Research Synthesis Methods (2026), **00**: 1–19
doi:10.1017/rsm.2025.10064

Research Synthesis Methods

RESEARCH ARTICLE

Evaluation of the replicability of systematic reviews with meta-analyses of the effects of health interventions

Daniel G. Hamilton[1], Joanne E. McKenzie[1], Phi-Yen Nguyen[1], Melissa L. Rethlefsen[2], Steve McDonald[1], Sue E. Brennan[1], Fiona M. Fidler[3], Julian P. T. Higgins[4], Raju Kanukula[5], Sathya Karunananthan[6], Lara J. Maxwell[7], David Moher[8], Shinichi Nakagawa[9], David Nunan[10], Peter Tugwell[11], Vivian A. Welch[12] and Matthew J. Page[1]

[1]School of Public Health and Preventive Medicine, Monash University, Australia
[2]Health Sciences Library and Informatics Center, University of New Mexico Health Sciences Center, USA
[3]School of History and Philosophy of Sciences, The University of Melbourne, Australia
[4]Population Health Sciences, University of Bristol, UK
[5]School of Architecture, Design and Planning, University of Sydney, Sydney, NSW, Australia
[6]Interdisciplinary School of Health Sciences, University of Ottawa Faculty of Health Sciences, Canada
[7]Faculty of Medicine, University of Ottawa, Canada
[8]Ottawa Methods Centre, Ottawa Hospital Research Institute, Canada
[9]Department of Biological Sciences, University of Alberta, Canada
[10]Nuffield Department of Primary Care, University of Oxford, UK
[11]Department of Medicine, University of Ottawa, Canada
[12]School of Epidemiology and Public Health, University of Ottawa, Canada

Corresponding author: Matthew J. Page; Email: matthew.page@monash.edu.

Received: 18 July 2025; **Revised:** 21 October 2025; **Accepted:** 26 November 2025

Keywords: literature search; meta-analysis; meta-research; replication; reproducibility; systematic review

Abstract

Systematic reviews are often characterized as being inherently replicable, but several studies have challenged this claim. The objective of the study was to investigate the variation in results following independent replication of literature searches and meta-analyses of systematic reviews. We included 10 systematic reviews of the effects of health interventions published in November 2020. Two information specialists repeated the original database search strategies. Two experienced review authors screened full-text articles, extracted data, and calculated the results for the first reported meta-analysis. All replicators were initially blinded to the results of the original review. A meta-analysis was considered not 'fully replicable' if the original and replicated summary estimate or confidence interval width differed by more than 10%, and meaningfully different if there was a difference in the direction or statistical significance. The difference between the number of records retrieved by the original reviewers and the information specialists exceeded 10% in 25/43 (58%) searches for the first replicator and 21/43 (49%) searches for the second. Eight meta-analyses (80%, 95% CI: 49–96) were initially classified as not fully replicable. After screening and data discrepancies were addressed, the number of meta-analyses classified as not fully replicable decreased to five (50%, 95% CI: 24–76). Differences were classified as meaningful in one blinded replication (10%, 95% CI: 1–40) and none of the unblinded replications (0%, 95% CI: 0–28). The results of systematic review

This article was awarded Open Data and Open Materials badges for transparent practices. See the Data availability statement for details.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

processes were not always consistent when their reported methods were repeated. However, these inconsistencies seldom affected summary estimates from meta-analyses in a meaningful way.

Highlights
What is already known?

- Systematic reviews are often characterized as being inherently replicable; however, several studies have challenged this claim.
- Few studies have examined where and why inconsistencies arise, and what their impact is, when replicating multiple systematic review processes.

What is new?

- Replication of published systematic review processes (database searches, full-text screening, data extraction, and meta-analysis) frequently produced results that were inconsistent with the original review.
- Following correction of replicator errors, the main drivers of variation in the results were incomplete reporting (e.g., unclear search methods, study eligibility criteria, and methods for selecting study results) and reviewer data extraction errors.
- Differences between the original reviewer's and replicators' summary estimates and confidence intervals were seldom meaningful.

1. Introduction

The findings of systematic reviews have the potential to shape clinical practice and future research efforts in health and medicine in a profound way. However, despite their potential, there is growing unease among researchers about the reliability of systematic reviews and other published research.[1–5] This crisis of confidence is commonly referred to as the 'replication crisis', or the 'reproducibility crisis', and is often attributed to low replication success rates in large-scale replications of high-profile experiments across multiple scientific fields,[6–10] including medicine.[11–15]

Systematic reviews with meta-analyses are often characterized as being inherently replicable.[16,17] However, some evidence has brought this claim into question. For example, Rethlefsen et al.[18] found that when repeating 88 database searches from 39 systematic reviews, only 41 (47%) searches retrieved a similar number of records (i.e., differed by less than 10% to the original). Similarly discordant results have been observed when replicating other important review processes, such as article screening,[19,20] data extraction,[21–26] and risk of bias assessments.[27,28] Many of these studies also evaluated the impact of discrepancies in these processes on the findings of reported meta-analyses.[19,21–26]

Only a limited number of studies have attempted to replicate multiple review processes (e.g., search, screening, and data collection) within systematic reviews of health interventions.[29,30] One study by Pieper et al.[29] repeated the literature searches of a systematic review of the effects of omega-3 polyunsaturated fatty acid intake in patients with chronic kidney disease, then screened, extracted data, and assessed the risk of bias of a random 25% of the records retrieved. The authors found that one of three database searches differed by more than 10% from the original, observed eight screening discrepancies out of the 214 records screened (four studies incorrectly included and four incorrectly excluded), but no issues with data extraction. Another relevant study by Ford et al.[30] replicated article screening, data extraction, and calculations for 16 meta-analyses from eight systematic reviews that examined the effects of pharmacological interventions for irritable bowel syndrome. The authors found six reviews had missed 22 eligible randomized trials and incorrectly included 35 trials. They also identified 80 data extraction errors across all reviews. These discrepancies and errors resulted in at least a 10% change to the summary estimate in 5 (31%) of 16 meta-analyses, and a reversal of the statistical significance in 4 (25%).

Investigation of the replicability of multiple systematic review processes is important for determining where and why inconsistencies arise, and what their impact is. However, previous studies of this type are limited by their focus on only one clinical topic and have gaps in the processes replicated

(e.g., focus on searching and screening but not meta-analysis calculations). To address this, we aimed to evaluate the extent and source of variation in results when teams of information specialists and systematic reviewers independently replicated the database searches and meta-analyses (including the full-text screening, data extraction, and calculation steps) of systematic reviews assessing the effects of health interventions.

2. Methods

This study was the final in a series of studies comprising the REPRISE (REProducibility and Replicability In Syntheses of Evidence) project.[31–36] An abbreviated version of the methods is provided here. The complete methods, including the list of deviations from the original study protocol[31] (Supplementary Table S1), are reported in the Supplementary Materials.

2.1. Sampling frame

Systematic reviews selected for this study were sampled from those identified in the first study of the REPRISE project, which assessed the reporting completeness of 300 systematic reviews with meta-analyses published in November 2020.[33] For this study, we selected for replication the reviews from the first REPRISE study which met the following inclusion criteria: (1) reported the full Boolean search strategy for each database searched, (2) reported the number of records retrieved by each search, (3) provided information on which studies were included in the review and in the first reported (i.e., 'index') pairwise meta-analysis, (4) reported summary statistics (e.g., means and standard deviations) or an effect estimate and associated measure of precision (e.g., standardized mean difference and 95% confidence interval) for each study in the index meta-analysis, and (5) did not include more than 10 studies in the index meta-analysis. For this study, we also excluded reviews which had been retracted and restricted our attention to meta-analyses undertaken in the frequentist framework.

2.2. Replication personnel

Two replications were completed in this study: (1) replications of the database literature searches and (2) replications of the index meta-analysis (including the full-text screening, data extraction, and calculation steps). We did not replicate other review processes (e.g., title and abstract screening, risk of bias assessments), but used the results of the original reviewers when required. For example, when reviewers excluded studies based on risk of bias, we used the reviewers' risk of bias assessment to determine study eligibility. A flow diagram depicting the procedure followed for each replication is shown in Figure 1. One author (MJP) with expertise in performing and evaluating the conduct and reporting of systematic reviews was responsible for the coordination of all replication activities. Replications of the literature searches were performed by two information specialists (MLR and SM), each with over 25 years' experience developing search strategies for systematic reviews, clinical guidelines, and other evidence syntheses. Replication of the meta-analyses were performed by two researchers with experience screening articles, collecting and preparing outcome data, and performing meta-analyses for more than three systematic reviews within the past 3 years (DGH and P-YN). None of the four replicators were involved in the conduct of any of the 10 systematic reviews selected for replication, any of the studies included within them, nor had performed a systematic review that addressed the same or similar questions. All replicators were instructed not to seek out any information on the 10 reviews (e.g., a PROSPERO entry, a review protocol, the review report) prior to submission of the results of their initial replications (i.e., were blinded to the results of the reviews). The blinded replications were conducted to provide a test of the replicability of the methods as documented in the review report. We then followed this up with unblinded replications which allowed us to test the replicability of both the methods and results, as well as investigate the causes of observed discrepancies, including assessing the impact of errors introduced by the replication team.

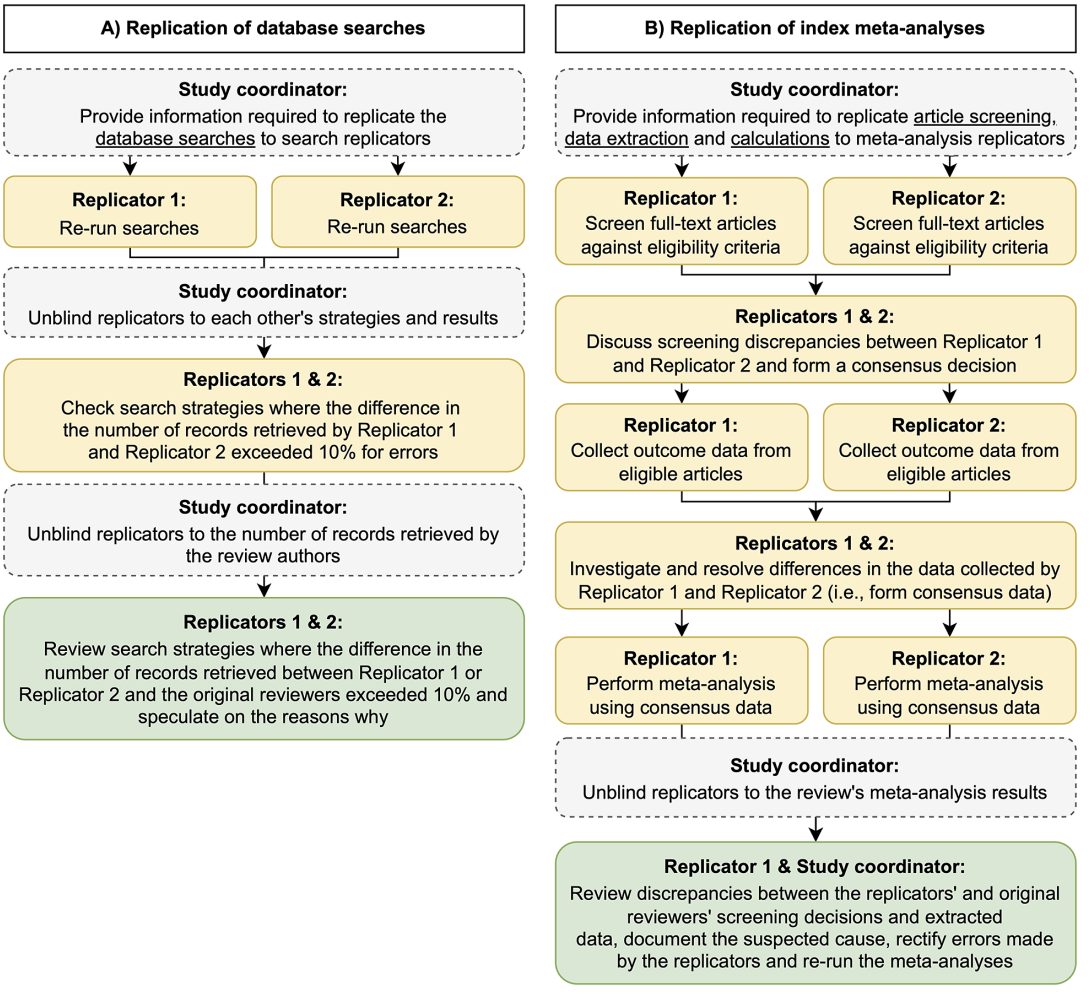

Figure 1. *Flow diagram outlining the process followed to replicate the selected review's (A) database searches and (B) index meta-analysis. Note that the yellow boxes refer to activities performed by replicators while blinded, and green boxes refer to activities occurring after unblinding.*

2.3. Extracted information for replicators

The study coordinator (MJP) assembled all information required to perform and evaluate the findings of the replications, which was subsequently checked for accuracy by one of the replicators (DGH) after they were unblinded. Exactly what information was provided to the replicators in the blinded phase and unblinded phase is reported in Table 1 and the project's OSF page.[37]

2.4. Replication methods

We recognize that the terminology for 'replication' is not standardized within and across disciplines.[38,39] In this study, we adopted the non-procedural definitions of replication advocated by Nosek and Errington[40] and Machery.[41] That is, replicators did not need to follow every single step exactly as reported in the original systematic review, but they were constrained by the original review question and were instructed to avoid making changes to the methods and concepts that might be reasonably judged to violate an attempt to answer that question.

Table 1. *Information extracted from selected systematic review reports and associated supplementary materials to guide replicators and compare results.*

Review process	Data item extracted	Data given to blinded replicators	Data given to unblinded replicators
Database searches	Line-by-line search strategy for each database consulted	Yes	Yes
	The platform used	Yes	Yes
	Date ranges when databases were last searched	Yes	Yes
	The number of records yielded by each search	No	Yes
Full-text article screening	The question addressed by the index meta-analysis	Yes	Yes
	Eligible participants, interventions, comparators, outcomes, study designs, languages of publication, report types for the index meta-analysis	Yes	Yes
	Reports of studies that were included in the index meta-analysis	Yes	Yes
	Reports of studies which were not included in the index meta-analysis (i.e., studies that were included in other analyses or excluded by the reviewers after full-text screening)	Yes	Yes
	Original reviewers' screening decisions for each study	No	Yes
Data extraction	Decision rules reported by the original systematic reviewers regarding which results to select from studies (e.g., which measurement scale, time point, or analysis sample was selected)	Yes	Yes
	Results of each study included in the index meta-analysis (i.e., summary statistics and effect estimate with a measure of precision)	No	Yes
Meta-analysis	The effect measure(s) used (e.g., risk ratio)	Yes	Yes
	Meta-analysis method used (e.g., inverse variance)	Yes	Yes
	Meta-analysis model used (e.g., random effects)	Yes	Yes
	Between-study variance estimator used (e.g., DerSimonian-Laird)	Yes	Yes
	Analysis software used (e.g., RevMan)	Yes	Yes
	Results of the meta-analysis (i.e., summary estimate and measure of precision) and associated heterogeneity statistics (e.g., chi^2 value, I^2, tau^2)	No	Yes

2.4.1. Database search replications

For the literature search replications, replicators were asked to re-run all database search strategies using the same date limits as reported in the original review. When able, searches were run using the same platform and exact same search string and date limits as reported in the original review. Where possible, searches were also run so as not to retrieve records that were published within the relevant date range

but indexed in the database after the original search was run. The two replicators performed all literature searches independently and documented which original search strategies they attempted to re-run, any assumptions made, and the number of records that each search yielded. Following submission of their search results, the replicators were unblinded to each other's search strategies and results, then given the opportunity to amend their searches if any errors were detected. Next, the replicators were unblinded to the number of records retrieved by the original reviewers, and where the replicators' search results differed from the original reviewers' by more than 10%, were asked to speculate on the likely causes for the discrepancy.

2.4.2. Index meta-analysis replications

For the meta-analysis replications, replicators were given the full-text reports of all articles included in the review (not just the index meta-analysis), as well as those cited as being excluded from the review, then asked to screen all reports against the eligibility criteria for the index meta-analysis. Replicators screened each article independently, recording their decision ('include', 'exclude', or 'unsure') and rationale, then met afterwards to discuss and resolve any discrepancies or uncertainties and form a consensus decision. In some situations, further information concerning the review's eligibility criteria was requested from the study coordinator to aid decision making (e.g., under what cut-off were doses considered 'low dose' for vitamin D supplementation), which was investigated and provided to the replicators when available in the systematic review report.

The data required for the meta-analysis were then independently extracted by both replicators from the body text, tables, or plots in the study reports, or Supplementary Materials of the studies deemed eligible. WebPlotDigitizer (v4) was used to extract data from plots. The algorithm proposed by Guyot et al.[42] was executed in R to derive approximations of the hazard ratios and their standard error from data extracted from Kaplan–Meier curves. If the required data were not reported, one author (DGH) attempted to contact the primary study authors (i.e., authors of the study included in the systematic review) to obtain the required information. If the replicators were unable to source required data for an eligible study (including from the study authors), it was classified as having missing results.

Any discrepancies between numerical data extracted from text and tables in the study report by the replicators were discussed and resolved. While, based on prior calibration testing, differences between replicators greater than 0.5 units for data extracted from plots where the length of the axis of interest ranged between 25 and 100 units were investigated and investigated for possible errors (e.g., axis alignment errors, time point selection errors). Additionally, percentage differences between hazard ratios, upper and lower confidence interval bounds, and standard errors derived from Kaplan–Meier curves greater than 5% were investigated. Once all data derived from plots were finalized (including the estimation of hazard ratios), the arithmetic mean of the two values were calculated and used for the meta-analysis.

After consensus was reached on the extracted data, replicators independently performed the calculations to generate meta-analysis results. Given, we did not have access to all software used by the reviewers, the replicators were instructed to use software they were most familiar with, namely, R (v4.3.1) using the meta package (v7.0.0) (DGH) and Stata 17 using the in-built meta command (P-YN). The results from both replicators are reported.

Following submission of the results of the article screening, data extraction, and calculation steps, the replicators were then unblinded, and discrepancies between the replicators and original reviewers in each step were identified and jointly investigated by two authors (DGH and MJP) to try to determine the cause. Where further data needed to be extracted (e.g., where an exclude decision was updated to include), this was performed by one author (DGH) and checked for accuracy by another (MJP). When the causes of discrepancies could not be confidently identified, one author (DGH) contacted the original reviewers to seek clarification. Once all discrepancies were reviewed and addressed, the results of the replications of each step were updated. These updated findings are referred to as the 'unblinded' results in Section 3.

2.5. Outcomes

The study's outcomes are reported in Table 2. For the literature searches, we calculated the number of times the percentage difference between the number of records retrieved by the original reviewers and replicators was less than or equal to 10%. The cut-off of 10% was chosen to allow for expected variation in the number of records retrieved due to database changes, as well as comparability with previous related research.[18] Similarly, meta-analyses were classified as 'fully replicable' if both the percentage difference between the original reviewers' and replicators' summary estimates *and* confidence interval widths did not exceed 10%. Furthermore, we classified replicated summary estimates as 'meaningfully' different if they were opposite in direction to the original, or had a difference in the statistical significance, or both, given in practice, these differences might lead reviewers to reach different conclusions. Percentage difference for each outcome described earlier was defined as the difference between the replicator's and the original reviewer's results divided by the original reviewer's result. For percentage differences between summary estimates, ratio measures were log-transformed (natural logarithm) prior to calculation.

2.6. Statistical analysis

Categorical data and continuous data are presented descriptively using (i) counts and proportions and (ii) medians, interquartile ranges (IQR), and ranges, respectively. All analyses were performed in R (v4.3.1). Confidence intervals around proportions for binary outcomes were calculated using the modified Wilson interval method as proposed by Brown et al.[43] using the DescTools package (v0.99.55). Unweighted Kappa statistics and 95% confidence intervals were calculated using the psych package (v2.4.6.26). Figures were generated via draw.io (Figure 1), R using the ggplot2 (v3.5.1) package (Figure 2), and Microsoft Excel (Figure 3A and B). All data, code, and materials needed to reproduce the findings of this research have been made publicly available on the Open Science Framework (https://osf.io/v3mnh/).[37]

2.7. Sensitivity analysis

We performed two post hoc sensitivity analyses to investigate the robustness of two replicated meta-analyses. The first sensitivity analysis removed hazard ratios estimated using Guyot and colleague's[42] algorithm from one meta-analysis that did not report the total number of events and the numbers at risk (other than at time zero) alongside the reported Kaplan–Meier curves. This was due to warnings provided the developers of the algorithm that its accuracy is likely to be poor (i.e., associated with a 10-fold increase in the estimated mean absolute error) when the aforementioned information has not been provided. The second sensitivity analysis included a result from a study eligible for another meta-analysis where the effect estimate was calculated using means and standard deviations estimated from reported medians and ranges via the application of normal and non-normal data transformation methods[44,45] via the estmeansd package (v1.0.1) in R.

3. Results

3.1. Characteristics of the included reviews

Ten of the 300 systematic reviews from the first REPRISE study met the inclusion criteria for this study.[46–55] The 10 reviews examined different interventions, including pharmaceuticals (e.g., methadone), behavioural interventions (e.g., sleep deprivation), surgical techniques (e.g., lymphadenectomy), and devices (e.g., laser therapy) (Table 3). Reviewers frequently limited eligible study designs to randomized trials (N = 7/10, 70%). Half registered the review's details or disseminated a protocol outlining the planned methods (N = 5/10, 50%). The software used most often to perform meta-analyses was RevMan (N = 5/10, 50%).

Table 2. *Study outcomes and analysis methods.*

Replication step and study outcome	Analysis method
Replication of database searches	
Number of database search strategies that could be re-run by the two replicators	We calculated the frequency and percentage (with 95% CIs) of search strategies that could be re-run
Agreement between the number of search records retrieved by each replicator and the original reviewers	We calculated the frequency and percentage (with 95% CIs) of search strategies for which the percentage difference* between the number of records retrieved by the original reviewers and replicators was less than or equal to 10%
Agreement between the number of search records retrieved by each replicator.	We calculated the frequency and percentage (with 95% CIs) of search strategies for which the percentage difference* between the number of records retrieved by the replicators was less than or equal to 10%.
Replication of meta-analyses: full-text screening	
Agreement between the replicators' consensus screening decisions and the original reviewers' screening decisions	We calculated for each systematic review the percentage agreement (with 95% CIs) in screening decisions and the unweighted Kappa statistic (with 95% CIs)
Reasons underpinning discrepancies between screening decisions	We inductively classified reasons for discrepancies into categories
Replication of meta-analyses: data extraction	
Number of studies where the replicators' consensus study effect estimate or its precision (collected from reports or calculated from reported summary statistics) differed to the effect estimate or its precision included in the original meta-analysis due to reasons other than software-related reasons	We calculated the frequency and percentage (with 95% CIs) of studies for which a study effect estimate *or* its precision differed in a non-trivial manner
Reasons underpinning discrepancies between study effect estimates	We inductively classified reasons for discrepancies into categories
Replication of meta-analyses: calculations	
Number of meta-analyses classified as 'fully replicable' (that is the percentage difference*[a] between the original reviewers' and replicators' summary estimates and confidence interval widths did not exceed 10%)	We calculated the frequency and percentage (with 95% CIs) of meta-analyses classified as fully replicable
Number of meta-analyses classified as 'meaningfully different' (that is replicated summary estimates were opposite in direction to the original or had a difference in the statistical significance (as defined by the original reviewers), or both	We calculated the frequency and percentage (with 95% CIs) of meta-analyses classified as meaningfully different

Note: Percentage difference for each outcome described earlier was defined as the difference between the replicators' and the original reviewer's results divided by the original reviewer's result. CI: confidence interval.

[a] For percentage differences between summary estimates, ratio measures were log-transformed (natural logarithm) prior to calculation.

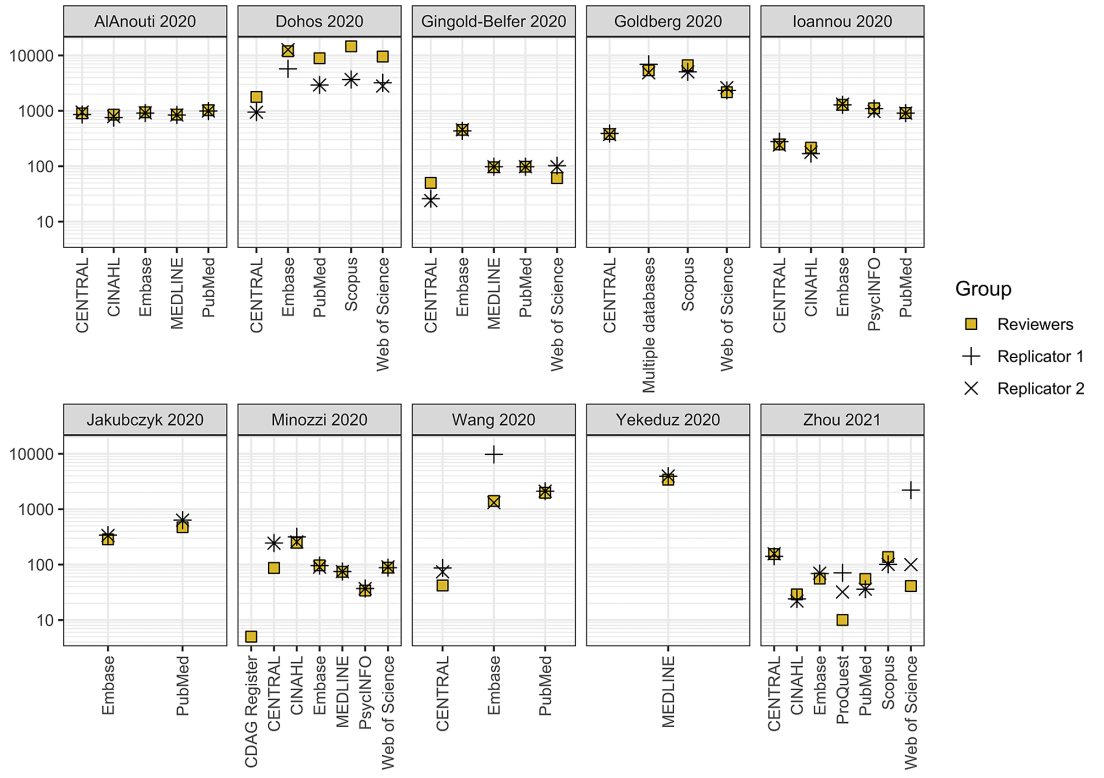

Figure 2. *Scatter plot of the number of records retrieved by the original reviewers (square symbols) and the first (plus signs) and second (cross signs) replicators by systematic review. Note: These numbers for Replicators 1 and 2 refer to the records retrieved following correction of initial errors they had made. The y-axis is presented on a logarithmic scale (base 10).*

3.2. Replication of literature searches

Across the 10 reviews, 46 bibliographic database search strategies were reported in sufficient detail to enable replication. A median of five databases (IQR: 3.5–5.8, range: 1–7) were searched per review, with the most frequently searched databases being CENTRAL, Embase, and PubMed (all $N = 8/10$, 80%). Both replicators were able to re-run all but one reported search strategy ($N = 45/46$, 98%, 95% CI: 89–100). The only search that could not be re-run was of the Cochrane Drugs and Alcohol Group's Specialized Register as it was inaccessible to the public. The number of records retrieved by the original reviewers and replicators by review and database are shown in Figure 2 and Supplementary Figure S1, respectively.

For the 43 searches where the number of records retrieved by the original reviewers was reported, the percentage difference in number of records for the original reviewers and each of the two replicators exceeded 10% for approximately half of the searches (search replicator 1: $N = 25/43$, 58%, 95% CI: 43–72; search replicator 2: $N = 21/43$, 49%, 95% CI: 35–63). In contrast, the difference between the number of records retrieved by the two replicators exceeded 10% in 14 of 45 cases (31%, 95% CI: 20–46). For no review or database did the results of all searches differ by less than 10%. The most common reasons for differences in the number of records retrieved between the replicators and original reviewers were thought to relate to syntax errors, undisclosed fields and limits, and differences in the platforms used, the database(s) searched within the platform and their default search settings (further details in Supplementary Table S2).

Table 3. *Characteristics of the reviews and index meta-analyses selected for replication.*

Author	Protocol registered	Full-text screeners per study	Data extractors per study	Population	Interventions	Outcome	Study designs	Total *N*	Summary statistics	Analysis software	Used PRISMA
AlAnouti et al.[46]	Yes	Two	Two	Metabolic syndrome	Vitamin D supplementation versus placebo	Fasting triglycerides	Trial	105	Yes	RevMan	Yes
Dohos et al.[47]	Yes	Two	Two	Inflammatory bowel disease	Immunomodulator withdrawal versus maintenance	Relapse rate	Trial	401	Yes	Stata	Yes
Gingold-Belfer et al.[48]	No	Unclear	Two	Refractory *Helicobacter pylori* infection	Rifabutin triple therapy versus any other treatment	Eradication success rate	Trial	395	No	CMA	Yes
Goldberg et al.[49]	Yes	Unclear	Two	Anxiety and depression	Psilocybin, ayahuasca, or LSD versus placebo or sub-clinical dose	Targeted symptoms	Trial and cohort	NR	No	R	Yes
Ioannou et al.[50]	No	Ten	Three	Depression	Sleep deprivation versus standard treatment	Hamilton Rating Scale	Trial	215	Yes	RevMan	Yes
Jakubczyk et al.[51]	No	Unclear	Unclear	No restrictions	Curcumin versus placebo	Total antioxidant capacity	Trial	NR	No	CMA	Yes
Minozzi et al.[52]	Yes	Two	Two	Opiate addiction during pregnancy	Methadone versus buprenorphine	Trial dropout rate	Trial	223	Yes	RevMan	No[a]
Wang et al.[53]	No	Two	Three	Oesophageal cancer	Two-field versus three-field lymphadenectomy	Overall survival	Cohort	NR	NA[b]	RevMan	Yes
Yekeduz et al.[54]	No	Two	Unclear	Cancer and COVID–19 infection	Chemotherapy versus no chemotherapy	All-cause mortality	Cohort and case control	NR	NA[c]	RevMan	Yes
Zhou et al.[55]	Yes	Two	Two	Diabetic limb ulcers	Low-level light therapy versus standard treatment (± placebo laser)	Percentage reduction in ulcer size	Trial	235	No	CMA	Yes

NR: not reported; CMA: comprehensive meta-analysis.

[a] This is not applicable as Cochrane reviewers are instructed to follow the MECIR standards.

[b] Not applicable as reviewers combined hazard ratios from survival analyses.

[c] Not applicable as reviewers combined adjusted odds ratios.

3.3. Replication of meta-analyses: full-text screening

In total, 185 articles were screened, of which 13 discrepancies occurred between the blinded replicators' consensus decision and the original reviewers (κ = 0.82, 95% CI: 0.72–0.91). At a review level, the replicators obtained perfect agreement or almost perfect agreement (i.e., a κ point estimate >0.80) with the original reviewers in half the reviews (Table 4). Following unblinding, when the 13 screening discrepancies were investigated, they were deemed to be due to the study coordinator giving incomplete or incorrect eligibility criteria to the replicators (N = 4/13, 31%), incomplete or ambiguous eligibility criteria in the review reports (N = 4/13, 31%), errors made by the replicators due lack of content expertise or failure to spot relevant information (N = 3/13, 23%), and unclear reasons (N = 2/13, 15%). Overall, there were 10 instances where the replicators changed their screening decision after being unblinded (five re-classified as eligible and five as ineligible), and three where they did not (one eligible and two ineligible). Further details on the circumstances behind the differing screening decisions are outlined in Supplementary Table S3.

3.4. Replication of meta-analyses: data extraction

Data were extracted from the 46 reports identified as eligible during the blinded phase, and the six re-classified as eligible in the unblinded phase. Data were mostly extracted directly from study reports (N = 35/52, 67%) or estimated from plots (N = 14/52, 27%). For the 40 reports which had data extracted by both the original reviewers and the blinded replicators, 27 discrepancies were observed. The reasons underpinning the observed discrepancies were deemed to be due to trivial differences related to the analysis software used (N = 9), variation when extracting data from plots (N = 7), incomplete methods for selecting study results in the review report (N = 7), reviewer errors (e.g., double counting of participants) (N = 3), and an unknown reason (N = 1). For the data extracted from studies re-classified as eligible in the unblinded phase, there were two discrepancies, one trivial software-related difference and one deemed to be due to under-specification of the methods for selecting study results in the review report. In total, 19 out of the 46 effect estimates and estimates of precision extracted by both the replicators (blinded and unblinded) and the original reviewers were classified as having differed due to non-software related reasons (41%, 95% CI: 28–56). Further details on the circumstances behind these 19 discrepancies are outlined in Supplementary Table S4.

3.5. Replication of meta-analyses: calculations

Both replicators were able to generate results for all 10 index meta-analyses using the data they extracted. Results of the original and replicated meta-analyses using data from the blinded and unblinded phases of the study are presented in Figure 3A and B (refer to Supplementary Figures S2–S11 and Supplementary Table S5 for more detailed results for each review). Using the data collected during the phase of the study where the replicators were blinded to the reviews' results, both replicators classified 8 of 10 meta-analyses (80%, 95% CI: 49–96) as not fully replicable (i.e., summary estimate or confidence interval length differed by more than 10% different to the original). One of 10 meta-analyses (10%, 95% CI: 1–40) was classified as meaningfully different from the original due to a difference in the statistical significance (Supplementary Table S6). This change in statistical significance was caused by the incorrect exclusion of a study by the replicators who were not informed by the study coordinator that the study was excluded due to the follow-up duration being too much longer than the other included studies.

When the 10 meta-analyses were performed with the unblinded data (i.e., data after discrepancies between the original reviewers and replicators were identified and addressed, and only reviewer errors, or unresolvable or trivial software-related discrepancies remained), the number of meta-analyses classified as not fully replicable decreased from eight to five (50%, 95% CI: 24–76) and the number considered meaningfully different from the original decreased from one to zero (0%, 95% CI: 0–28). Numerous factors contributed to the ultimate variation observed between the original reviewers'

Table 4. *Comparison of full-text article screening decisions between the reviewers and replicators' consensus judgements.*

Author	Articles screened	Number of studies included in the index meta-analysis			Agreement between replicators' consensus judgements and original reviewers' judgements (κ[a], 95% CI)		Percentage agreement between replicators' consensus judgements and original reviewers' judgements (%, 95% CI)	
		Original reviewers	Blinded replicators	Unblinded replicators	Blinded replicators	Unblinded replicators	Blinded replicators	Unblinded replicators
AlAnouti et al.[46]	7	2	3	2	0.70 (0.17, 1.00)	1.00 (1.00, 1.00)	86% (49%, 99%)	100% (65%, 100%)
Dohos et al.[47]	19	7	5	7	0.76 (0.45, 1.00)	1.00 (1.00, 1.00)	89% (69%, 98%)	100% (83%, 100%)
Gingold-Belfer et al.[48]	33	4	4	4	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	100% (90%, 100%)	100% (90%, 100%)
Goldberg et al.[49]	24	5	5	5	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	100% (86%, 100%)	100% (86%, 100%)
Ioannou et al.[50]	33	6	3	5	0.62 (0.24, 1.00)	0.89 (0.68, 1.00)	91% (76%, 97%)	97% (85%, 100%)
Jakubczyk et al.[51]	4	3	3	3	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	100% (51%, 100%)	100% (51%, 100%)
Minozzi et al.[52]	25	3	3	3	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	100% (87%, 100%)	100% (87%, 100%)
Wang et al.[53]	12	8	11	8	0.31 (−0.18, 0.80)	1.00 (1.00, 1.00)	75% (47%, 91%)	100% (76%, 100%)
Yekeduz et al.[54]	16	4	5	4	0.85 (0.56, 1.00)	1.00 (1.00, 1.00)	94% (72%, 100%)	100% (81%, 100%)
Zhou et al.[55]	12	6	5	6	0.50 (0.02, 0.98)	0.67 (0.24, 1.00)	75% (47%, 91%)	83% (55%, 97%)

CI: confidence interval.

[a] Kappa coefficients (κ) were interpreted as poor ($\leq$0.00), slight (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), almost perfect (0.81–0.99), or perfect (1.00).

	Included studies	Summary estimate	95% CI	P value (overall)	I^2	P value (chi-square)	Plot
Dohos et al (2020)*							
Original reviewers	7	1.85	(1.44, 2.38)	0.00	0	0.83	
Blinded: Replicator 1	5	1.84	(1.34, 2.55)	0.00	0	0.82	
Blinded: Replicator 2	5	1.84	(1.34, 2.54)	0.00	0	0.82	
Unblinded: Replicator 1	7	1.94	(1.49, 2.54)	0.00	0	0.68	
Unblinded: Replicator 2	7	1.94	(1.49, 2.54)	0.00	0	0.68	
Yekeduz et al (2020)^							
Original reviewers	4	1.85	(1.26, 2.71)	0.00	0	0.55	
Blinded: Replicator 1	5	1.53	(1.12, 2.10)	0.01	17	0.31	
Blinded: Replicator 2	5	1.53	(1.12, 2.10)	0.01	17	0.31	
Unblinded: Replicator 1	4	1.85	(1.26, 2.71)	0.00	0	0.55	
Unblinded: Replicator 2	4	1.85	(1.26, 2.71)	0.00	0	0.55	
Minozzi et al (2020)*							
Original reviewers	3	0.66	(0.37, 1.20)	0.17	18	0.29	
Blinded: Replicator 1	3	0.66	(0.37, 1.20)	0.17	18	0.30	
Blinded: Replicator 2	3	0.66	(0.37, 1.19)	0.17	18	0.30	
Unblinded: Replicator 1	3	0.66	(0.37, 1.20)	0.17	18	0.30	
Unblinded: Replicator 2	3	0.66	(0.37, 1.19)	0.17	18	0.30	
Wang et al (2020)#							
Original reviewers	8	1.05	(0.90, 1.21)	0.56	0	0.78	
Blinded: Replicator 1	11	1.06	(0.94, 1.18)	0.34	79	0.00	
Blinded: Replicator 2	11	1.06	(0.95, 1.18)	0.33	79	0.00	
Unblinded: Replicator 1	8	1.07	(0.93, 1.23)	0.35	0	0.58	
Unblinded: Replicator 2	8	1.07	(0.93, 1.23)	0.36	0	0.57	
Gingold-Belfer et al (2020)^							
Original reviewers	4	0.89	(0.44, 1.79)	0.73	68	0.02†	
Blinded: Replicator 1	4	0.87	(0.43, 1.76)	0.70	65	0.04	
Blinded: Replicator 2	4	0.87	(0.43, 1.76)	0.70	65	0.04	
Unblinded: Replicator 1	4	0.87	(0.43, 1.76)	0.70	65	0.04	
Unblinded: Replicator 2	4	0.87	(0.43, 1.76)	0.70	65	0.04	

*Risk ratio, ^Odds Ratio, #Hazard ratio. †Value was estimated in R. Plot details: Range: 0.33 to 3, Scale transformation: Log (base 10), Dotted line: Summary estimate = 1 (Null value)

Figure 3A. *Replication of meta-analyses of (A) ratio measures and (B) difference measures: Comparison of the results of the original reviewer's and the replicators' (blinded and unblinded) index meta-analyses.*

and blinded and unblinded replicators' summary estimates (Supplementary Tables S7 and S8). After correction of replicator errors, the main drivers of variation in the summary estimates from the unblinded replications were unclear or ambiguous eligibility criteria in the review report, under-specification of the methods for selecting study results, and reviewer data extraction errors.

For the sensitivity analyses that removed potentially unreliable hazard ratios from one meta-analysis and included estimated summary statistics for another (Supplementary Figures S12 and S13), the number of meta-analyses based on the blinded data that were classified as not fully replicable did not change. However, summary estimates from the unblinded replications were not robust to the sensitivity analyses, with the percentage classified as not fully replicable increasing from 50% to 60%. The percentage of replicated meta-analyses considered to be meaningfully different from the original also increased from 0% to 20% due to the first summary estimate changing from statistically significant to non-significant and the second reversing direction.

	Included studies	Summary estimate	95% CI	P value (overall)	I^2	P value (chi-square)	Plot
Meta-analyses of standardised mean differences (SMD)							
Ioannou et al (2020)*							
Original reviewers	6	-0.29	(-0.84, 0.25)	0.29	70	0.01	
Blinded: Replicator 1	2	-0.32	(-1.06, 0.42)	0.40	0	0.58	
Blinded: Replicator 2	2	-0.33	(-1.02, 0.36)	0.35	0	0.55	
Unblinded: Replicator 1	5	-0.27	(-0.96, 0.43)	0.45	74	0.00	
Unblinded: Replicator 2	5	-0.28	(-0.96, 0.41)	0.43	75	0.00	
Zhou et al (2021)*							
Original reviewers	6	2.81	(1.14, 4.48)	0.00	95†	0.00†	
Blinded: Replicator 1	5	3.62	(2.31, 4.93)	0.00	83	0.00	
Blinded: Replicator 2	5	3.62	(2.33, 4.91)	0.00	84	0.00	
Unblinded: Replicator 1	6	3.38	(2.21, 4.55)	0.00	81	0.00	
Unblinded: Replicator 2	6	3.38	(2.22, 4.53)	0.00	83	0.00	
Jakubczuk et al (2020)#							
Original reviewers	3	2.70	(0.06, 5.34)	0.05	98	0.00	
Blinded: Replicator 1	3	2.68	(0.08, 5.28)	0.04	98	0.00	
Blinded: Replicator 2	3	2.70	(0.06, 5.34)	0.05	98	0.00	
Unblinded: Replicator 1	3	2.68	(0.08, 5.28)	0.04	98	0.00	
Unblinded: Replicator 2	3	2.70	(0.06, 5.34)	0.05	98	0.00	
Goldberg et al (2020)*							
Original reviewers	5	1.08	(0.74, 1.43)	0.00†	0	0.45†	
Blinded: Replicator 1	5	0.84	(0.31, 1.37)	0.00	45	0.12	
Blinded: Replicator 2	5	0.84	(0.31, 1.38)	0.00	47	0.12	
Unblinded: Replicator 1	5	0.84	(0.31, 1.37)	0.00	45	0.12	
Unblinded: Replicator 2	5	0.84	(0.31, 1.38)	0.00	47	0.12	
Meta-analyses of mean differences (MD)							
AlAnouti et al (2020)^							
Original reviewers	2	30.67	(4.89, 56.45)	0.02	86	0.01	
Blinded: Replicator 1	3	0.37	(-9.19, 9.92)	0.94	85	0.00	
Blinded: Replicator 2	3	0.46	(-9.05, 9.97)	0.93	86	0.00	
Unblinded: Replicator 1	2	29.47	(4.24, 54.71)	0.02	87	0.01	
Unblinded: Replicator 2	2	30.02	(4.94, 55.11)	0.02	87	0.01	

*Hedge's g, #Cohen's d, ^Mean difference in serum triglycerides (mg/dL), †Values were estimated in R. SMD plot scale range: -1.2 to 5.6, MD scale range: -10 to 60, Scale transformations: None, Dotted line: Summary estimate = 0 (Null value).

Figure 3B. *(Continued)*

4. Discussion

This study found variation in results when replicating the literature searches and first reported meta-analyses of 10 systematic reviews, which is consistent with previous research. For the literature search replications, while all but one could be successfully re-run, the replicators were only able to retrieve a similar number of records for around half the searches (49% and 58%), which is consistent with two other similar studies (47% and 33%, respectively).[18,29] The percentage of discrepancies occurring in the data extraction step that were determined to be due to errors made by the reviewers (11%, $N = 3/27$) is consistent with an investigation of data extraction errors within 201 systematic reviews of adverse events[26] (17%, $N = 1762/10{,}386$). Furthermore, when applying our definitions for replication success to the studies by Ford et al.,[30] Carrol et al.,[24] and Xu et al.,[26] they would have classified 31%, 50%, and 42% of their meta-analyses of binary data as not fully replicable, respectively (our estimate is

50%). The same studies would have also only classified 25%, 0%, and less than 10% as 'meaningfully' different from the original meta-analysis, respectively (our estimate is 0%). Similar findings have also been observed among replications of meta-analyses of continuous and ordinal data.[22,23]

4.1. Implications of the research

When reflecting on the causes of variation observed when replicating the literature searches and meta-analyses of systematic reviews, and how it might be minimized, three key implications emerge from our research. Our findings suggest that most reported database searches are likely not replicable, and that replication success is likely to be largely driven by the reporting completeness of the search strategy. Therefore, it is likely that search replicability could be enhanced if key details outlined in the PRISMA 2020[56] and PRISMA-S[57] statements (e.g., exact databases and platforms searched, field, and limits used) were more routinely reported.

Inadequate specification of the eligibility criteria for each synthesis, along with guidance on how to select results when there is multiplicity, explained the variation between the original reviewers' and replicators' summary estimates in some cases. The influence of incomplete reporting of eligibility criteria on replicability has been previously foreshadowed.[58,59] For example, in a recent evaluation of 41 systematic reviews of health interventions by Cumpston et al.,[59] the authors estimated that less than a third of the reviews provided enough information to allow someone to replicate decisions about which included studies were eligible for each synthesis. In our sample, we note that the provision of information that clarified acceptable and unacceptable co-interventions, outlined rules for selecting results, and avoided ambiguous/conflicting instructions would have led to greater consistency between the original reviewers' and replicators' results. Such issues could be addressed through greater uptake of resources which assist review authors with developing and reporting their research questions and eligibility criteria for all planned and performed syntheses, such as the Cochrane Handbook for Systematic Reviews of Interventions[60] and the recently developed InSynQ (Intervention Synthesis Questions) checklist.[61]

Finally, more thorough data documentation practices (e.g., archival of, at minimum, the extracted summary statistics) might have yielded more consistent outcomes between the replicators and original reviewers. For example, greater availability of summary statistics from the studies included in the meta-analyses would likely have helped the replicators resolve discrepancies between replicators and original reviewers, particularly those related to the selection of results. However, improvements to documentation practices likely require top-down intervention (e.g., journal, funder, or institutional mandates) given previous research has shown, in the absence of a mandatory sharing policy, less than 1% of systematic reviewers make their data publicly available.[62]

4.2. Strengths and limitations

This study had several strengths. We improved on previous replication research[21,22,26,29,30] by initially blinding the replicators to ensure they were not influenced by the original reviewers' results. Importantly, in contrast to replication work involving participants, where experimental set ups can be sensitive to seemingly innocuous contextual factors (i.e., 'hidden moderators'),[63] we were able to directly observe, investigate, and resolve discrepancies, often with the assistance of the original reviewers. However, we also note several limitations. First, though consistent with the study by Ford and colleagues,[30] for logistical reasons our sample size was small, which has led to large imprecision in our estimates of the percentage of meta-analysis results that were fully replicable or were meaningfully different to the original. Information initially provided to the replicators was collected by one investigator only which resulted in the introduction of some errors. However, these errors were identified and addressed through the unblinded analysis. The study adopted a crude measure to classify 'meaningful' differences in summary estimates from meta-analyses which ignores clinically important changes in effect size and precision that could impact the certainty of evidence. However,

given the minor changes in the summary estimates and confidence intervals overall, it is unlikely that our conclusions would have changed had we used a measure which factored in minimally important differences. The study also restricted its scope to the first reported meta-analysis of systematic reviews that met a high threshold for reporting quality and employed highly experienced replicators. Therefore, our results may not generalize to reviews that have less complete reporting or be comparable to findings of other replication initiatives where replicators have minimal experience performing systematic reviews. We also did not replicate other important review processes (e.g., title and abstract screening) and at times relied on information reported by the original reviewers to guide our decision making (e.g., risk of bias judgements).

4.3. Future directions

For this study, we adopted language consistent with most systematic reviewers' understanding of the term 'replication'[35] as well as definitions proposed by the National Academies of Sciences, Engineering, and Medicine.[64] However, we note that the terms 'replicability' and 'reproducibility' are still often used interchangeably.[35,65] Consequently, as suggested by others,[35,65,66] standardization of these terms, in addition to the continued discussion of other important challenges associated with performing replication studies, such as conflicts of interest,[35,67,68] shortage of funding opportunities,[35,69,70] determining when replications are informative and when they are redundant,[66,71,72] and addressing difficulties associated with publishing them[35,68] should also be a research priority. We also highlight a lack of evidence on the replicability of other evidence synthesis methods (both quantitative and qualitative) beyond pairwise meta-analysis of aggregate data. There is also a need to assess the replicability of AI-supported/generated reviews given their increasing prevalence in the literature.[73]

Finally, while the inability to replicate published findings has been a focal point of the replication crisis, it is but only one factor that contributes to our confidence in the reliability of a scientific claim. For example, as Errington et al.[15] keenly note, 'a finding can be both replicable and invalid at the same time'. Therefore, while we have shown that summary estimates from meta-analyses were seldom 'meaningfully' affected by discrepancies occurring in earlier stages of the review, we have not demonstrated that the summary estimates themselves are valid.[74] As such, the performance of replications that address methodological shortcomings observed in original reviews would provide another useful line of research into the reliability of evidence generated from systematic reviews.

5. Conclusion

Our results show that the results of systematic review processes are not always consistent when their methods were repeated. However, these inconsistencies seldom affected summary estimates from meta-analyses in a meaningful way. Lack of detail about the search methods, eligibility criteria and methods concerning which results to extract from eligible studies were impediments to replicability. Future work should investigate methods to improve author engagement with resources designed to improve the reporting of systematic reviews.

Acknowledgements. We thank Dr Rebecca Hamilton, Dr Khalia Primer, and Dr Jason Wasiak for their assistance with the study.

Author contributions. Conceptualization, supervision: JEM and MJP. Data curation, project administration: DGH and MJP. Formal analysis: DGH, JEM, PN, and MJP. Funding acquisition: MJP. Investigation: DGH, PN, MLR, SM, and MJP. Methodology, writing—review and editing: DGH, JEM, PN, MLR, SM, SB, FMF, JPH, RK, SK, LJM, DM, SN, DN, PT, VAW, and MJP. Resources: DGH, JEM, PN, MLR, SM, and MJP. Software: DGH and PN. Validation, writing—original draft: DGH. Visualization: DGH, JEM, and MJP.

Competing interest statement. The authors declare that no competing interests exist.

Data availability statement. All data, code, and materials associated with this project are publicly available on the Open Science Framework under a CC-BY license (DOI: https://doi.org/10.17605/OSF.IO/V3MNH).

Funding statement. This research was funded by an Australian Research Council Discovery Early Career Researcher Award (DE200101618), held by MJP. MJP is supported by a National Health and Medical Research Council Investigator Grant (GNT2033917). JEM is supported by a National Health and Medical Research Council Investigator Grant (GNT2009612). P-YN is supported by a Monash Graduate Scholarship and a Monash International Tuition Scholarship. JPTH is supported in part by the National Institute for Health and Care Research (NIHR203807, NIHR153861, and NIHR200181). The funders had no role in the study design, decision to publish, or preparation of the manuscript.

Ethics approval and consent to participate. Ethics approval was not required for this research.

Study registration. The protocol for the study is publicly available at https://doi.org/10.1186/s13643-021-01670-0.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/rsm.2025.10064.

References

[1] Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533: 452–454. https://doi.org/10.1038/533452a.
[2] Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294: 218–228. https://doi.org/10.1001/jama.294.2.218.
[3] Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2: e124. https://doi.org/10.1371/journal.pmed.0020124.
[4] Mobley A, Linder SK, Braeuer R, Ellis LM, Zwelling L. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLoS One*. 2013;8: e63221. https://doi.org/10.1371/journal.pone.0063221.
[5] National Health and Medical Research Council. Survey of research culture in Australian NHMRC-funded institutions: survey findings report. *ORIMA Res*. 2020. https://www.nhmrc.gov.au/file/15322/download?token=SpPpeE6j.
[6] Hartshorne J, Schachner A. Tracking replicability as a method of post-publication open evaluation. Front Comput Neurosci 2012; 6. https://doi.org/10.3389/fncom.2012.00008.
[7] Nosek BA, Hardwicke TE, Moshontz H, et al. Replicability, robustness, and reproducibility in psychological science. *Annu Rev Psychol*. 2022;73(1): 719–748. https://doi.org/10.1146/annurev-psych-020821-114157.
[8] Chang AC, Li P. Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not". *Finance and Economics Discussion Series 2015–083.* Washington: Board of Governors of the Federal Reserve System; 2015. http://dx.doi.org/10.17016/FEDS.2015.083.
[9] Camerer CF, Dreber A, Forsell E, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016;351: 1433–1436. https://doi.org/10.1126/science.aaf0918.
[10] Camerer CF, Dreber A, Holzmeister F, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat Hum Behav*. 2018;2: 637–644. https://doi.org/10.1038/s41562-018-0399-z.
[11] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*. 2011;10(9): 712. https://doi.org/10.1038/nrd3439-c1.
[12] Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature*. 2012;483(7391): 531–533. https://doi.org/10.1038/483531a.
[13] Perrin S. Preclinical research: make mouse studies work. *Nature*. 2014;507(7493): 423–425. https://doi.org/10.1038/507423a.
[14] Steward O, Popovich PG, Dietrich WD, Kleitman N. Replication and reproducibility in spinal cord injury research. *Exp Neurol*. 2012;233(2): 597–605. https://doi.org/10.1016/j.expneurol.2011.06.017.
[15] Errington TM, Mathur M, Soderberg CK, et al. Investigating the replicability of preclinical cancer biology. *eLife*. 2021;10: e71601. https://doi.org/10.7554/eLife.71601.
[16] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd; 2009. https://doi.org/10.1002/9780470743386.
[17] Siddaway AP, Wood AM, Hedges LV. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu Rev Psychol*. 2019;70: 747–770. https://doi.org/10.1146/annurev-psych-010418-102803.
[18] Rethlefsen ML, Brigham TJ, Price C, et al. Systematic review search strategies are poorly reported and not reproducible: a cross-sectional metaresearch study. *J Clin Epidemiol*. 2024;166. https://doi.org/10.1016/j.jclinepi.2023.111229.
[19] Pham MT, Waddell L, Rajić A, Sargeant JM, Papadopoulos A, McEwen SA. Implications of applying methodological shortcuts to expedite systematic reviews: three case studies using systematic reviews from Agri-food public health. *Res Synth Methods*. 2016;7: 433–446. https://doi.org/10.1002/jrsm.1215.
[20] Gartlehner G, Affengruber L, Titscher V, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *J Clin Epidemiol*. 2020;121: 20–28. https://doi.org/10.1016/j.jclinepi.2020.01.005.
[21] Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *J Clin Epidemiol*. 2005;58: 741–742. https://doi.org/10.1016/j.jclinepi.2004.11.024.

[22] Gøtzsche PC, Hróbjartsson A, Marić K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA*. 2007;298: 430–437. https://doi.org/10.1001/jama.298.4.430.
[23] Tendal B, Higgins JPT, Jüni P, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ*. 2009;339: b3128. https://doi.org/10.1136/bmj.b3128.
[24] Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. *BMC Res Notes*. 2013;6: 539. https://doi.org/10.1186/1756-0500-6-539.
[25] Bonetti AF, Tonin FS, Lucchetta RC, Pontarolo R, Fernandez-Llimos F. Methodological standards for conducting and reporting meta-analyses: ensuring the replicability of meta-analyses of pharmacist-led medication review. *Res Soc Adm Pharm*. 2022;18: 2259–2268. https://doi.org/10.1016/j.sapharm.2021.06.002.
[26] Xu C, Yu T, Furuya-Kanamori L, et al. Validity of data extraction in evidence synthesis practice of adverse events: reproducibility study. *BMJ*. 2022;377: e069155. https://doi.org/10.1136/bmj-2021-069155.
[27] Könsgen N, Barcot O, Heß S, et al. Inter-review agreement of risk-of-bias judgments varied in Cochrane reviews. *J Clin Epidemiol*. 2020;120: 25–32. https://doi.org/10.1016/j.jclinepi.2019.12.016.
[28] Bertizzolo L, Bossuyt P, Atal I, Ravaud P, Dechartres A. Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews: a research on research study using cross-sectional design. *BMJ Open*. 2019;9: e028382. https://doi.org/10.1136/bmjopen-2018-028382.
[29] Pieper D, Heß S, Faggion CM. A new method for testing reproducibility in systematic reviews was developed, but needs more testing. *BMC Med Res Methodol*. 2021;21: 157. https://doi.org/10.1186/s12874-021-01342-6.
[30] Ford AC, Guyatt GH, Talley NJ, Moayyedi P. Errors in the conduct of systematic reviews of pharmacological interventions for irritable bowel syndrome. *Am J Gastroenterol* 2010;105: 280–288. https://doi.org/10.1038/ajg.2009.658.
[31] Page MJ, Moher D, Fidler FM, et al. The REPRISE project: protocol for an evaluation of REProducibility and replicability in syntheses of evidence. *Syst Rev*. 2021;10: 112. https://doi.org/10.1186/s13643-021-01670-0.
[32] Page MJ, Nguyen P-Y, Hamilton DG, et al. Data and code availability statements in systematic reviews of interventions were often missing or inaccurate: a content analysis. *J Clin Epidemiol*. 2022;147: 1–10. https://doi.org/10.1016/j.jclinepi.2022.03.003.
[33] Nguyen P-Y, Kanukula R, McKenzie JE, et al. Changing patterns in reporting and sharing of review data in systematic reviews with meta-analysis of the effects of interventions: cross sectional meta-research study. BMJ. 2022: e072428. https://doi.org/10.1136/bmj-2022-072428.
[34] Nguyen P-Y, McKenzie JE, Hamilton DG, et al. Systematic reviewers' perspectives on sharing review data, analytic code, and other materials: a survey. *Cochrane Evid Synth Methods*. 2023;1: e12008. https://doi.org/10.1002/cesm.12008.
[35] Nguyen P-Y, McKenzie JE, Hamilton DG, et al. Systematic reviewers' perspectives on replication of systematic reviews: a survey. *Cochrane Evid Synth Methods*. 2023;1: e12009. https://doi.org/10.1002/cesm.12009.
[36] Nguyen PY, McKenzie JE, Alqaidoom Z, Hamilton DG, Moher D, Page MJ. Reproducibility of meta-analytic results in systematic reviews of interventions: meta-research study. *BMJ Med*. 2025;4(1): e002024. https://doi.org/10.1136/bmjmed-2025-002024.
[37] Hamilton DG, Page MJ. REPRISE Project | Study 4 | Evaluation of the replicability of systematic reviews and meta-analyses of health interventions. Open Science Framework 2024. https://doi.org/10.17605/OSF.IO/V3MNH.
[38] Barba LA. Terminologies for reproducible research. arXiv:1802.03311. 2018.
[39] Vachon B, Curran JA, Karunananthan S, et al. Replication research series-paper 1: a concept analysis and meta-narrative review established a comprehensive theoretical definition of replication research to improve its use. *J Clin Epidemiol*. 2021;129: 176–187.
[40] Nosek BA, Errington TM. What is replication? *PLoS Biol*. 2020;18(3): e3000691.
[41] Machery E. What is a replication? *Philos Sci*. 2020;87(4): 545–567. https://doi.org/10.1086/709701.
[42] Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12: 9. https://doi.org/10.1186/1471-2288-12-9.
[43] Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci*. 2001;16(2): 101–117.
[44] McGrath S, Katzenschlager S, Zimmer AJ, Seitel A, Steele R, Benedetti A. Standard error estimation in meta-analysis of studies reporting medians. *Stat Methods Med Res*. 2023;32: 373–388. https://doi.org/10.1177/09622802221139233.
[45] Cai S, Zhou J, Pan J. Estimating the sample mean and standard deviation from order statistics and sample size in meta-analysis. *Stat Methods Med Res*. 2021;30: 2701–2719. https://doi.org/10.1177/0962280211047348.
[46] AlAnouti F, Abboud M, Papandreou D, Mahboub N, Haidar S, Rizk R. Effects of vitamin D supplementation on lipid profile in adults with the metabolic syndrome: a systematic review and meta-analysis of randomized controlled trials. *Nutrients*. 2020;12: 3352. https://doi.org/10.3390/nu12113352.
[47] Dohos D, Hanák L, Szakács Z, et al. Systematic review with meta-analysis: the effects of immunomodulator or biological withdrawal from mono- or combination therapy in inflammatory bowel disease. *Aliment Pharmacol Ther*. 2021;53: 220–233. https://doi.org/10.1111/apt.16182.
[48] Gingold-Belfer R, Niv Y, Levi Z, Boltin D. Rifabutin triple therapy for first-line and rescue treatment of *Helicobacter pylori* infection: a systematic review and meta-analysis. *J Gastro Hepatol*. 2021;36: 1392–1402. https://doi.org/10.1111/jgh.15294.
[49] Goldberg SB, Shechet B, Nicholas CR, et al. Post-acute psychological effects of classical serotonergic psychedelics: a systematic review and meta-analysis. *Psychol Med*. 2020;50: 2655–2666. https://doi.org/10.1017/S003329172000389X.

[50] Ioannou M, Wartenberg C, Greenbrook JTV, et al. Sleep deprivation as treatment for depression: systematic review and meta-analysis. *Acta Psychiatr Scand*. 2021;143: 22–35. https://doi.org/10.1111/acps.13253.
[51] Jakubczyk K, Drużga A, Katarzyna J, Skonieczna-Żydecka K. Antioxidant potential of curcumin—a meta-analysis of randomized clinical trials. *Antioxidants*. 2020;9: 1092. https://doi.org/10.3390/antiox9111092.
[52] Minozzi S, Amato L, Jahanfar S, Bellisario C, Ferri M, Davoli M. Maintenance agonist treatments for opiate-dependent pregnant women. *Cochrane Database Syst Rev* 2020;2020. https://doi.org/10.1002/14651858.CD006318.pub4.
[53] Wang J, Yang Y, Shafiulla Shaik M, et al. Three-field versus two-field lymphadenectomy for Esophageal squamous cell carcinoma: a meta-analysis. *J Surg Res*. 2020;255: 195–204. https://doi.org/10.1016/j.jss.2020.05.057.
[54] Yekedüz E, Utkan G, Ürün Y. A systematic review and meta-analysis: the effect of active cancer treatment on severity of COVID-19. *Eur J Cancer*. 2020;141: 92–104. https://doi.org/10.1016/j.ejca.2020.09.028.
[55] Zhou Y, Chia HWA, Tang HWK, et al. Efficacy of low-level light therapy for improving healing of diabetic foot ulcers: a systematic review and meta-analysis of randomized controlled trials. *Wound Repair Regen*. 2021;29: 34–44. https://doi.org/10.1111/wrr.12871.
[56] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372: n71. https://doi.org/10.1136/bmj.n71.
[57] Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Syst Rev*. 2021;10: 39. https://doi.org/10.1186/s13643-020-01542-z.
[58] Hacke C, Nunan D. Discrepancies in meta-analyses answering the same clinical question were hard to explain: a meta-epidemiological study. *J Clin Epidemiol*. 2020;119: 47–56. https://doi.org/10.1016/j.jclinepi.2019.11.015.
[59] Cumpston MS, McKenzie JE, Ryan R, Thomas J, Brennan SE. Critical elements of synthesis questions are incompletely reported: survey of systematic reviews of intervention effects. *J Clin Epidemiol*. 2023;163: 79–91. https://doi.org/10.1016/j.jclinepi.2023.09.013.
[60] McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnston RV, Thomas J. Chapter 3: defining the criteria for including studies and how they will be grouped for the synthesis [last updated August 2023]. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 6.5*. Cochrane; 2024. www.training.cochrane.org/handbook.
[61] Brennan SE, Cumpston MS, Ryan R, McKenzie JE. InSynQ (intervention synthesis questions) checklist and guide for developing and reporting the questions addressed in systematic reviews of interventions. Version 1.0. 2023. Accessed October 17, 2024. https://www.insynq.info/.
[62] Hamilton DG, Hong K, Fraser H, Rowhani-Farid A, Fidler F, Page MJ. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. *BMJ*. 2023;382: e075767. https://doi.org/10.1136/bmj-2023-075767.
[63] Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA. Contextual sensitivity in scientific reproducibility. *Proc Natl Acad Sci*. 2016;113: 6454–6459. https://doi.org/10.1073/pnas.1521897113.
[64] National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC: National Academies Pres; 2019.
[65] Puljak L, Pieper D. Replicability in the context of systematic reviews: a call for a framework with considerations regarding duplication, overlap, and intentionality. *J Clin Epidemiol*. 2022;142: 313–314. https://doi.org/10.1016/j.jclinepi.2021.11.014
[66] Tugwell P, Welch VA, Karunananthan S, et al. When to replicate systematic reviews of interventions: consensus checklist. *BMJ*. 2020;370: m2864. https://doi.org/10.1136/bmj.m2864
[67] Munafò M. Conflicts of interest and solicited replication attempts. *Nicotine Tob Res*. 2016;18(4): 377–378. https://doi.org/10.1093/ntr/ntw031
[68] Hamilton DG, Fraser H, Hoekstra R, Fidler F. Journal policies and editors' opinions on peer review. *elife* 2020;9:e62529. https://doi.org/10.7554/eLife.62529.
[69] Drude NI, Gamboa LM, Danziger M, Dirnagl U, Toelch U. Improving preclinical studies through replications. *elife*. 2021;10: e62101. https://doi.org/10.7554/eLife.62101
[70] Ioannidis JP. *Failure to Replicate: Sound the Alarm*. *Cerebrum*. 2015;2015: cer-12a-15.
[71] Moher D. The problem of duplicate systematic reviews. *BMJ*. 2013;347: f5040. https://doi.org/10.1136/bmj.f5040.
[72] Karunananthan S, Maxwell LJ, Welch V, et al. When and how to replicate systematic reviews. *Cochrane Database Syst Rev*. 2020;2020: MR000052. https://doi.org/10.1002/14651858.MR000052.
[73] Siemens W, Von Elm E, Binder H, et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. *BMJ EBM*. 2025: bmjebm-2024-113320. https://doi.org/10.1136/bmjebm-2024-113320.
[74] Wanous JP, Sullivan SE, Malinak J. The role of judgment calls in meta-analysis. *J Appl Psychol*. 1989;74(2): 259–264. https://doi.org/10.1037/0021-9010.74.2.259.

Cite this article: Hamilton DG, McKenzie JE, Nguyen P-Y., Rethlefsen ML, McDonald S, Brennan SE, Fidler FM, Higgins JPT, Kanukula R, Karunananthan S, Maxwell LJ, Moher D, Nakagawa S, Nunan D, Tugwell P, Welch VA, Page MJ. Evaluation of the replicability of systematic reviews with meta-analyses of the effects of health interventions. *Research Synthesis Methods*. 2026;00: 1–19. https://doi.org/10.1017/rsm.2025.10064