects (β = −0.079 to −0.064), whereas human guidance was associated with larger anxiety effects (β = 0.095–0.124). For each association described above, the 95% credible intervals excluded zero under all three frameworks. Female proportion and mean participant age showed no clear associations. High-income-country setting was negatively associated with post-intervention anxiety only under the delivery modality framework, with no consistent association across frameworks. Autonomy showed no clear association with intervention effects. Some behavior change technique coefficients had credible intervals that excluded zero, but the findings depended on the model specification and interval level (Supplementary Table 9).

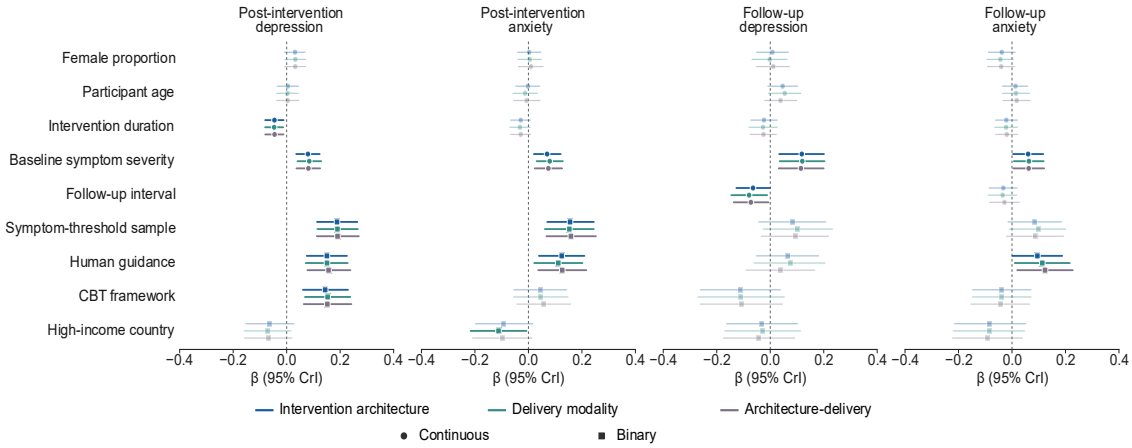


Fig. 7 | Bayesian moderator associations across the three classification frameworks. Points show posterior mean aggregate-level meta-regression coefficients (β) with 95% credible intervals for the four outcome-stage combinations. Positive coefficients indicate larger intervention effects, and dashed lines mark $\beta = 0$. Each coefficient was estimated in a separate Bayesian random-effects meta-regression that included the active intervention node under the indicated classification framework and an analysis-unit random intercept; moderators were not entered simultaneously. Continuous moderators were standardized within each outcome-stage combination and classification framework, whereas binary coefficients represent the contrast between the presence and absence of the coded characteristic. Circles denote continuous moderators and squares denote binary moderators. CBT, cognitive behavioral therapy; CrI, credible interval.

Risk of bias and confidence in the evidence

Overall risk of bias was judged high in 301 studies, unclear in 102 and low in one (Supplementary Table 10 and Supplementary Fig. 1). These judgments were driven primarily by reliance on self-reported outcomes, the practical difficulty of blinding participants and personnel, incomplete outcome data and limited reporting of allocation procedures. Across the three classification frameworks, CINeMA rated confidence as very low for all 94 comparisons with Standard care (Supplementary Table 11 and Supplementary Fig. 2).

Transitivity and inconsistency

Transitivity. Potential effect modifiers differed across intervention nodes, including intervention duration, human guidance and follow-up interval (Supplementary Tables 12a–12c). These descriptive comparisons informed the plausibility of transitivity but did not establish it.

Assessment was particularly limited for sparsely populated architecture–delivery modality nodes.

Inconsistency. Global design-by-treatment tests indicated inconsistency in 10 of the 12 outcome-by-framework networks (Supplementary Table 13). Node splitting identified one nominal direct–indirect disagreement among 247 comparisons: LLM-based Web versus Standard care for post-intervention depression in the architecture–delivery modality network (difference = −0.596; P = 0.043). This comparison involved the highest-ranked node in that network. Loop-specific analyses identified 12 nominally significant loops among 232 tests, including five involving a highest-ranked delivery modality node. These local tests were unadjusted diagnostic analyses and were treated as exploratory. The small number of nominal local disagreements does not negate the global inconsistency detected in most networks.

Sensitivity analyses

The overall pooled effect remained positive when analyses were restricted to the PHQ-9 and GAD-7, when one instrument was selected per independent study sample and outcome domain, or when the assumed within-arm instrument correlation was varied from $\rho = 0$ to 0.90; all pooled 95% credible intervals excluded zero (Supplementary Table 14). Varying the pre–post correlation from $r = 0$ to 0.75, selecting the earliest or longest eligible follow-up assessment, and using a common pooled baseline standard deviation yielded pooled estimates of $g = 0.332$–0.343; all 12 networks retained the same highest-ranked node under each corresponding assumption. Results from frequentist network models and omission-rule sensitivity analyses are reported in Supplementary Tables 15 and 16 and Supplementary Fig. 3. Under the

alternative half-Normal heterogeneity prior with standard deviation 0.5, the highest-ranked node remained unchanged in all 12 networks (Supplementary Table 5).

Excluding context-specific anxiety outcomes retained the highest-ranked node in all six anxiety networks and left the two highest-ranked nodes unchanged. Rule-based replaced Therapist-led at rank three in the post-intervention anxiety architecture network, and Rule-based Web replaced Module-based App at rank three in the follow-up anxiety architecture–delivery modality network; the corresponding primary and sensitivity credible intervals overlapped (Supplementary Table 17). Adjustment for symptom-threshold sampling retained the primary highest-ranked node in all 12 networks. The medication-adjusted complete-case analysis changed the delivery-modality leader from Web to App for both post-intervention outcomes, but the same changes occurred in the matched unadjusted complete-case analyses, suggesting that the ranking changes reflected restriction to studies with medication information rather than covariate adjustment itself (Supplementary Table 18).

Small-study effects. Exploratory Egger regression indicated funnel-plot asymmetry ($P < 0.001$). In the conventional random-effects model used for these diagnostics, trim-and-fill changed the pooled estimate from $g = 0.306$ to $g = 0.168$ after imputing 208 effect-size estimates. The non-favoring-effect subset yielded $g = -0.114$. These analyses were exploratory diagnostics and did not replace the primary multilevel estimate of $g = 0.338$ (Supplementary Table 19 and Supplementary Fig. 4).

Discussion

Across 404 randomized studies involving 110,037 participants, digital interventions showed modest symptom improvements relative to primary controls. Web ranked highest at post-

intervention and App at follow-up, while LLM-based interventions received the highest rankings across architecture analyses. These rankings describe the relative positions of interventions within the fitted networks. Their clinical interpretation is limited by global inconsistency in 10 of the 12 networks, very low confidence in comparisons with Standard care and sparse follow-up evidence for several configurations.

In line with previous evidence that digital interventions can reduce depressive and anxiety symptoms[14,16,34–36], we observed a modest positive overall effect of approximately one-third of a standard deviation. This average effect does not by itself establish clinically meaningful improvement at the individual level; modest effects may nevertheless be relevant when interventions are delivered at scale, although symptom-change trials do not by themselves establish population-level preventive impact. The pooled estimates were nearly identical in the Bayesian and frequentist multilevel models and remained positive across sensitivity analyses using alternative effect-size constructions and correlation assumptions.

The architecture–delivery framework distinguishes how therapeutic content is selected, structured or generated from the channel used to deliver it. An app can provide fixed modules, rule-based interactions or generated dialogue, while a given architecture can operate through more than one channel. The joint classification retains the combinations evaluated in trials and shows how the comparative evidence is distributed across them. It compares intervention packages and does not isolate the independent causal effects of architecture and delivery.

Web ranked highest at post-intervention and App at follow-up, but the Web–App contrasts were uncertain. The two assessment stages were informed by different sets of studies, and follow-up intervals varied across interventions. These findings therefore do not show that Web

produces faster improvement or that App produces more durable effects. The App node also included locally installed software and was not restricted to smartphone applications. Its ranking cannot be attributed specifically to mobile access or use in everyday settings. Comparisons with the Face-to-face delivery node showed no clear differences for Web at post-intervention or App at follow-up, but did not establish equivalence.

LLM-based interventions ranked highest across both symptom domains and assessment stages. Compared with Therapist-led interventions, the 95% credible intervals crossed zero for both post-intervention outcomes and follow-up anxiety. The follow-up depression contrast favored LLM-based interventions (g = 0.407, 95% CrI, 0.036 to 0.776), with a similar finding for LLM-based App versus Therapist-led face-to-face interventions in the joint analysis. Follow-up evidence was limited, however: the architecture estimates were based on six network study units involving 298 participants for depression and seven units involving 542 participants for anxiety. The nominal direct–indirect disagreement for LLM-based Web versus Standard care further limits interpretation of its post-intervention depression ranking. The observed follow-up depression difference does not establish a general advantage over therapist-led care across outcomes or assessment stages.

Evaluation of LLM-based digital interventions should also extend beyond symptom outcomes. The evidence base for these interventions is recent and has expanded rapidly, limiting the extent to which current estimates can be generalized across model generations and implementation periods. LLM-based interventions are also heterogeneous, differing in their underlying model and version, prompting procedures, memory functions, therapeutic constraints and degree of human oversight[37,38]. Although open-ended generation may enhance

contextual responsiveness, it may also increase variation in the relevance, consistency and clinical appropriateness of responses[39,40]. Future trials should therefore report these technical and therapeutic characteristics in sufficient detail and assess factually misleading or clinically inappropriate outputs, automation bias, crisis detection and escalation, privacy, accountability and pathways to professional care alongside symptom outcomes[41–43].

Greater mean baseline symptom severity was associated with larger effects in all four outcome-stage combinations (Supplementary Table 8). Symptom-threshold recruitment and human guidance were associated with larger post-intervention effects for both depression and anxiety, and guidance was also associated with larger follow-up anxiety effects. The finding for guidance is consistent with previous evidence favoring supported over fully self-guided digital interventions[16,27]. Associations involving CBT framework, intervention duration and follow-up interval varied across outcomes and assessment stages, while age, female proportion and country income showed little consistent association. These study-level associations do not identify which individual patients benefit or establish a causal effect of human guidance.

The exploratory autonomy and behavior change technique analyses do not isolate the effects of individual intervention components. Because they used aggregate-level coding, the specification-dependent associations should not be interpreted as causal effects of individual techniques.

Several limitations affect the interpretation of these findings. First, most outcomes were assessed using validated self-report measures, which may be affected by recall and reporting biases. Second, participant characteristics, comparator content and assessment timing differed across trials. Standard care combined waitlist with usual or routine care, and follow-up

estimates summarized eligible later assessments rather than effects at a common fixed interval. These differences limit the comparability of network estimates. Funnel-plot asymmetry also suggested possible small-study effects. Third, risk-of-bias and CINeMA assessments indicated limitations in the underlying evidence, particularly for fine-grained comparisons between intervention configurations. Fourth, evidence for LLM-based interventions was limited, particularly at follow-up, and rapid changes in model capabilities restrict generalization across models and implementation periods. Fifth, the classification framework simplified multicomponent interventions and did not fully capture the intensity and function of human guidance or differences in adaptivity, personalization and safety design. Finally, moderator analyses were based on aggregate study-level data, and incomplete medication reporting limited assessment of residual confounding. Assessment retention, intervention completion and the interval between intervention completion and assessment were incompletely reported. Retention was used as a proxy for acceptability but did not measure active intervention use. Heterogeneous usage metrics precluded quantitative synthesis of behavioral engagement.

Digital interventions were associated with modest improvements in depressive and anxiety symptoms relative to primary controls. Distinguishing intervention architecture from delivery modality shows which configurations have been evaluated and where comparative evidence remains sparse. The observed rankings identify candidates for direct comparison but do not establish a preferred technology or delivery channel for clinical care.

Methods

Protocol and registration

This systematic review and network meta-analysis followed PRISMA 2020 and the PRISMA extension for network meta-analysis[44,45] (Supplementary Table 20). The protocol was prospectively registered in PROSPERO (CRD420251242964).

Several deviations from the registered protocol were made before the final analyses. First, the final search date was extended from 1 December 2025 to 31 May 2026 to incorporate newly eligible evidence. Second, CINeMA replaced GRADE as the framework for evaluating confidence in network estimates. Third, intervention nodes were refined after data extraction and organized within three complementary classification frameworks; the final definitions and coding rules were documented before outcome modeling. Fourth, additional sensitivity analyses were added to assess robustness to outcome-instrument selection, within-study dependence assumptions and alternative analytical specifications.

Information sources and search strategy

This review retained 27 eligible studies from an earlier version of the same review, all of which were reassessed against the final eligibility criteria. We searched PubMed, Embase, Web of Science, PsycINFO and MEDLINE from inception to 31 May 2026. Controlled vocabulary and free-text terms covered depression and anxiety, digital or remote delivery, psychological interventions and randomized allocation. We also screened targeted websites and the reference

lists of eligible reports and relevant reviews. Full search strategies and database-specific record counts are provided in Supplementary Table 21.

Eligibility criteria

We included randomized trials in which individuals or clusters were allocated to at least two conditions. Participants of any age, with or without a formal diagnosis, were eligible. At least one arm had to evaluate a psychological, behavioral, supportive or rehabilitative intervention delivered primarily through a digital or remote channel and report depression or anxiety measured using an established domain-specific instrument or subscale[46]. Depression or anxiety could be a primary or secondary outcome. Eligible comparators included another eligible digital intervention, Therapist-led psychological interventions, Education control, Placebo control or Standard care, as defined below.

Trials were required to report extractable arm-level data at post-intervention or follow-up. Formal diagnosis was not required: samples selected using a clinician-established diagnosis, a validated instrument-specific symptom threshold or neither criterion were eligible. Context-specific anxiety outcomes in medical, procedural or performance settings were retained when measured using an established anxiety instrument or subscale and were examined in a dedicated exclusion sensitivity analysis.

English-language journal articles, conference proceedings and publicly accessible preprints were eligible when their methods and outcome data were sufficient for evaluation and effect-size calculation. Concomitant medication, treatment or background services were permitted when the incremental intervention effect could be isolated. Virtual-reality and serious-game

interventions required an explicit therapeutic aim, an eligible depression or anxiety outcome and classification within the finalized frameworks. We excluded non-randomized studies, abstract-only reports, assessment-only technologies, laboratory inductions without therapeutic intent, standalone cognitive training or rehabilitation without a psychological therapeutic component and non-therapeutic games.

Study selection

Records were imported into EndNote 19 and deduplicated. Titles and abstracts and then full-text reports were independently assessed by two reviewers, with reasons recorded for full-text exclusions. Disagreements were resolved by consensus and, when necessary, referral to a senior author. Multiple reports from the same trial were linked and treated as one study, while companion reports were retained when they supplied additional intervention, outcome or follow-up information. Fig. 2 summarizes study selection, and Supplementary Table 22 lists the included reports.

Data extraction

Five investigators extracted study-, participant-, intervention- and outcome-level data using a standardized arm-level form, with all entries independently verified against the source reports by a second investigator. Multiple reports from the same trial were linked before extraction and overlapping reports were consolidated so that each independent study sample was represented once.

For each eligible endpoint, we extracted the analyzed sample size, means and measures of dispersion, or equivalent statistics sufficient to calculate an effect size. We also extracted data required to characterize acceptability, compliance and assessment timing. Study characteristics are summarized in Supplementary Table 1, and study-level classifications are reported in Supplementary Table 2.

Unclear entries were checked against primary reports, registrations, protocols, supplementary materials and linked reports, and study authors were contacted when clarification or numerical data were required.

Intervention classification and comparator coding

Classification framework. Each active arm was assigned an intervention architecture code, a delivery modality code and their joint cross-classification (Table 1). Architecture described how therapeutic content was selected, structured or generated and comprised Therapist-led, Module-based, Rule-based and LLM-based nodes. Delivery described the primary channel and comprised Face-to-face, Telephone, SMS, Web and App. App included mobile or tablet applications and standalone locally installed software accessed outside a browser. The App node was therefore not restricted to smartphone interventions. This distinction accords with existing behavioral intervention technology frameworks[21].

Coding hierarchy. Detailed intervention descriptions took precedence over generic labels. Multicomponent interventions were classified by the primary mechanism through which therapeutic content was selected, structured or generated; when mechanisms overlapped, LLM

generation took precedence over rule-based interaction, which took precedence over modular delivery. Delivery modality followed the predominant channel.

CBT and human guidance. CBT was coded only when an eligible source explicitly described the intervention as CBT-based. Human guidance captured planned feedback, motivation, accountability or clinical input within an otherwise digital intervention and excluded technical or administrative contact. Guidance did not change the architecture to Therapist-led unless a qualified professional delivered the therapeutic content primarily in real time.

Adjudication and comparators. C.L. and G.L. independently coded every eligible arm, with unresolved disagreements adjudicated by A.C. Standard care comprised waitlist and usual or routine care conditions without an additional study-provided structured psychological intervention beyond outcome assessment and otherwise available services. Education control provided information or psychoeducation without structured therapeutic components, whereas Placebo control was attention-, expectancy- or activity-matched without the active components under evaluation. Therapist-led and head-to-head digital comparators were treated as active comparators. Study-level assignments are reported in Supplementary Table 2 and node distributions in Supplementary Tables 12a–12c.

Outcomes and effect-size derivation

Outcome definitions. The four primary outcome-stage combinations were post-intervention depression, post-intervention anxiety, follow-up depression and follow-up anxiety. Post-intervention was defined as the first eligible assessment conducted at treatment completion or the closest eligible assessment thereafter; later eligible assessments were classified as follow-

519 up. All eligible depression and anxiety instruments and follow-up assessments were retained,
520 and Supplementary Table 3 lists the instruments.

521 For outcome instrument ℓ in arm a of study i, all scales were oriented so that positive Hedges'
522 g values denoted greater improvement. Change was defined as the baseline score minus the
523 eligible endpoint score and standardized using the baseline standard deviation from the same
524 arm[47,48].

525 $$\Delta_{ia\ell} = M_{ia\ell,0} - M_{ia\ell,t}, d_{ia\ell} = \frac{\Delta_{ia\ell}}{\mathrm{SD}_{ia\ell,0}} \quad (1)$$

526 The standardized change was corrected for small-sample bias using Hedges' correction[47].
527 Eligible author-reported standardized effects and standard errors were retained when these
528 were the eligible source estimates.

529 $$J(n_{ia}) = 1 - \frac{3}{4n_{ia}-9}, g_{ia\ell} = J(n_{ia})d_{ia\ell} \quad (2)$$

530 For reconstructed arm-level changes, the primary sampling variance used a pre-post
531 correlation of $r = 0.25$. Substitution gives $2(1 - r)/n = 1.5/n$. Sensitivity analyses repeated the
532 reconstruction at $r = 0$, 0.50 and 0.75. Reported standard errors were retained when supplied
533 with an eligible standardized effect.

534 $$v_{ia\ell} = \frac{2(1-r)}{n_{ia}} + \frac{{g_{ia\ell}}^2}{2n_{ia}}, \mathrm{SE}_{ia\ell} = \sqrt{v_{ia\ell}} \quad (3)$$

535 When an arm reported more than one eligible record for the same outcome domain and
536 outcome stage, record-level effects were combined by generalized least squares[49]. Follow-up
537 records could differ by instrument and eligible assessment time. A compound-symmetry
538 covariance matrix was used, with the diagonal equal to each sampling variance and off-
539 diagonal elements determined by the assumed within-arm record correlation ρ.

$$V_{a,lm} = \begin{cases} \mathrm{SE}_{al}^2, & l = m \\ \rho \mathrm{SE}_{al} \mathrm{SE}_{am}, & l \neq m \end{cases} \quad (4)$$

Let g be the vector of eligible record-level Hedges' g values, V its covariance matrix and 1 a conformable vector of ones. The primary analysis used $\rho = 0.60$ and combined records within arm, outcome domain and outcome stage; follow-up aggregates could therefore include multiple instruments and multiple eligible assessment times. Sensitivity analyses repeated the aggregation at $\rho = 0$, 0.30 and 0.90. A separate analysis retained one instrument family per independent study sample and outcome domain using a result-independent hierarchy: the author-designated primary outcome, a domain-specific measure, then the most frequently used eligible measure. Fixed-instrument analyses retained PHQ-9 records for depression and GAD-7 records for anxiety. Further analyses selected the earliest or longest eligible follow-up time with maximal arm coverage, used $r = 0$, 0.50 or 0.75, and replaced arm-specific baseline standard deviations with a common pooled baseline standard deviation across matched study, instrument and assessment sets. The pooled standard deviation was weighted by arm-specific degrees of freedom. Results are reported in Supplementary Table 14.

$$g^* = \frac{1^T V^{-1} g}{1^T V^{-1} 1}, v^* = \frac{1}{1^T V^{-1} 1} \quad (5)$$

For each eligible intervention-comparator pair s, the direct effect was the difference between the aggregated arm-level changes. I and C denote intervention and comparator arms. Arm-level sampling errors were treated as independent when the contrast variance was constructed. Multi-arm trials were retained as connected study designs; shared-control dependence was handled within the network models, and analysis-unit clustering was retained in direct-effect models.

$$y_{is} = {g_{iI}}^* - {g_{iC}}^*, v_{is} = {\mathrm{SE}_{iI}}^2 + {\mathrm{SE}_{iC}}^2 \quad (6)$$

Process measures and assessment timing. Post-intervention retention was used as a proxy for acceptability and was calculated as the number retained at assessment divided by the number randomized to the arm. Compliance was calculated as completed divided by planned intervention dose, capped at 100%. The interval from intervention completion to outcome assessment was recorded only when explicitly reported. These measures were summarized descriptively and are reported in Supplementary Table 4.

Risk-of-bias assessment

Two reviewers (C.L. and G.L.) independently assessed each study using an adapted version of the original Cochrane Collaboration risk-of-bias tool for randomized trials[50], retained for consistency across the earlier and newly included studies. Disagreements were resolved through discussion or referral to a third reviewer. Overall risk was classified as low when all domains were low, high when any domain was high and unclear otherwise; domain definitions and study-level judgments are reported in Supplementary Table 10 and Supplementary Fig. 1.

Statistical analysis

Overall multilevel meta-analysis. The overall estimand comprised direct contrasts of active digital interventions versus primary controls: Standard care, Education control and Placebo control. Therapist-led and head-to-head digital comparators were active comparators and were excluded from this estimand. An analysis unit was an independent study sample; studies with multiple independent samples could contribute more than one unit. Within a unit, arms assigned to the same intervention architecture node were combined to avoid double counting.

The frequentist model was fitted by restricted maximum likelihood, with random intercepts for the independent analysis unit and effect-size row[51,52]. In equation (7), u_j represents between-unit heterogeneity, w_{ij} residual between-effect heterogeneity and ε_{ij} sampling error. Inference used t-based tests and 95% confidence intervals.

$$y_{ij} = \mu + u_j + w_{ij} + \varepsilon_{ij} \quad (7)$$

A parallel Bayesian multilevel model used the observed sampling variance plus a residual row-level heterogeneity variance σ_r^2 and an analysis-unit-level random-effect variance σ_s^2. Posterior uncertainty was summarized using 95% credible intervals. The pooled mean had a Normal prior with mean 0 and standard deviation 1,000, and both heterogeneity standard deviations had Uniform(0, 2) priors.

$$y_{ij} \sim N\left(\mu + u_j, v_{ij} + {\sigma_r}^2\right), u_j \sim N(0, {\sigma_s}^2) \ (8)$$

The Bayesian overall model used three chains, 1,000 adaptation iterations, 2,000 warm-up iterations, 6,000 sampling iterations and thinning by 3.

$$\mu \sim N(0, 10^6), \sigma_s, \sigma_r \sim \text{Uniform}(0,2) \ (9)$$

An additive frequentist meta-regression tested differences by assessment stage and symptom domain.

Network meta-analysis. The generalized-least-squares arm aggregates for each analysis unit, outcome domain and stage were the network input; individual instruments and assessment times were not entered separately. Separate random-effects consistency models[53,54] were fitted for the four outcome-stage combinations under Delivery modality, Intervention architecture and Architecture–delivery modality. Standard care was the reference node; multi-arm trials were retained as connected designs and same-node arms within an analysis unit were combined.

In equation (10), t_i and c_i denote the compared nodes, d denotes a basic treatment effect, u_i is a comparison-level random effect and τ is the common between-study heterogeneity standard deviation.

$$y_i \sim N\left(d_{t_i} - d_{c_i} + u_i, v_i\right), u_i \sim N(0, \tau^2), d_{\mathrm{SC}} = 0 \quad (10)$$

Primary Bayesian models used *a* normal likelihood, identity link, random effects, Normal priors with mean 0 and standard deviation 30 for basic relative treatment effects and $\tau \sim$ Uniform(0, 2). Models used four chains, 5,000 adaptation iterations, 40,000 sampling iterations and thinning by 10. Convergence and model fit were assessed using *R*-hat[55,56], effective sample size, posterior mean residual deviance relative to the number of data points, and the deviance information criterion[57]. All 12 networks were also refitted using *a half-Normal prior for τ with standard deviation 0.5* to assess prior sensitivity (Supplementary Table 5). Posterior node effects are reported as means with 95% credible intervals. Relative rankings were summarized by the surface under the cumulative ranking curve (SUCRA)[58]. Posterior rank probabilities are reported in Supplementary Table 6; rankings were interpreted alongside effect estimates and credible intervals.

$$\mathrm{SUCRA}_t = \frac{\sum_{r=1}^{T-1} \Pr(R_t \leq r)}{T-1} = \frac{T - E(R_t)}{T-1} \quad (11)$$

We also fitted frequentist random-effects network meta-analysis models to assess sensitivity to the inferential framework with the same network definitions and Standard care as the reference. Estimates are reported with 95% confidence intervals and ranked using P-scores[59], where Φ is the standard normal cumulative distribution function. Bayesian network results are reported in Supplementary Table 6; frequentist sensitivity results are reported in Supplementary

Table 15 and Supplementary Fig. 3.

$$P_t = \frac{1}{T-1}\sum_{s\neq t}^{T} \Phi\left(\frac{d_t - d_s}{\mathrm{SE}(d_t - d_s)}\right) \quad (12)$$

Moderator analyses

Prespecified moderator analyses. Study-, sample- and intervention-level moderators were examined in separate Bayesian random-effects meta-regressions of direct contrasts between active digital interventions and primary controls, fitted for each outcome-stage combination and classification framework. Same-node active arms were combined within analysis units. Each model included the corresponding intervention node as a categorical covariate, one moderator and an analysis-unit random intercept; moderators were not fitted simultaneously.

Continuous moderators were female proportion, mean participant age, intervention duration, mean baseline symptom severity and, for follow-up outcomes, follow-up interval[60–64]. Binary moderators were symptom-threshold sample, human guidance, use of a CBT framework and high-income-country setting, classified using the World Bank fiscal-year 2026 definition[65]. Symptom-threshold samples required a clinician-established diagnosis or a validated instrument-specific threshold; mixed samples were excluded from this binary comparison. Continuous moderators were standardized within each outcome-stage combination and classification framework, whereas binary coefficients contrasted the presence versus absence of the coded characteristic. Coefficients were interpreted as aggregate-level associations. Complete coefficient estimates and model diagnostics are reported in Supplementary Table 8.

Models required at least 12 effect-size estimates from eight independent analysis units, with both binary levels represented in at least two analysis units. The intercept had a Normal (0, 1)

prior, random-effect standard deviations had Cauchy (0, 1) priors and regression coefficients used flat priors. Four chains used 3,000 iterations per chain, including 1,000 warm-up iterations, adapt_delta = 0.99 and max_treedepth = 15. Posterior means, 95% credible intervals and posterior probabilities of the estimated direction are reported; coefficients are aggregate-level associations rather than individual-level causal effects (Supplementary Table 8).

$$y_i \mid \theta_i \sim \text{Normal}(\theta_i, s_i^2), \theta_i = \alpha + \delta_{\text{node}[i]} + \beta x_i + u_{\text{study}[i]}, u_j \sim \text{Normal}(0, \tau^2) \quad (13)$$

Exploratory autonomy and behavior change technique analyses. These analyses used the same primary-control contrasts and Bayesian random-effects structure. Autonomy was entered as the prespecified 1–10 score in a separate model for each framework and outcome. Following Noetel et al.[66], behavior change techniques coded as present in at least five analysis units and absent in at least four were entered simultaneously; identical coding vectors were removed. Intervention architecture was the primary node adjustment, with delivery modality and architecture–delivery modality used as framework-sensitivity adjustments. Simultaneous models used four chains of 5,000 iterations, including 2,000 warm-up iterations, and 95% credible intervals. Exploratory one-technique-at-a-time models were adjusted for Intervention architecture and used four chains of 3,000 iterations, including 1,000 warm-up iterations, with 99% credible intervals to reduce overinterpretation under multiple comparisons. Random-effects metaCART[67] with permutation testing, using default settings and seed 42, examined potential combinations and interactions among eligible techniques. All estimable coefficients are reported in Supplementary Table 9.

Exploratory Bayesian network meta-regressions used a shared treatment-by-covariate coefficient relative to Standard care. Symptom-threshold sample compared Yes with No after

667 excluding Mixed and records with missing covariate data; medication use compared explicitly
668 reported Yes with No after excluding records with missing covariate data. Covariate-specific
669 posterior node effects were marginalized over the observed covariate prevalence to obtain
670 adjusted rankings. For medication use, a matched unadjusted model was fitted to each
671 complete-case network to distinguish covariate adjustment from case restriction. These
672 analyses used four chains, 5,000 adaptation iterations, 40,000 sampling iterations and thinning
673 by 10 (Supplementary Table 18).

674 $$\theta_{i,tc} = d_t - d_c + (\beta_t - \beta_c)\left(X_i - \overline{X}\right) \quad (14)$$

675 ## Transitivity, inconsistency, sensitivity and confidence assessment

676 ***Transitivity.*** We appraised transitivity descriptively[68] using node-level distributions of mean
677 participant age, female proportion, intervention duration, mean baseline symptom severity,
678 symptom-threshold sample, medication reporting, human guidance, CBT framework, country
679 income classification and, for follow-up outcomes, follow-up interval. Study-level comparator
680 and clinical-context classifications, outcome-instrument distributions and context-specific
681 anxiety sensitivity analyses were also reviewed. These distributions informed plausibility
682 rather than providing formal statistical tests.

683 ***Inconsistency.*** Global incoherence was assessed using design-by-treatment interaction
684 models[69]. Local disagreement was evaluated using direct–indirect node splitting[70] and loop-
685 specific inconsistency factors[71]. Nominal, unadjusted two-sided P values were treated as
686 diagnostic rather than confirmatory (Supplementary Table 13).

Sensitivity analyses and small-study effects. Comparisons with therapist-led interventions are reported in Supplementary Table 7. Sensitivity analyses examined alternative effect-size construction, outcome and follow-up selection, dependence assumptions, heterogeneity priors, omission rules, frequentist specifications, context-specific anxiety outcomes and covariate-adjusted rankings. Funnel plots and Egger regression[72] were restricted to eligible direct digital-intervention versus primary-control effects and interpreted as exploratory; trim-and-fill[73] and the non-favoring subset analysis were treated as complementary diagnostics rather than corrections (Supplementary Tables 5 and 14–19; Supplementary Fig. 4).

$$y_i = \alpha + \gamma \mathrm{SE}_i + \varepsilon_i \quad (15)$$

Confidence assessment. CINeMA[74,75] was used to assess confidence in the 94 comparisons with Standard care across the 12 networks, using network-specific random-effects study contribution matrices. The six domains were within-study bias, reporting bias, indirectness, imprecision, heterogeneity and incoherence. Imprecision and heterogeneity were judged using Bayesian 95% credible intervals and posterior prediction intervals, respectively, against the prespecified Hedges' g equivalence range of −0.20 to 0.20 (Supplementary Table 11 and Supplementary Fig. 2).

Use of AI tools in manuscript preparation

ChatGPT was used to suggest revisions to the manuscript's wording. The authors reviewed the suggestions and take responsibility for the final text.

Data availability

Data supporting the findings of this study are available via the Open Science Framework at https://osf.io/ak6wv/overview?view_only=5d52e82b6b6a4be3bfdedfd2b7c1807e.

Code availability

Code supporting the findings of this study is available via the Open Science Framework at https://osf.io/ak6wv/overview?view_only=5d52e82b6b6a4be3bfdedfd2b7c1807e.

References

1. Santomauro, D. F. et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 398, 1700–1712 (2021).
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 9, 137–150 (2022). https://doi.org/10.1016/S2215-0366(21)00395-3.
3. World Health Organization. World mental health today: latest data (World Health Organization, 2025). https://www.who.int/publications/i/item/9789240113817.
4. Chesney, E., Goodwin, G. M. & Fazel, S. Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry 13, 153–160 (2014). https://doi.org/10.1002/wps.20128.
5. Patel, V. et al. The Lancet Commission on global mental health and sustainable development. Lancet 392, 1553–1598 (2018).
6. Clement, S. et al. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. Psychol. Med. 45, 11–27 (2015). https://doi.org/10.1017/S0033291714000129.
7. Evans-Lacko, S. et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health surveys. Psychol. Med. 48, 1560–1571 (2018). https://doi.org/10.1017/S0033291717003336.

8. Torous, J., Myrick, K. & Aguilera, A. The need for a new generation of digital mental health tools to support more accessible, effective and equitable care. World Psychiatry 22, 1–2 (2023). https://doi.org/10.1002/wps.21058.
9. World Health Organization. Global strategy on digital health 2020-2025 (World Health Organization, 2021). https://www.who.int/publications/b/57220.
10. Lattie, E. G., Stiles-Shields, C. & Graham, A. K. An overview of and recommendations for more accessible digital mental health services. Nat. Rev. Psychol. 1, 87–100 (2022). https://doi.org/10.1038/s44159-021-00003-1.
11. Lehtimaki, S., Martic, J., Wahl, B., Foster, K. T. & Schwalbe, N. Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview. JMIR Ment. Health 8, e25847 (2021).
12. Torous, J. et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 20, 318–335 (2021).
13. Crocamo, C. et al. Digital health interventions for mental health disorders: an umbrella review of meta-analyses of randomised controlled trials. Lancet Digit. Health 7, 100878 (2025).
14. Philippe, T. J. et al. Digital Health Interventions for Delivery of Mental Health Care: Systematic and Comprehensive Meta-Review. JMIR Ment. Health 9, e35159 (2022).
15. Furukawa, T. A. et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. Lancet Psychiatry 8, 500–511 (2021). https://doi.org/10.1016/S2215-0366(21)00077-8.

16. Karyotaki, E. et al. Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. JAMA Psychiatry 78, 361–371 (2021).
17. Li, S. et al. Efficacy of digital interventions for smoking cessation by type and method: a systematic review and network meta-analysis. Nat. Hum. Behav. 9, 2054–2065 (2025). https://doi.org/10.1038/s41562-025-02295-2.
18. Zhu, H. et al. A systematic review and Bayesian network meta-analysis on the efficacy and potential of mobile interventions for stress management. Nat. Hum. Behav. 9, 1431–1441 (2025).
19. Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M. & Househ, M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. J. Med. Internet Res. 22, e16021 (2020).
20. Hua, Y. et al. A scoping review of large language models for generative tasks in mental health care. npj Digit. Med. 8, 230 (2025).
21. Mohr, D. C., Schueller, S. M., Montague, E., Burns, M. N. & Rashidi, P. The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J. Med. Internet Res. 16, e146 (2014). https://doi.org/10.2196/jmir.3077.
22. Hutchby, I. Technologies, texts and affordances. Sociology 35, 441–456 (2001). https://doi.org/10.1177/S0038038501000219.

23. Leonardi, P. M. When flexible routines meet flexible technologies: affordance, constraint, and the imbrication of human and material agencies. MIS Q. 35, 147–167 (2011). https://doi.org/10.2307/23043493.

24. Nahum-Shani, I. et al. Just-in-time adaptive interventions in mobile health: key components and design principles for ongoing health behavior support. Ann. Behav. Med. 52, 446–462 (2018). https://doi.org/10.1007/s12160-016-9830-8.

25. Plessen, C. Y., Panagiotopoulou, O. M., Tong, L., Cuijpers, P. & Karyotaki, E. Digital mental health interventions for the treatment of depression: A multiverse meta-analysis. J. Affect. Disord. 369, 1031–1044 (2025).

26. Bower, P. et al. Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. BMJ 346, f540 (2013). https://doi.org/10.1136/bmj.f540.

27. Baumeister, H., Reichler, L., Munzinger, M. & Lin, J. The impact of guidance on internet-based mental health interventions: a systematic review. Internet Interv. 1, 205–215 (2014). https://doi.org/10.1016/j.invent.2014.08.003.

28. Kim, J. et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment. Health 10, e43066 (2023). https://doi.org/10.2196/43066.

29. Linardon, J., Cuijpers, P., Carlbring, P., Messer, M. & Fuller-Tyszkiewicz, M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 18, 325–336 (2019).

30. Zainal, N. H., Soh, C. P., Van Doren, N. & Benjet, C. Do the effects of internet-delivered cognitive-behavioral therapy last after a year and beyond? A meta-analysis of 154 randomized controlled trials. Clin. Psychol. Rev. 114, 102518 (2024). https://doi.org/10.1016/j.cpr.2024.102518.
31. Dugas, M., Gao, G. G. & Agarwal, R. Unpacking mHealth interventions: a systematic review of behavior change techniques used in randomized controlled trials assessing mHealth effectiveness. Digit. Health 6, 2055207620905411 (2020). https://doi.org/10.1177/2055207620905411.
32. Michie, S. et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann. Behav. Med. 46, 81–95 (2013). https://doi.org/10.1007/s12160-013-9486-6.
33. Ryan, R. M. & Deci, E. L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. 55, 68–78 (2000). https://doi.org/10.1037/0003-066X.55.1.68.
34. Fatouros, P. et al. Randomized controlled study of a digital data-driven intervention for depressive and generalized anxiety symptoms. npj Digit. Med. 8, 113 (2025).
35. Moshe, I. et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol. Bull. 147, 749–786 (2021).
36. Fu, Z., Burger, H., Arjadi, R. & Bockting, C. L. H. Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis. Lancet Psychiatry 7, 851–864 (2020).

37. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat. Med. 31, 60–69 (2025). https://doi.org/10.1038/s41591-024-03425-5.

38. Huo, B. et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw. Open 8, e2457879 (2025). https://doi.org/10.1001/jamanetworkopen.2024.57879.

39. Guo, Z. et al. Large language models for mental health applications: systematic review. JMIR Ment. Health 11, e57400 (2024). https://doi.org/10.2196/57400.

40. Stade, E. C. et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. npj Ment. Health Res. 3, 12 (2024). https://doi.org/10.1038/s44184-024-00056-z.

41. Ong, J. C. L. et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit. Health 6, e428–e432 (2024). https://doi.org/10.1016/S2589-7500(24)00061-X.

42. De Freitas, J. & Cohen, I. G. The health risks of generative AI-based wellness apps. Nat. Med. 30, 1269–1275 (2024). https://doi.org/10.1038/s41591-024-02943-6.

43. CHART Collaborative. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw. Open 8, e2530220 (2025). https://doi.org/10.1001/jamanetworkopen.2025.30220.

44. Hutton, B. et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann. Intern. Med. 162, 777–784 (2015). https://doi.org/10.7326/M14-2385.

841 45. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting
842 systematic reviews. BMJ 372, n71 (2021). https://doi.org/10.1136/bmj.n71.
843 46. Bandelow, B. et al. Efficacy of treatments for anxiety disorders: a meta-analysis. Int. Clin.
844 Psychopharmacol. 30, 183–192 (2015). https://doi.org/10.1097/YIC.0000000000000078.
845 47. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. Effect sizes based on
846 means. In Introduction to Meta-Analysis 21–32 (Wiley, 2009).
847 https://doi.org/10.1002/9780470743386.ch4.
848 48. Morris, S. B. Estimating effect sizes from pretest-posttest-control group designs. Organ.
849 Res. Methods 11, 364–386 (2008). https://doi.org/10.1177/1094428106291059.
850 49. Berkey, C. S., Anderson, J. J. & Hoaglin, D. C. Multiple-outcome meta-analysis of
851 clinical trials. Stat. Med. 15, 537–557 (1996). https://doi.org/10.1002/(SICI)1097-
852 0258(19960315)15:5<537::AID-SIM176>3.0.CO;2-S.
853 50. Higgins, J. P. T. et al. The Cochrane Collaboration's tool for assessing risk of bias in
854 randomised trials. BMJ 343, d5928 (2011). https://doi.org/10.1136/bmj.d5928.
855 51. Assink, M. & Wibbelink, C. J. M. Fitting three-level meta-analytic models in R: a step-
856 by-step tutorial. Quant. Methods Psychol. 12, 154–174 (2016).
857 https://doi.org/10.20982/tqmp.12.3.p154.
858 52. Van den Noortgate, W., López-López, J. A., Marín-Martínez, F. & Sánchez-Meca, J.
859 Three-level meta-analysis of dependent effect sizes. Behav. Res. Methods 45, 576–594
860 (2013). https://doi.org/10.3758/s13428-012-0261-6.

53. van Valkenhoef, G., Lu, G., de Brock, E. O., Hillege, H., Ades, A. E. & Welton, N. J. Automating network meta-analysis. Res. Synth. Methods 3, 285–299 (2012). https://doi.org/10.1002/jrsm.1054.

54. Dias, S., Sutton, A. J., Ades, A. E. & Welton, N. J. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med. Decis. Making 33, 607–617 (2013). https://doi.org/10.1177/0272989X12458724.

55. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–472 (1992). https://doi.org/10.1214/ss/1177011136.

56. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7, 434–455 (1998). https://doi.org/10.1080/10618600.1998.10474787.

57. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. Bayesian measures of model complexity and fit. J. R. Stat. Soc. B 64, 583–639 (2002). https://doi.org/10.1111/1467-9868.00353.

58. Salanti, G., Ades, A. E. & Ioannidis, J. P. A. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J. Clin. Epidemiol. 64, 163–171 (2011). https://doi.org/10.1016/j.jclinepi.2010.03.016.

59. Rücker, G. & Schwarzer, G. Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Med. Res. Methodol. 15, 58 (2015). https://doi.org/10.1186/s12874-015-0060-8.

60. Furukawa, T. A. et al. Initial severity of depression and efficacy of cognitive-behavioural therapy: individual-participant data meta-analysis of pill-placebo-controlled trials. Br. J. Psychiatry 210, 190–196 (2017). https://doi.org/10.1192/bjp.bp.116.187773.

61. Sextl-Plötz, T. et al. A systematic review of predictors and moderators of treatment outcomes in internet- and mobile-based interventions for depression. Internet Interv. 37, 100760 (2024). https://doi.org/10.1016/j.invent.2024.100760.

62. Reins, J. A. et al. Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother. Psychosom. 90, 94–106 (2021). https://doi.org/10.1159/000507819.

63. Weitz, E. S. et al. Baseline depression severity as moderator of depression outcomes between cognitive behavioral therapy versus pharmacotherapy: an individual patient data meta-analysis. JAMA Psychiatry 72, 1102–1109 (2015). https://doi.org/10.1001/jamapsychiatry.2015.1516.

64. Cuijpers, P. et al. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. J. Affect. Disord. 159, 118–126 (2014). https://doi.org/10.1016/j.jad.2014.02.026.

65. Metreau, E., Young, K. E. & Eapen, S. G. Understanding country income: World Bank Group income classifications for FY26 (1 July 2025–30 June 2026). World Bank Data Blog (2025). https://blogs.worldbank.org/en/opendata/understanding-country-income--world-bank-group-income-classifica.

903 66. Noetel, M. et al. Effect of exercise for depression: systematic review and network meta-
904 analysis of randomised controlled trials. BMJ 384, e075847 (2024).
905 https://doi.org/10.1136/bmj-2023-075847.
906 67. Li, X., Dusseldorp, E. & Meulman, J. J. A flexible approach to identify interaction effects
907 between moderators in meta-analysis. Res. Synth. Methods 10, 134–152 (2019).
908 https://doi.org/10.1002/jrsm.1334.
909 68. Salanti, G. Indirect and mixed-treatment comparison, network, or multiple-treatments
910 meta-analysis: many names, many benefits, many concerns for the next generation
911 evidence synthesis tool. Res. Synth. Methods 3, 80–97 (2012).
912 https://doi.org/10.1002/jrsm.1037.
913 69. Higgins, J. P. T. et al. Consistency and inconsistency in network meta-analysis: concepts
914 and models for multi-arm studies. Res. Synth. Methods 3, 98–110 (2012).
915 https://doi.org/10.1002/jrsm.1044.
916 70. van Valkenhoef, G., Dias, S., Ades, A. E. & Welton, N. J. Automated generation of node-
917 splitting models for assessment of inconsistency in network meta-analysis. Res. Synth.
918 Methods 7, 80–93 (2016). https://doi.org/10.1002/jrsm.1167.
919 71. Veroniki, A. A., Vasiliadis, H. S., Higgins, J. P. & Salanti, G. Evaluation of inconsistency
920 in networks of interventions. Int. J. Epidemiol. 42, 332–345 (2013).
921 https://doi.org/10.1093/ije/dys222.
922 72. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected
923 by a simple, graphical test. BMJ 315, 629–634 (1997).
924 https://doi.org/10.1136/bmj.315.7109.629.

925 73. Duval, S. & Tweedie, R. Trim and fill: a simple funnel-plot-based method of testing and
926 adjusting for publication bias in meta-analysis. Biometrics 56, 455–463 (2000).
927 https://doi.org/10.1111/j.0006-341X.2000.00455.x.
928 74. Nikolakopoulou, A. et al. CINeMA: an approach for assessing confidence in the results of
929 a network meta-analysis. PLoS Med. 17, e1003082 (2020).
930 https://doi.org/10.1371/journal.pmed.1003082.
931 75. Papakonstantinou, T., Nikolakopoulou, A., Higgins, J. P. T., Egger, M. & Salanti, G.
932 CINeMA: software for semiautomated assessment of confidence in network meta-
933 analysis. Campbell Syst. Rev. 16, e1080 (2020). https://doi.org/10.1002/cl2.1080.
934

Acknowledgements

This research was supported by grants from the National Natural Science Foundation of China (32171040 and 32371105). The funder had no role in study design, data collection, analysis, interpretation, manuscript preparation or the decision to submit.

Author contributions

C.L. and G.L. conceived the study and developed the design. C.L., G.L. and X.L. screened titles and abstracts. C.L., G.L., X.L., Z. Zhang and P.Y. assessed full texts and extracted data. C.L. and G.L. independently coded all eligible intervention arms, and A.C. adjudicated unresolved disagreements. C.L. and G.L. performed the statistical analyses and drafted the manuscript. J.Q. and A.C. provided senior supervision and methodological guidance. Z. Zeng, X.C. and Y.L. contributed to the methodological interpretation and contextualization of the findings. All authors interpreted the findings, critically revised the manuscript and approved the submitted version.

Competing interests

The authors declare no competing interests.

Table 1 | Definitions and evidence volume of intervention classification frameworks and network nodes.

Network node	Definition	Study arms, n
Intervention architecture		
Therapist-led	Therapeutic content was delivered primarily in real time by a therapist or other qualified professional; technology, when used, served mainly as a communication channel.	28
Module-based	Predefined therapeutic or educational modules were completed in a fixed or semi-fixed sequence without open-ended generative responses.	340
Rule-based	Responses were selected using predefined scripts, algorithms, decision trees, menus or keyword-triggered rules, without LLM generation.	53
LLM-based	A large language model generated context-sensitive, open-ended responses or adaptive therapeutic dialogue.	41
Delivery modality		
Face-to-face	Therapeutic contact was delivered primarily through in-person sessions.	28
Telephone	Therapeutic contact was delivered primarily through synchronous voice calls.	1
SMS	Primary delivery was through short message service or equivalent short text messaging rather than a Web or App interface.	29
Web	The intervention was accessed primarily through a browser-based website or online platform.	315
App	Interventions delivered through mobile or tablet applications or through standalone locally installed software outside a browser.	89
Architecture–delivery modality		
Joint cross-classification	Each active arm received one architecture and one primary delivery code; their combination defined the joint node (for example, LLM-based App). Only observed combinations entered a network.	—
Comparator nodes		
Standard care	Comparator conditions with no additional study-provided structured psychological intervention beyond outcome assessment and otherwise available services.	308
Education control	Information or psychoeducation without the structured therapeutic components of an active psychological intervention.	57
Placebo control	An attention-, expectancy- or activity-matched condition without the active psychological components under evaluation.	43

Study arms, n, denotes unique randomized arms after deduplication within joint architecture-by-delivery combinations; framework-specific analytical pooling was not used for these counts. An arm represented in multiple outcome domains or assessment points was counted once. Every active arm contributes once to intervention architecture and once to delivery modality; each section therefore totals 462 and the sections should not be summed. Comparator nodes total 408. Joint-node distributions are reported in Supplementary Table 12c. LLM, large language model; SMS, short message service.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationnpjDMfinal.docx