BMJ Mental Health

Methodological quality in randomised clinical trials of mental health apps: systematic review and longitudinal analysis

Jake Linardon [1], Qiang Xie [2,3,4], Caroline Swords [2,3], John Torous [5], Shufang Sun [6], Simon B Goldberg [2,3]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjment-2025-301595).

[1]SEED Lifespan Strategic Research Centre, School of Psychology, Faculty of Health, Deakin University, Geelong, Victoria, Australia
[2]Department of Counselling Psychology, University of Wisconsin–Madison, Madison, Wisconsin, USA
[3]University of Wisconsin–Madison, Madison, Wisconsin, USA
[4]Center for Healthy Minds, University of Wisconsin – Madison, Madison, Wisconsin, USA
[5]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA
[6]Department of Behavioral and Social Sciences, School of Public Health, Brown University, Providence, Rhode Island, USA

Correspondence to
Dr Jake Linardon, Deakin University, Melbourne, Victoria, Australia; jake.linardon@deakin.edu.au
Dr Simon B Goldberg; sbgoldberg@wisc.edu

QX and CS contributed equally.

Received 3 February 2025
Accepted 18 March 2025

To cite: Linardon J, Xie Q, Swords C, *et al*. *BMJ Ment Health* 2025;**28**:1–8.

ABSTRACT

Question This study investigated the methodological rigour of randomised controlled trials (RCTs) of mental health apps for depression and anxiety, and whether quality has improved over time.

Study selection and analysis RCTs were drawn from the most recent meta-analysis of mental health apps for depression and anxiety symptoms. 20 indicators of study quality were coded, encompassing risk of bias, participant diversity, study design features and app accessibility measures. Regression models tested associations between year of publication and each quality indicator.

Findings 176 RCTs conducted between 2011 and 2023 were included. Methodological concerns were common for several quality indicators (eg, <20% were replication trials, <35% of trials reported adverse events). Regression models revealed only three significant changes over time: an increase in preregistration (OR=1.27; 95% CI 1.10, 1.46) and reporting of adverse events (OR=1.32; 95% CI 1.11, 1.56), and a decrease in apps reported to be compatible with iOS and/or Android (OR=0.78; 95% CI 0.64, 0.96). Results were unchanged when excluding outliers. Results were similar when excluding three high-quality studies published between 2011 and 2013, with additional evidence for an increase in modern missing data methods (OR=1.22; 95% CI 1.04, 1.42) and studies reporting intention-to-treat analysis (OR=1.20; 95% CI 1.03, 1.39).

Conclusions Findings provide minimal evidence of improvements in the quality of clinical trials of mental health apps, highlighting the need for higher methodological standards in future research to ensure the reliability and generalisability of evidence for these digital tools.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Smartphone apps for mental health have been widely studied, with randomised controlled trials (RCTs) showing promising efficacy. However, methodological concerns persist, including weak control groups, high risk of bias, small samples, high dropout rates and inadequate adverse event reporting. Despite ongoing critiques, no formal assessment has examined whether trial quality has improved over time.

WHAT THIS STUDY ADDS

⇒ This study provides the first empirical evaluation of whether the methodological quality of RCTs on mental health apps has improved over time. By assessing key trial characteristics, it identifies persistent weaknesses and areas where progress has been made, offering insights into the evolving rigour of this research field.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights persistent methodological shortcomings in trials of mental health apps, emphasising the need for more rigorous study standards. By identifying gaps in trial quality, the findings may influence research priorities, encourage policy changes to enforce stricter methodological guidelines and ultimately improve the reliability of evidence used to inform clinical practice and digital health policy.

INTRODUCTION

There has been a remarkable surge in scientific interest in mental health interventions delivered via smartphone applications (apps). App-based interventions offer capabilities that are not possible with in-person treatment, such as delivering personalised and just-in-time resources around the clock by leveraging passive, active and metadata continuously collected from its users.[1] Hundreds of randomised controlled trials (RCTs) evaluating mental health apps have been conducted, with promising evidence highlighting their efficacy on symptom reduction.[2–4]

Notwithstanding these encouraging findings, concerns have been raised regarding the methodological quality of this field. The earliest reviews of trials of mental health apps highlighted several methodological issues, including the overuse of inactive/passive controls over active/placebo controls,[5 6] the preponderance of trials with considerable risk of bias,[7 8] high dropout rates coupled with inappropriate handling of missing data (complete case analyses),[9 10] small sample sizes,[7] lack of longer follow-ups[11] and/or inadequate reporting of adverse events.[11] These factors are widely known to reduce study quality and lead to an overestimation of the effects of psychological interventions.[12–14] More recent syntheses of this field[2 3 15–17] continue to emphasise these same

methodological shortcomings, raising concerns about the empirical status of mental health apps.

Despite these persistent challenges, there has been no formal evaluation of whether the methodological rigour of RCTs on mental health apps has improved over time. To assess whether these ongoing concerns are valid, we conducted an empirical investigation into whether the quality of trials of apps targeting depression and anxiety has shown improvement over time. Beyond examining the trial features discussed above (ie, use of passive controls, high risk of bias, significant dropout and handling of missing data, small sample sizes, short follow-ups and inadequate adverse event reporting), we also assessed whether improvements have occurred in other trial characteristics that may influence the quality and robustness of evidence regarding the use of apps. These include: (1) preregistration; (2) representation of racial and ethnic minority groups; (3) study uptake rates; (4) app compatibility across iOS and/or Android platforms; (5) public availability of the apps; (6) replication trials (ie, tested an app that had been evaluated previously); and (7) provision of compensation for app use (which we viewed as an undesirable characteristic).

Here we elaborate on the features of trial quality and the rationale for their significance. Rigorous control groups can provide a more accurate estimate of the true effects of apps, as they help distinguish specific intervention effects from non-specific factors (placebo effects, demand characteristics).[17] Trials with lower risk of bias are important for ensuring reliable and accurate results, minimising the impact of confounding factors (eg, unconcealed random allocation, unmasked outcome assessments, etc) and enhancing the credibility of findings for evidence-based clinical practice.[18] Lower attrition can minimise threats to the internal, external and statistical validity of findings.[19] Larger sample sizes can enhance the reliability of the results and improve statistical power.[20] Longer follow-ups are important for understanding whether the effects of the app intervention remain stable, diminish or continue to improve over time. Reporting adverse events ensures patient safety, provides a comprehensive risk-benefit assessment and supports informed decision-making in clinical practice.[15] Preregistration is necessary for preventing selective reporting, reducing bias and building trust in research findings.[21] Sampling racial and ethnic minorities ensures the findings are generalisable, addresses health disparities and promotes equity in the development and evaluation of apps.[22] High uptake rates minimise selection and recruitment bias while also reflecting an app's feasibility and practical relevance.[23] Delivery of publicly available, empirically supported apps compatible across iOS and/or Android enhances the generalisability and real-world relevance of trials, ensuring the interventions are accessible, inclusive and validated across diverse populations and settings. Finally, not compensating participants for engagement helps prevent incentivised behaviour that could artificially inflate adherence rates, thus ensuring that the findings are more reflective of app use in real-world environments.

The present research examined the extent to which RCTs of depression and anxiety apps have demonstrated increased rigour over time across various methodological features, many of which have been explicitly identified as areas for improvement. Based on trends seen in other bodies of literature (eg, mindfulness research),[24] we did not expect improvement in study quality over time.

METHOD

Search strategy and study selection

This research involved a preregistered (https://osf.io/pmrqg?view_only=5b2849901df84757806429b3a03858f9) secondary analysis of the 176 RCTs included in the most recent review of mental health apps for depression and anxiety symptoms.[2] Study data and analysis code are available (https://uwmadison.box.com/s/w16jhkaf871qdnincdamhyr6gqv3dx65). The one deviation we made from our preregistration was the inclusion of a sensitivity analysis with three early high-quality studies excluded. We included this sensitivity analysis to evaluate whether changes in study quality over time were unduly influenced by these points (ie, whether a lack of improvement was driven by having three early high-quality studies).

Four major online databases (Medline, PsycINFO, Web of Science and ProQuest Dissertations) were searched using a combination of terms related to smartphones, RCTs and anxiety/depression (all search terms can be found in Linardon *et al*).[2] Review papers and reference lists of eligible trials were also hand searched. Record screening and selection were performed by two independent researchers. Studies were included if they were RCTs that tested the effects of a stand-alone, app-based, smartphone intervention against either a control condition (passive, care as usual, placebo, etc) or an active therapeutic comparison (eg, face-to-face counselling), and assessed symptoms of depression and anxiety as either primary or secondary outcomes. There were no sample restrictions. Published and unpublished trials were eligible for inclusion. Apps incorporated within a broader treatment plan (eg, pharmacotherapy, psychotherapy, computerised programmes), as well as text message-only interventions, were not included.

Quality assessment

Five criteria from the Cochrane Risk of Bias[18] tool were used to assess for risk of bias: (1) random sequence generation, (2) allocation concealment, (3) blinding of participants or personnel, (4) blinding of outcome assessments (or use of self-report instruments, which does not require any interaction with an assessor) and (5) completeness of outcome data (ie, use of intention-to-treat (ITT) analyses). Each domain was rated as high risk, low risk or unclear risk. The quality assessments were performed in the Linardon *et al*[2] meta-analysis by two independent researchers in consultation with the senior author, where any minor disagreements were resolved by consensus.

Data extraction

Standardised spreadsheets were developed for coding relevant study-level information. Beyond routine sample, intervention and trial features extracted by Linardon *et al*,[2] for this study, we also extracted the following data from eligible trials: (1) trial year (year that appeared on the trial publication); (2) total number of participants randomised; (3) use of active control (ie, not a waitlist or assessment only condition, which we defined as 'no-treatment control' per Goldberg *et al*[17]); (4) whether the trial was preregistered; (5) use of modern missing data methods for handling missingness (eg, multiple imputation, maximum likelihood estimation); (6) percentage of racial/ethnic minorities; (7) attrition rate at post-test; (8) length of longest follow-up; (9) study uptake rate (ie, proportion of those screening as eligible who are randomised); (10) use of an iOS and/or Android-supported app; (11) delivery of a publicly available app to enable replication; (12) whether the study is a replication (ie, tested an app that had been evaluated previously);

Table 1 Descriptive statistics of study characteristics

Variable	Missing	Mean	SD	Min	Max	k	%
Sample size	0	243.90	471.65	18	5017		
No treatment control	0					80	45.45
Preregistered	0					112	63.64
Modern missing	10					100	60.24
RoB domain 1 low	0					125	71.02
RoB domain 2 low	0					46	26.14
RoB domain 3 low	0					45	25.57
RoB domain 4 low	0					176	100.00
RoB domain 5 low	0					112	63.64
RoB all domains low	0					10	5.68
% REM	97	33.44	23.92	0.00	100.00		
Attrition rate	8	20.38	17.98	0.00	86.5		
Follow-up length	4	2.57	2.38	0.00	12.00		
Uptake rate	22	48.72	31.93	1.93	100		
App iOS and/or Android	0					141	80.11
App available	0					91	51.70
Replication	0					33	18.75
Usage paid	62					11	9.65
AE reported	0					61	34.66
% REM reported	0					79	44.89

Missing indicates the number of studies where this item was not reported. k indicates the number of studies with this item coded as 'yes'. RoB items 1 through 5: sequence generation, allocation concealment, blinding of personnel, blinding of outcome/self-report and use of intention-to-treat (ITT) analysis, respectively. Follow-up length indicates postrandomisation follow-up length in months. Mean and SD reported for continuous variables, k and per cent reported for dichotomous variables. 10 studies reported no missing data, so the method for handling missing data was not coded. Eight studies did not report their sample size at post-treatment, 22 studies did not report the number of participants who were screened and 62 studies did not report whether participants were paid for using the app. When studies included multiple control groups, No treatment was coded as 'no' if any of the control groups were more rigorous (eg, active control).
AE, adverse event; Max, maximum; Min, minimum; REM, racial/ethnic minority; RoB, Risk of Bias.

(13) whether participants were compensated for using the app; and (14) reporting of adverse events. Two independent coders (QX and CS) trained by the senior author (SBG) extracted these data and any disagreements were resolved through consensus.

Data analysis

Since we aimed to test whether the methodological rigour of app trials has improved over time, the various design features extracted served as the dependent variables while the year of publication served as the independent variable. Ordinary least squares and logistic regression models were used to assess changes over time using the R statistical software. Given years are an easily interpreted unit, unstandardised coefficients were computed for continuous outcomes. ORs were computed for logistic regression models. We conducted sensitivity analyses to see whether the results remain stable when removing the presence of outliers (±3 SD from the mean) and when removing three high-quality studies from 2011 to 2013 (ie, early high-quality studies). Using the 'pwr.r.test' function in R,[25] our sample of 176 studies provided power to detect small to moderate magnitude associations between time and study characteristics (r=0.21). All analyses were performed using R V.4.4.1.[26]

RESULTS

Study characteristics

Descriptive statistics of the 176 studies are reported in table 1 and a flow chart is shown in online supplemental figure 1. The trial year of publication ranged from 2011 to 2023, with the following frequencies: 2011–2013 (k=1 trial each), 2014–2015 (k=2 trials each), 2016 (k=6), 2017 (k=8), 2018 (k=13), 2019 (k=23), 2020 (k=26), 2021 (k=24), 2022 (k=47) and 2023 (k=22). The average sample size was 243.90 participants (SD=471.65). Slightly less than half of studies used no treatment control groups defined as waitlists or assessment-only conditions; other types of comparison groups delivered included information resources, placebos (non-therapeutic apps, ecological momentary assessments), care as usual or active psychological interventions (face-to-face or web-based treatments)—the full list of comparison groups can be seen in Supplementary Table 1 in Linardon *et al* (https://osf.io/ufb2w/). A majority of studies were preregistered and used modern missing data methods (63.64% and 60.24%, respectively). Very few studies (5.68%) had low risk of bias across all five domains assessed. Slightly less than half of the studies (44.89%) reported the percentage of racial/ethnic minority participants. Within studies reporting the percentage of racial/ethnic minority participants, 33.44% (SD=23.92%) of participants were racial/ethnic minorities, on average. Average attrition at post-treatment was 20.38% (SD=17.98%). Average time to last follow-up assessment was 2.57 months (SD=2.38) postrandomisation. Average study uptake was 48.72% (SD=31.93%). Most studies specified the app under study was compatible with iOS and/or Android (80.11%). Approximately half (51.70%) indicated the app was available to the public. A minority of studies (18.75%) were replications. A very small number of studies (9.65%) indicated they paid participants to use the app. 27 studies reported offering professional guidance for app use. The most commonly used instruments to assess depression were the Patient Health Questionnaire variants (k=59), Depression Anxiety Stress Scale (DASS; k=27), Hospital Anxiety and Depression Scale (HADS; k=17), the Beck Depression Inventory (k=15) and the Center for Epidemiologic Studies Depression Scale (k=8), while the

Table 2 Ordinary least squares and logistic regression results examining changes in study characteristics over time

Variable	Estimate (95% CI)	SE	Log odds	Log odds SE	OR (95% CI)	t/z	P value
Sample size	24.78 (–4.90, 54.47)	15.15				1.64	0.104
No treatment control			0.09	0.07	1.10 (0.96, 1.25)	1.38	0.168
Preregistered			0.24	0.07	1.27 (1.10, 1.46)	3.29	0.001
Modern missing			0.10	0.07	1.11 (0.97, 1.27)	1.50	0.134
RoB domain 1 low			0.08	0.07	1.08 (0.94, 1.24)	1.09	0.274
RoB domain 2 low			–0.02	0.07	0.98 (0.85, 1.13)	–0.24	0.812
RoB domain 3 low			–0.12	0.07	0.88 (0.77, 1.02)	–1.75	0.081
RoB domain 4 low			NA	NA	NA	NA	NA
RoB domain 5 low			0.13	0.07	1.14 (1.00, 1.30)	1.95	0.051
RoB all domains low			–0.11	0.12	0.90 (0.70, 1.14)	–0.89	0.373
% REM	–1.45 (–3.71, 0.80)	1.15				–1.26	0.210
Attrition rate	–0.83 (–2.00, 0.33)	0.60				–1.40	0.163
Follow-up length	0.14 (–0.01, 0.29)	0.08				1.83	0.069
Uptake rate	0.43 (–1.79, 2.66)	1.14				0.38	0.703
App iOS and/or Android			–0.24	0.10	0.78 (0.64, 0.96)	–2.33	0.020
App available			0.02	0.06	1.02 (0.90, 1.16)	0.31	0.758
Replication			0.05	0.09	1.05 (0.89, 1.25)	0.57	0.567
Usage paid			–0.12	0.12	0.89 (0.71, 1.12)	–1.00	0.317
Adverse event reported			0.28	0.09	1.32 (1.11, 1.56)	3.21	0.001
% REM reported			–0.04	0.06	0.96 (0.84, 1.09)	–0.67	0.505

P values from regression models.
t/z indicates test statistic for linear and logistic regression models, respectively. RoB items 1 through 5: sequence generation, allocation concealment, blinding of personnel, blinding of outcome/self-report and use of intention-to-treat (ITT) analysis, respectively. Follow-up length indicates postrandomisation follow-up length in months.
NA, not available (due to all studies being rated as low risk of bias for RoB 4); REM, racial/ethnic minority; RoB, Risk of Bias.

most commonly used instruments to assess anxiety were the Generalized Anxiety Disorder Scale (k=45), the DASS (k=23), the HADS (k=19) and the State-Trait Anxiety Scale (k=19). See Supplementary Table 1 in Linardon *et al* for more details on the measures used across trials (https://osf.io/ufb2w/).

Main analyses

Regression model results are reported in table 2 and figures 1 and 2. Only three domains showed changes over time. Preregistration (OR=1.27; 95% CI 1.10, 1.46) and reporting of adverse events (OR=1.32; 95% CI 1.11, 1.56) have both become more common. Fewer studies over time have indicated that the app under study was available in the iOS and/or Android stores (OR=0.78; 95% CI 0.64, 0.96). The remaining study characteristics did not show changes over time.

Sensitivity analyses

Outliers were detected for total sample size (three high outliers), attrition rate (one high outlier) and follow-up length (three high outliers). Significance tests were unchanged in this sensitivity analysis. Mirroring the primary analysis, there was no evidence for improvements in these study characteristics over time (table 3).

Results were generally similar in a sensitivity analysis that excluded three early high-quality studies (published between 2011 and 2023). Preregistration (OR=1.45; 95% CI 1.22, 1.71) and reporting of adverse events (OR=1.31; 95% CI 1.10, 1.56) were again more common over time. Reporting that the app under study is available in the iOS and/or Android stores was again less common over time (OR=0.79; 95% CI 0.64, 0.98). In this sensitivity analysis, use of modern missing data methods (OR=1.22; 95% CI 1.04, 1.42) and likelihood of a study being rated as low risk of bias through use of ITT analyses (OR=1.20, 95% CI 1.03, 1.39) became more common over time. The remaining characteristics did not show changes over time (see online supplemental table 1).

DISCUSSION

We examined the methodological rigour of trials evaluating apps for depression or anxiety and whether quality has improved over time. We included 176 RCTs conducted between 2011 and 2023 identified in the most recent review.[2] We examined the association between publication year and 20 facets of study quality, encompassing indicators of risk of bias, participant diversity, study design features and app accessibility measures. In our primary analyses, three statistically significant associations were found, indicating an increase in trial preregistration and reporting of adverse events, and a decrease in studies reporting the app under study is available for iOS and/or Android. There was evidence that use of modern methods to handle missing data and low risk of bias related to using ITT (rather than completer) analyses was increasing in sensitivity analyses that removed three early high-quality trials. Overall, findings provide limited evidence indicative of improvements in the quality of clinical trials of mental health apps in some domains, reinforcing concerns about the state of the evidence in this field and highlighting urgent need for change.[5 11 27 28]

It is promising to observe an increase in the preregistration of clinical trials of mental health apps. Preregistration is crucial for helping to mitigate biases and selective reporting, ultimately enhancing the reliability, replicability and trustworthiness of findings regarding the clinical benefits of mental health apps. This trend also aligns with broader patterns of increased preregistration observed across medicine and psychology.[29 30] We suspect that there may be a few reasons for this shift. Researchers may have become more aware of and receptive to open science principles, possibly driven by the replication crisis observed across numerous scientific fields.[31 32] Alternatively, the increase

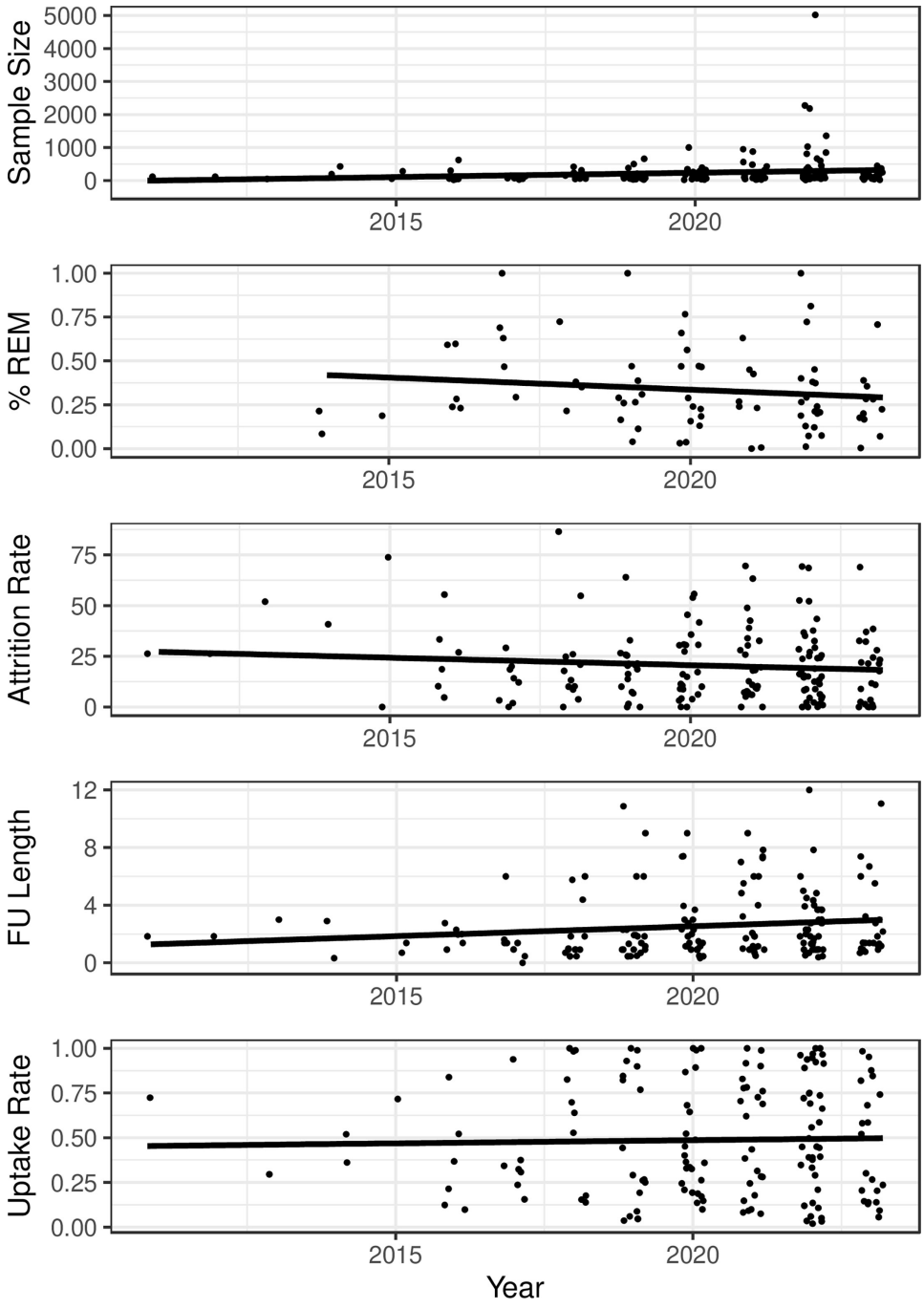

Figure 1 Change in continuous study quality features over time. FU Length, postrandomisation follow-up length in months; REM, racial/ethnic minority participants.

could be driven by the growing number of high-impact scientific journals that now mandate trial preregistration as a condition of publication, or to the pressures from ethical, institutional or funding bodies encouraging preregistration to ensure the robustness and rigour of the research they support.[31] Whatever the reason, this is an encouraging trend that should enhance the transparency and reproducibility of clinical trials for mental health apps.

We found evidence of an increase in the reporting of adverse events in trials of mental health apps. This finding aligns with recent calls to prioritise the documentation of risks and harms in trials of digital health therapeutics.[15 33 34] This finding is also encouraging, as reporting adverse events is critical for ensuring the safety of these tools and upholding ethical standards by fostering transparency and supporting informed decision-making by researchers, clinicians and patients. It is now important for future research to understand the mechanisms behind possible adverse events, determining whether they are directly attributable to the functionality or content of the app, the use of a digital device in general, or influenced by other participant or contextual factors.[15]

When excluding three early high-quality studies, we found evidence that trials are handling missing data more appropriately, that is, using modern missing data methods (eg, maximum likelihood, multiple imputation) and conducting ITT (rather than completer) analyses. Given attrition is common in app

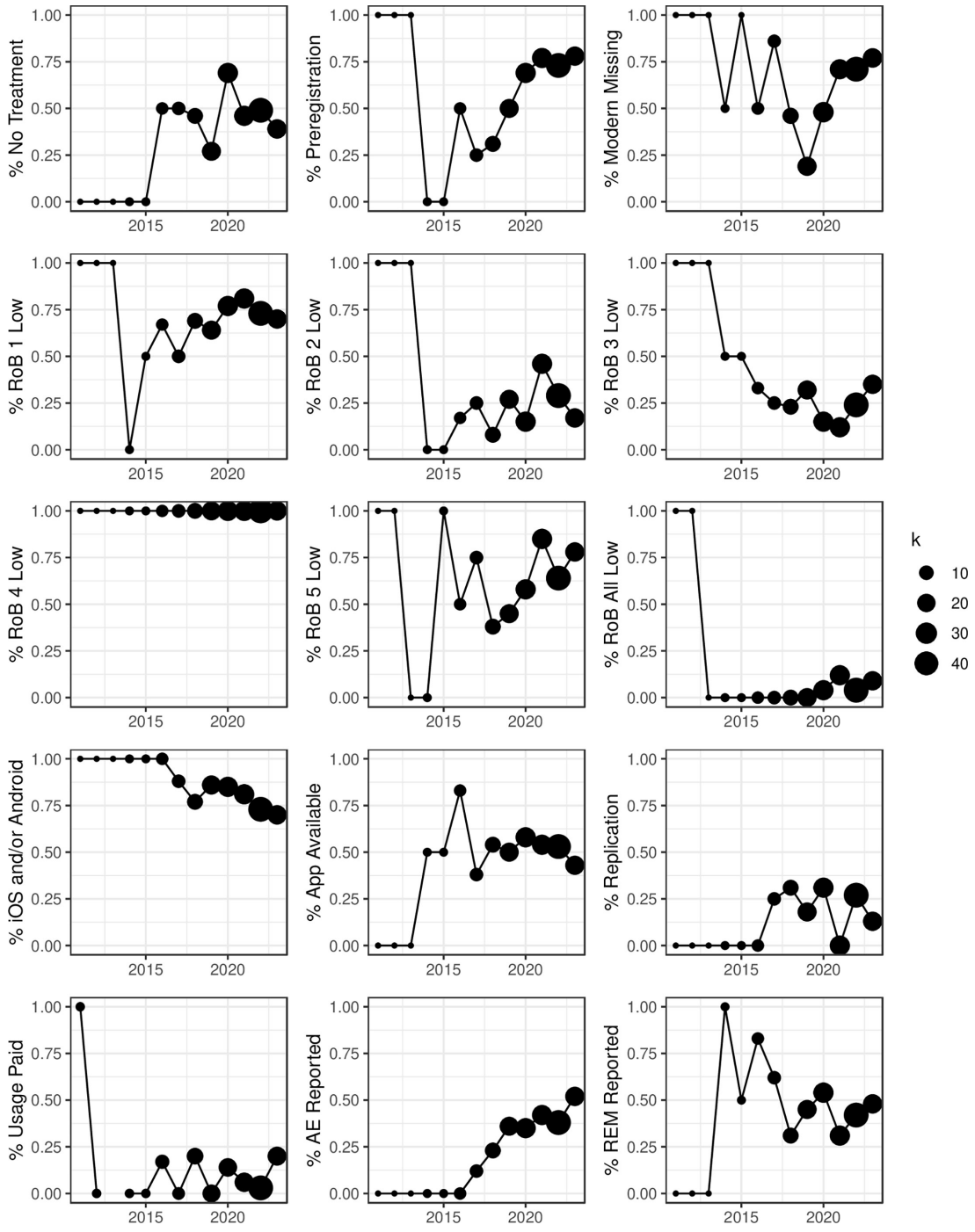

Figure 2 Change in dichotomous study quality features over time. RoB items 1 through 5: sequence generation, allocation concealment, blinding of personnel, blinding of outcome/self-report and use of intention-to-treat (ITT) analysis, respectively. % No Treatment indicates the percentage of trials that used a no treatment control condition, comprising a waitlist or assessment only condition. k indicates the number of studies. AE, adverse event; REM, racial/ethnic minority participants; RoB, Risk of Bias.

Table 3 Sensitivity analysis with outliers removed

Variable	Estimate (95% CI)	SE	t	P value
Sample size	24.78 (−4.91, 54.47)	15.15	1.64	0.104
Attrition rate	−0.83 (−2.01, 0.35)	0.60	−1.4	0.163
Follow-up length	0.14 (−0.02, 0.3)	0.08	1.83	0.069

P values from regression models.
t indicates test statistic for linear regression models. Models rerun only for variables that had outliers. Follow-up length indicates postrandomisation follow-up length in months.

RCTs,[9] appropriately handling missing data is important for drawing reliable conclusions. Ideally, future studies will manage to decrease study attrition (a feature that has not shown improvement over time) as well as use modern methods for exploring the potential influence of data missing not at random (eg, pattern mixture modelling).[35]

Notwithstanding positive trends in preregistration, reporting of adverse events, use of modern missing data methods and use of ITT analyses, other methodological features that characterise high-quality trials have not shown evidence of increased

adoption, despite ongoing calls for their implementation. For example, there have been repeated efforts to highlight persistent issues related to the (1) excessive use of waitlist controls, (2) underpowered pilot nature of many available trials, (3) high attrition rates that compromise interpretation of study findings and (4) lack of longer follow-up assessments.[1 11 17] These shortcomings are likely influenced by a combination of factors. One contributing factor may be increased publication pressure, where scientists are incentivised to prioritise quantity over quality. This environment can deincentivise conducting larger, more time-intensive trials that better meet the standards of high-quality research. Another factor may relate to budgetary and funding constraints; implementing features that characterise higher quality trials requires additional resources, including money and personnel (eg, more participants, resources for active controls, participant reimbursement), which may not always be feasible, especially for smaller research teams or those operating in underfunded environments. Likewise, the budget and resources necessary to build and maintain functional digital mental health systems may be greater than most investigators realise.[36] Alternatively, the heavy focus on developing and trialling new apps (rather than working with existing apps that can be customised) may encourage more exploratory pilot testing, as is typically recommended in established frameworks that set out the phases of intervention evaluation.[37] If this is the case, it is possible that those apps that demonstrate feasibility will be subject to more rigorous evaluation in large-scale, confirmatory trials in future.

Another concerning finding was the lack of increased replication efforts. With the exception of a few commercially available apps like Headspace and PTSD Coach, few identified apps have been tested for efficacy across multiple settings and participant groups. This is likely because many apps tested in clinical trials are developed for research purposes and are not commercially available for broader use or independent validation from other research teams. Furthermore, the field and funders may prioritise the creation of entirely new apps over refining and augmenting existing ones with a promising evidence base. Establishing publication and funding standards to prioritise replication is essential to advancing the field. Replication efforts would strengthen trust in these tools and help identify the specific conditions under which apps are most safe and effective, paving the way for a more personalised approach to mental healthcare.

The current findings must be interpreted within the context of their limitations. First, findings regarding the lack of improvement in methodological rigour in clinical trials of apps for depression and anxiety cannot be generalised to other psychiatric conditions. There is some evidence that trials on digital health technologies in patients with schizophrenia are conducted with greater rigour, including more oversight and risk assessments, given heightened concerns for potential adverse events.[15] It would be useful to investigate whether the methodological quality of digital health trials for other psychiatric conditions has improved over time. Second, our analyses were based solely on the information provided in the published report, so it is possible that certain design features may have been implemented (eg, participant payment, iOS and/or Android compatibility, etc) but were not explicitly reported. Third, although we analysed a large number of trial quality features, there are potentially many more relevant design features that we did not consider, which could be increasing over time (eg, Consolidated Standards of Reporting Trials compliance, conflict of interest declarations, adherence monitoring efforts, intervention fidelity measures, etc). Fourth, our findings reflect trends in trial quality within this growing literature and do not speak to the quality of individual trials. It is important to note that, despite the overall trends, the increased publication of trials in this field has also given rise to a number of high-quality individual studies. Arguably, these high-quality trials can be examined individually and in aggregate (eg, via meta-analysis) to make reliable inferences. Fifth, we used the publication year as a proxy for trial timing, which may not always reflect the actual year the trial was initiated. However, this decision was necessary to ensure consistency across studies, as many trials did not report their start date or have a preregistered protocol.

Another limitation of our study is that we did not include a measure of engagement as an indicator of trial quality, instead opting for study attrition (a related but distinct construct[38]). This decision was driven by the inconsistent reporting and definition of engagement across trials, making meaningful between-study comparisons difficult and degree of missing data significant. The lack of standardised reporting on engagement is a well-recognised issue in the field and continues to hinder progress.[39 40] New methods like digital phenotyping may help provide new objective data on screen use. Recent pilot research[41] suggests that the correlation between engagement as measured by digital phenotyping and as measured by self-reported scales may be minimal, suggesting each method is capturing unique data. New unified frameworks to assess engagement will also be critical.[42] Beyond measurement, it is equally important for future research to focus on improving engagement with digital interventions, as greater engagement may lead to stronger clinical benefits.[43] Emerging design strategies such as digital navigator coaching,[44] gamification principles,[45] tailored email prompts[46] and just-in-time intervention strategies[47] show promise in increasing engagement but require rigorous evaluation in large-scale randomised trials.

CONCLUSION

In conclusion, our analysis of 176 trials of apps targeting depression and anxiety revealed minimal evidence of increased methodological rigour over time. While promising increases in preregistration, adverse event reporting and potentially in handling of missing data were observed, other key indicators of trial quality showed no significant improvement over time. This research highlights key areas for improvement which may encourage the next generation of research to overcome the methodological limitations identified here. Doing so can expedite the safe and ethical integration and dissemination of digital health tools into clinical practice and broader society, ensuring they effectively reach and benefit those who need mental health support.

X Qiang Xie @QiangXie12, John Torous @JohnTorousMD and Simon B Goldberg @SGoldbergPhD

Contributors Conceptualisation: JL, JT, SS, SBG. Data curation: QX, CS, SBG. Data analysis: SBG. Writing: JL, JT, SBG. Editing: JT, QX, CS, JT, SS, SBG. JL is responsible for the overall content as guarantor. JL accepts full responsibility for the finished work and/or the conduct of the study, had access to the data and controlled the decision to publish.

Funding JL holds a National Health and Medical Research Council Investigator Grant (APP1196948). CS was supported by the National Institute of Mental Health (Award No MH018931). SS was supported by the National Center for Complementary and Integrative Health (NCCIH) (Award No K23AT011173 and R24AT012845). SBG was supported by the NCCIH (Award No K23AT010879 and R24AT012845).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Analysis codes are available (https://uwmadison.box.com/s/w16jhkaf871qdnincdamhyr6gqv3dx65).

ORCID iDs
Jake Linardon http://orcid.org/0000-0003-4475-7139
Qiang Xie http://orcid.org/0000-0003-3968-9629
Caroline Swords http://orcid.org/0009-0005-6373-4997
John Torous http://orcid.org/0000-0002-5362-7937
Shufang Sun http://orcid.org/0000-0001-5215-128X
Simon B Goldberg http://orcid.org/0000-0002-6888-0126

REFERENCES

1 Torous J, Bucci S, Bell IH, *et al*. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 2021;20:318–35.
2 Linardon J, Torous J, Firth J, *et al*. Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. A meta-analysis of 176 randomized controlled trials. *World Psychiatry* 2024;23:139–49.
3 Fuhrmann LM, Weisel KK, Harrer M, *et al*. Additive effects of adjunctive app-based interventions for mental disorders - A systematic review and meta-analysis of randomised controlled trials. *Internet Interv* 2024;35:100703.
4 Bae H, Shin H, Ji H-G, *et al*. App-Based Interventions for Moderate to Severe Depression: A Systematic Review and Meta-Analysis. *JAMA Netw Open* 2023;6:e2344120.
5 Firth J, Torous J, Nicholas J, *et al*. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry* 2017;16:287–98.
6 Firth J, Torous J, Nicholas J, *et al*. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *J Affect Disord* 2017;218:15–22.
7 Donker T, Petrie K, Proudfoot J, *et al*. n.d. Smartphones for Smarter Delivery of Mental Health Programs: A Systematic Review. *J Med Internet Res*15:e247.
8 Linardon JCA. Can Acceptance, Mindfulness, and Self-Compassion Be Learned by Smartphone Apps? A Systematic and Meta-Analytic Review of Randomized Controlled Trials. *Behav Ther* 2020;51:646–58.
9 Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. *J Consult Clin Psychol* 2020;88:1–13.
10 Torous J, Lipschitz J, Ng M, *et al*. Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. *J Affect Disord* 2020;263:413–9.
11 Linardon J, Cuijpers P, Carlbring P, *et al*. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. *World Psychiatry* 2019;18:325–36.
12 Cuijpers P, van Straten A, Bohlmeijer E, *et al*. The effects of psychotherapy for adult depression are overestimated: a meta-analysis of study quality and effect size. *Psychol Med* 2010;40:211–23.
13 Cuijpers P, Miguel C, Harrer M, *et al*. The overestimation of the effect sizes of psychotherapies for depression in waitlist controlled trials: a meta-analytic comparison with usual care controlled trials. *Epidemiol Psychiatr Sci* 2024;33:e56.
14 Michopoulos I, Furukawa TA, Noma H, *et al*. Different control conditions can produce different effect estimates in psychotherapy trials for depression. *J Clin Epidemiol* 2021;132:59–70.
15 Linardon J, Fuller-Tyszkiewicz M, Firth J, *et al*. Systematic review and meta-analysis of adverse events in clinical trials of mental health apps. *NPJ Digit Med* 2024;7:363.
16 Linardon J, Firth J, Torous J, *et al*. Efficacy of mental health smartphone apps on stress levels: a meta-analysis of randomised controlled trials. *Health Psychol Rev* 2024;18:839–52.
17 Goldberg SB, Sun S, Carlbring P, *et al*. Selecting and describing control conditions in mobile health randomized controlled trials: a proposed typology. *NPJ Digit Med* 2023;6.
18 Higgins J, Green S. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2011.
19 Eysenbach G. The Law of Attrition. *J Med Internet Res* 2005;7:e11.
20 Lipsey MW, Wilson D. *Practical meta-analysis*. Sage Publications, 2001.
21 P Simmons J, D. Nelson L, Simonsohn U. Pre-registration: Why and How. *J Consum Psychol* 2021;31:151–62.
22 Ramos G, Chavira DA. Use of Technology to Provide Mental Health Care for Racial and Ethnic Minorities: Evidence, Promise, and Challenges. *Cogn Behav Pract* 2022;29:15–40.
23 O'Connor S, Hanlon P, O'Donnell CA, *et al*. Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC Med Inform Decis Mak* 2016;16:120:120:.
24 Goldberg SB, Tucker RP, Greene PA, *et al*. Is mindfulness research methodology improving over time? A systematic review. *PLoS ONE* 2017;12:e0187298.
25 Champely S, Ekstrom C, Dalgaard P, *et al*. Basic functions for power analysis. R Package Version 2018; 2018.2.
26 R: a language and environment for statistical computing. Vienna, Austria R Foundation for Statistical Computing; 2024.
27 Kumar A, Ross JS, Patel NA, *et al*. Studies of prescription digital therapeutics often lack rigor and inclusivity: study examines prescription digital therapeutics standards. *Health Aff* 2023;42:1559–67.
28 Prentice C, Peven K, Zhaunova L, *et al*. Methods for evaluating the efficacy and effectiveness of direct-to-consumer mobile health apps: a scoping review. *BMC Digit Health* 2024;2:31.
29 Nosek BA, Lindsay DS. Preregistration becoming the norm in psychological science. APS observer 2018.:31.
30 Hardwicke TE, Wagenmakers E-J. Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nat Hum Behav* 2023;7:15–26.
31 Spitzer L, Mueller S. Registered report: Survey on attitudes and experiences regarding preregistration in psychological research. *PLoS One* 2023;18:e0281086.
32 Maxwell SE, Lau MY, Howard GS. Is psychology suffering from a replication crisis? What does 'failure to replicate'really mean? *Am Psychol* 2015;70:487–98.
33 Allan S, Ward T, Eisner E, *et al*. Adverse Events Reporting in Digital Interventions Evaluations for Psychosis: A Systematic Literature Search and Individual Level Content Analysis of Adverse Event Reports. *Schizophr Bull* 2024;50:1436–55.
34 Bradstreet S, Allan S, Gumley A. Adverse event monitoring in mHealth for psychosis interventions provides an important opportunity for learning. *J Ment Health* 2019;28:461–6.
35 Goldberg SB, Bolt DM, Davidson RJ. Data Missing Not at Random in Mobile Health Research: Assessment of the Problem and a Case for Sensitivity Analyses. *J Med Internet Res* 2021;23:e26749.
36 Owen JE, Kuhn E, Jamison AL. The ncptsd model for digital mental health: a public sector approach to development, evaluation, implementation, and optimization of resources for helping trauma survivors. *PsyArXiv* [Preprint].
37 Skivington K, Matthews L, Simpson SA, *et al*. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ* 2021;374:n2061.
38 Nwosu A, Boardman S, Husain MM, *et al*. Digital therapeutics for mental health: Is attrition the Achilles heel? *Front Psychiatry* 2022;13:900615.
39 Boucher EM, Raiker JS. n.d. Engagement and retention in digital mental health interventions: a narrative review. *BMC Digit Health*2:52.
40 Ng MM, Firth J, Minen M, *et al*. User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity. *Psychiatr Serv* 2019;70:538–44.
41 Dwyer B, Flathers M, Burns J, *et al*. Assessing Digital Phenotyping for App Recommendations and Sustained Engagement: Cohort Study. *JMIR Form Res* 2024;8:e62725.
42 Perski O, Blandford A, West R, *et al*. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Transl Behav Med* 2017;7:254–67.
43 Donkin L, Christensen H, Naismith SL, *et al*. A systematic review of the impact of adherence on the effectiveness of e-therapies. *J Med Internet Res* 2011;13:e52.
44 Perret S, Alon N, Carpenter-Song E, *et al*. Standardising the role of a digital navigator in behavioural health: a systematic review. *Lancet Digit Health* 2023;5:e925–32.
45 Looyestyn J, Kernot J, Boshoff K, *et al*. Does gamification increase engagement with online programs? A systematic review. *PLoS One* 2017;12:e0173403.
46 Agachi E, Bijmolt THA, van Ittersum K, *et al*. The Effect of Periodic Email Prompts on Participant Engagement With a Behavior Change mHealth App: Longitudinal Study. *JMIR Mhealth Uhealth* 2023;11:e43033.
47 Nahum-Shani I, Smith SN, Spring BJ, *et al*. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Ann Behav Med* 2018;52:446–62.