Journal of Clinical Epidemiology 166 (2024) 111229

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Systematic review search strategies are poorly reported and not reproducible: a cross-sectional metaresearch study

Melissa L. Rethlefsen[a,b,*], Tara J. Brigham[c], Carrie Price[d], David Moher[e], Lex M. Bouter[f,g], Jamie J. Kirkham[h], Sara Schroter[i,j], Maurice P. Zeegers[b,k]

[a]*Health Sciences Library & Informatics Center, University of New Mexico, MSC 09 5100, 1 University of New Mexico, Albuquerque, NM 87131-0001, USA*
[b]*Department of Epidemiology, Maastricht University, Maastricht, The Netherlands*
[c]*Library Services-Florida, Mayo Clinic Libraries, Mayo Clinic, 4500 San Pablo Road, Jacksonville, FL 32224, USA*
[d]*Albert S. Cook Library, Towson University, 8000 York Road, Towson, MD 21252, USA*
[e]*Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, The Ottawa Hospital, General Campus, Centre for Practice Changing Research Building, 501 Smyth Road, PO BOX 201B, Ottawa, Ontario K1H 8L6, Canada*
[f]*Department of Epidemiology and Data Science, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1089a, 1081 HV Amsterdam, The Netherlands*
[g]*Department of Philosophy, Faculty of Humanities, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands*
[h]*Centre for Biostatistics, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK*
[i]*BMJ, BMA House, Tavistock Square, London WC1H 9JR, UK*
[j]*Faculty of Public Health & Policy, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK*
[k]*MBP Holding, Heerlen, The Netherlands*

Accepted 27 November 2023; Published online 3 December 2023

Abstract

Objectives: To determine the reproducibility of biomedical systematic review search strategies.

Study Design and Setting: A cross-sectional reproducibility study was conducted on a random sample of 100 systematic reviews indexed in MEDLINE in November 2021. The primary outcome measure is the percentage of systematic reviews for which all database searches can be reproduced, operationalized as fulfilling six key Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension (PRISMA-S) reporting guideline items and having all database searches reproduced within 10% of the number of original results. Key reporting guideline items included database name, multi-database searching, full search strategies, limits and restrictions, date(s) of searches, and total records.

Results: The 100 systematic review articles contained 453 database searches. Only 22 (4.9%) database searches reported all six PRISMA-S items. Forty-seven (10.4%) database searches could be reproduced within 10% of the number of results from the original search; six searches differed by more than 1,000% between the originally reported number of results and the reproduction. Only one systematic review article provided the necessary search details to be fully reproducible.

Conclusion: Systematic review search reporting is poor. To correct this will require a multifaceted response from authors, peer reviewers, journal editors, and database providers.

Keywords: Systematic reviews; Search strategies; Reporting guidelines; Transparency; Reproducibility; Database searches

1. Introduction

Systematic reviews provide synthesized evidence using robust, prespecified methods to reduce bias and enable better decision-making by healthcare professionals, patients, and policy makers [1,2]. Transparent reporting is essential because it enables readers to evaluate the value of systematic reviews and to identify potential sources of bias that impact the review's findings [3]. A sensitive literature search encompassing multiple information sources is

Registration: This study is registered using Open Science Framework (https://doi.org/10.17605/OSF.IO/KBVSR).

* Corresponding author. Health Sciences Library & Informatics Center, University of New Mexico, MSC 09 5100, 1 University of New Mexico, Albuquerque, NM 87131-0001, USA. Tel: +1-505-272-2311.

E-mail addresses: mlrethlefsen@gmail.com; m.rethlefsen@maastrichtuniversity.nl (M.L. Rethlefsen).

https://doi.org/10.1016/j.jclinepi.2023.111229

What is new?

Key Findings

- Most systematic review searches are not reproducible. Only 1% of systematic reviews in our sample was fully reproducible for all database searches.
- Systematic review searches remain poorly reported. The vast majority of systematic reviews (91%) did not even provide enough details to identify the databases/platforms used for all database searches.

What this adds to what was known?

- Published search strategies may appear reproducible, but reproducing them can reveal wide disparities between numbers of results expected and achieved.

What is the implication and what should change now?

- Systematic review searches must be acknowledged as essential data and code to preserve, document, and share to improve transparency.

central to the methodology of systematic reviews, and the rigor of information retrieval process is key to the reliability and generalizability of a systematic review [1,2]. The search must be transparently and completely reported. It should be possible to reproduce the search and to assess whether this leads to similar results.

Systematic reviews are often poorly reported, which impacts reproducibility and leaves readers unable to assess potential biases and usability [4−8]. Systematic review searches are particularly vulnerable to incomplete reporting [9−13]. Estimates of systematic review search reproducibility vary widely, in part because many researchers have developed their own criteria for estimating reproducibility [10,11,14−16]. In 2021, PRISMA-S, an extension of PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses), a guideline for reporting systematic reviews, was designed to help systematic review teams report their searches for maximum transparency and reproducibility [17]. PRISMA-S offers a standard for the details systematic review searches need to describe to be reproducible [17]. However, this is not enough to determine actual search reproducibility; reproducing the search is required.

We sought to determine the reproducibility of systematic review searches. Understanding the reproducibility of published searches can serve as a baseline for improvement. Information sources and methods used vary widely across systematic reviews, but for the vast majority, searching one or more databases is the primary way to identify studies. In this study, we focus solely on database searches and not on supplementary search methods to enable equitable assessment. The primary outcome of this study is the percentage of systematic reviews for which all database searches can be reproduced. Secondary outcomes include:

(1) The percentage of systematic reviews for which
   (a) one database search can be fully reproduced
   (b) more than one database search can be fully reproduced
(2) The percentage difference between the number of search results reported in a systematic review vs. the number of results in the reproduced search
(3) The number of systematic reviews meeting individual elements needed for reproducibility for one or more databases

In our study, we used the definition of "reproducibility" from the National Academies of Science, Engineering, and Medicine, "Reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis" [18].

2. Materials and methods

Before data collection, the study protocol was registered on the Open Science Framework (OSF) [19]. The registration was amended during the study to reflect minor changes in the protocol (see study protocol amendments section).

2.1. Identification and selection of articles

We aimed for a cohort of 100 systematic reviews, selected randomly from a full month of publications indexed in MEDLINE, to try to make the sample as representative as possible in terms of specialty and topic, and being feasible for the research team to replicate. We first searched for systematic reviews indexed in MEDLINE in one calendar month, November 2021. Because it can take 6 months or more to be indexed in MEDLINE, this enabled us to get a snapshot of items from a range of publication dates. We only included publications in English for pragmatic reasons. We adapted the search published in Page et al. [7] and conducted it on December 20, 2021, using Ovid MEDLINE ALL <1946 to December 17, 2021>; the complete search strategy is available in the study's OSF Project site [20]. Results were deduplicated using Covidence [21].

To be considered a systematic review, articles needed to meet the PRISMA-P definition of a systematic review [22], namely they must explicitly state methods to locate studies (i.e., the search), methods to select studies (i.e., screening using eligibility criteria), and methods to synthesize the studies, whether qualitatively or quantitatively. In addition, the article had to explicitly state that one or more literature databases (e.g., MEDLINE) were searched as part of the

methods. We used the definition of a database search from PRISMA-S, ''a literature database designed to search journal literature'' [23]. We did not include searches on single-publisher websites, trial registries, or other types of information source. We excluded articles that solely searched PubMed before May 2020 [24], due to the structural changes to the PubMed database in May 2020 [25]. We excluded scoping reviews, evidence maps, and other review types that use iterative search methods. We had no restrictions to the study design and the study question of the primary studies the systematic review had included.

Titles and abstracts were screened in duplicate using Covidence [21] to assess whether articles appeared to meet the definition of a systematic review. If at least one reviewer (M.L.R., T.J.B., C.P., or P.M.) agreed that the studies met the eligibility criteria, they were included. This liberal screening method, as used in Page et al. [7], helps to account for the difficulty in determining whether something is a ''true'' systematic review, as no single definition of a systematic review exists. After the title/abstract screening, we randomly ordered the remaining articles using Microsoft Excel's RAND function syntax [26]. We reviewed the full text of studies in the randomly ordered list in duplicate using all inclusion and exclusion criteria until 100 studies were identified for inclusion. Conflicts were resolved through discussion between the two reviewers and a third reviewer where needed. We opted to use a sample of 100 articles to provide a workable set for database search reproductions.

2.2. Data collection: data extraction phase

Data from each article in the sample were extracted in duplicate using a custom extraction form in Covidence [21]. Before beginning data extraction, reviewers (M.L.R., T.J.B., C.P., and P.M.) trained together using a set of systematic reviews outside the study sample to ensure there was a shared understanding of each data extraction field. A copy of the complete data extraction form is available in the project's OSF repository [20]. Pairs of reviewers extracted information on the details of the search and search reporting (e.g., databases searched, number of results per database, database search dates, use of limits/restrictions, presence of a PRISMA flow diagram, etc.) and adherence to the six PRISMA-S items noted in Table S1 [20]. The six PRISMA-S criteria included those directly related to database searches: database name (Item 1), multi-database searching (Item 2), full search strategies (Item 8), limits and restrictions (Item 9), dates of searches (Item 13), and total records (Item 15). Data were extracted from the published article and supplementary materials; protocols were not considered. Consensus was achieved through discussion where necessary. Using the extracted data, we calculated database search and systematic review search adherence to the six PRISMA-S items. We considered a database search to be reproducible if it meets six PRISMA-S items, can be re-executed without editing in the named database and platform, and the number of results retrieved is within 10% of the original search results. We considered a systematic review search to be reproducible if these elements are met for all databases searched.

2.3. Data collection: reproduction stage

To conduct the search reproductions and to extract data relating to them, one assessor (M.L.R.) extracted data and conducted the initial reproduction. A second assessor (T.J.B., C.P., or J.R.) validated the data and reproduction. To conduct each reproduction, we used the same databases and platforms as used by the original systematic review teams used, where known. At this stage, we excluded database searches conducted in Japanese and Chinese platforms that required non-Latin characters or operated differently depending on geographical location. We also excluded database searches when we were unable to locate an assessor with access to the database and platform specified. Each search was reproduced by copying and pasting the search directly from the original article when possible. We also reproduced searches that were incomplete, but were described well enough to reconstruct a full search strategy. We anticipated that expertise would be required to reproduce many searches. If it was necessary to apply expert knowledge (e.g., selecting the appropriate platform or database to search) or if errors were evident and required fixing, these interventions were undertaken and noted.

We applied the same database limits as the original search, and searches were restricted to database records entered on or before the date of the last search, as reported by the systematic review authors. If only the month was reported, the search was re-executed with a date limit set on the last day of the reported month. Estimated search dates were used for any search date that was unclearly reported. For databases or platforms without the capacity to limit by database entry date, publication date limits were applied to remove recent records from the search results.

The assessor captured the details of the original search, plus details of the reproduced search, including assumptions, the complete reproduction search strategy, notes on where expert knowledge was applied, details of the database and platform used, search date, and number of results retrieved. In addition, we captured screenshots of each search for additional records. Second assessors added additional observations. Standardized data were captured using a Qualtrics form after the completion of each reproduction and validation [20]. If there was not enough detail to reproduce a search, even with expert knowledge, the search was considered impossible to reproduce.

2.4. Data analyses

To test for differences between the original search (as reported) and the reproduced search, we determined how

many records we retrieved vs. how many had been previously reported. We calculated the percentage difference between the original and reproduction searches (difference in number of results identified/number of results identified by the original search ×100%). If expert knowledge was required to rerun the search, we did not calculate the difference due to our inability to assess whether the reproduction was accurate. For many database searches, the number of results per database was not reported; for these, we could not assess the difference between the initial search and its reproduction. We used a 10% or less difference in results as an indicator of reproducibility to allow for known variability of database results over time and the potential for unclear search date reporting [27–29]. Data were collected at the database search level for all outcomes and at the systematic review level for search and systematic review characteristics. Binary (yes/no) data on PRISMA-S and the reproduction difference for each database search were collated for each systematic review to determine systematic review search outcomes.

2.5. Study protocol amendments

Minor changes were made to data collection elements between protocol registration and data collection. We clarified when database searches would not be reproduced. In addition, we transitioned to providing search reproduction documentation notes on some variables instead of capturing standardized data. During the analysis, we added an additional secondary outcome, the number of database searches meeting individual elements needed for reproducibility, as collecting these data was required to fulfill the other outcomes. The final protocol is available on the OSF Project site [20]. We conducted a post-hoc sensitivity analysis to analyze whether changing the required percentage difference (predetermined at 10%) would alter results of our primary outcome (number of reproducible systematic reviews) or secondary outcome (number of database searches meeting individual elements needed for reproducibility).

3. Results

The search retrieved 8,905 results; after removing duplicates, 8,640 results remained for title/abstract screening. Four thousand one hundred twenty-four systematic review articles remained after title/abstract screening and were randomly ordered for full-text screening. One hundred and sixty-three articles were reviewed in full text before 100 articles that met all eligibility criteria were identified. The 100 articles represented 78 different journals (Table S2 [20]). All articles except one were published in 2021. Most reported searching three or more databases (91%; 91/100) and presented PRISMA flow diagrams (99%; 99/100). Twenty-three percent (23/100) did not provide a search strategy. Forty-four percent (44/100) had a registered or published study protocol (Table S3 [20]).

3.1. Reporting of PRISMA-S items

The final set of 100 systematic reviews contained 453 database searches (range: 1−14 databases; median: four databases per article). Of those, complete database information, including naming the database and platform (PRISMA-S item 1), was available for 47.2% (214/453) (Table 1). Only 4.9% (22/453) database searches clearly reported all six PRISMA-S items. Least commonly reported were item 9, limits and restrictions, and item 13, dates of searches. Limits and restrictions were fully reported for 22.1% (100/453) of database searches, and the exact date of the search was provided for 22.7% (103/453) database searches.

Six (6%; 6/100) systematic reviews clearly reported all six PRISMA-S items for at least one database search (Table 1) and five of these six systematic reviews reported all PRISMA-S items for more than one database. Only two (2%; 2/100) systematic reviews fully reported all six PRISMA-S items for all of their searched databases. The most commonly fully reported PRISMA-S item was item 2 (multi-database searching) for which either a ‘‘yes’’ or ‘‘not applicable’’ response was considered meeting criteria (Table S1 [20]). ‘‘Not applicable,’’ which meant that they did not conduct a search in a platform where multiple databases could be searched simultaneously, applied to 41 of the 45 systematic review articles meeting criteria, leaving only four systematic reviews which clearly reported conducting a multi-database search. The second most commonly reported was item 15, total records, where 43 (43%; 43/100) articles reported the number of results for each database searched for the systematic review. Conversely, only nine (9%; 9/100) systematic reviews reported item 1 (database name) for all databases searched.

3.2. Reproduction

Of the 453 database searches, 64.2% (291/453) provided enough information for us to attempt a reproduction. We were unable to run 3.8% (17/453) of the searches, leaving 60.5% (274/453) remaining searches to execute (Fig. 1). 31.3% (142/453) searches required expert knowledge to recreate, most commonly selecting the platform to use (70.4%; 100/142). 64.1% (91/142) of the searches requiring expert knowledge needed intervention in multiple aspects of the search. These interventions included selecting a platform or database(s); fixing, adapting, or adding Boolean logic; adding limitations; fixing search syntax; selecting fields to search; and/or other changes. Detailed notes, including descriptions of all interventions taken, and screenshots for each attempted and/or completed reproduction are available on the OSF Project site [20].

Table 1. Number and percentage of database searches and systematic review searches meeting criteria, as well as the number of systematic review searches with one or more database searches meeting criteria

	Database searches meeting criteria (n = 453)		Systematic reviews with one or more database searches meeting criteria (n = 100)		Systematic review searches meeting criteria (n = 100)	
	n	%	*n*	%	*n*	%
PRISMA-S Items						
Item 1: Database name	214	47.2%	90	90.0%	9	9.0%
Item 2: Multi-database searching	368	81.2%	99	99.0%	45	45.0%
Item 8: Full search strategies	133	29.4%	40	40.0%	15	15.0%
Item 9: Limits and restrictions	100	22.1%	21	21.0%	20	20.0%
Item 13: Dates of searches	103	22.7%	25	25.0%	19	19.0%
Item 15: Total records	226	49.9%	58	58.0%	43	43.0%
All six PRISMA-S Items	22	4.9%	6	6.0%	2	2.0%
Reproducibility						
Reproduction results within 10% of original	47	10.4%	22	22.0%	1	1.0%
Total						
All seven criteria (PRISMA-S and Reproduction)	16	3.5%	6	6.0%	1	1.0%

There were 453 total database searches in 100 systematic review searches. For example, 22 systematic reviews had one or more database searches with reproduction results within 10% of the original, but only one systematic review search had all database searches with reproduction results within 10% of the original.

Abbreviation: PRISMA-S, Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension.

Results were not reported for all database searches. We were able to compare the number of results of 88 reproduced searches (from 39 systematic reviews) to the number of results retrieved by the original database searches. Forty-seven (10.4%; 47/453) database searches were able to be reproduced within 10% of the number of results from the original search; 3 (0.7%; 3/453) reproductions matched the results numbers exactly. For 41 database search reproductions, the results varied more than 10% between original and reproduction. Six of these searches differed by more than 1,000% between the originally reported number of results and the reproduction, although most (73.2%; 30/41) varied positively or negatively between 10.1% and 100.0% (Fig. 2). Twenty-two percent (22/100) of systematic reviews had one or more database searches that were able to be reproduced within 10% of the originally reported results, and only one systematic review [30] had all of its database searches reproduced within 10% of the originally reported results (Table 1). This paper by Nguyen et al. also fully reported all six PRISMA-S items, thus being the only fully reproducible systematic review in our study [30].

The post-hoc sensitivity analysis examining the impact of altering the threshold of 10% difference in results showed minor differences in outcomes. A 5% maximum difference reduced the number of reproducible systematic reviews to zero, and a 50% maximum difference increased it to two reproducible systematic review searches (Table S4 [20]).

3.3. Characteristics of searches: guidelines and librarian/information specialist involvement

We examined the characteristics of searches to identify commonalities in reproducible and irreproducible searches. Although there were few clear-cut commonalities in either group, some trends were visible. For example, database searches with 10% or less difference fulfilled the criteria for on average 4.9 PRISMA-S items. Database searches where it was impossible to attempt a reproduction only fulfilled the criteria for on average 1.7 PRISMA-S items (Table S5 [20]). Librarians or information specialists were co-authors on 14.9% (7/47) of searches with 10% or less difference in results, but also co-authored 23.5% (4/17) of database searches that would not run. Similarly, 27.7% (13/47) systematic reviews which we reproduced with 10% or less difference mentioned or acknowledged librarians or information specialists, but they were also mentioned or acknowledged in 41.2% (7/17) of the 17 searches which would not run (Fig. 3).

3.4. Characteristics of searches: search strategy location and format

The location and format of the search strategies differed between reproducible and irreproducible searches. Fully reproducible database searches and those

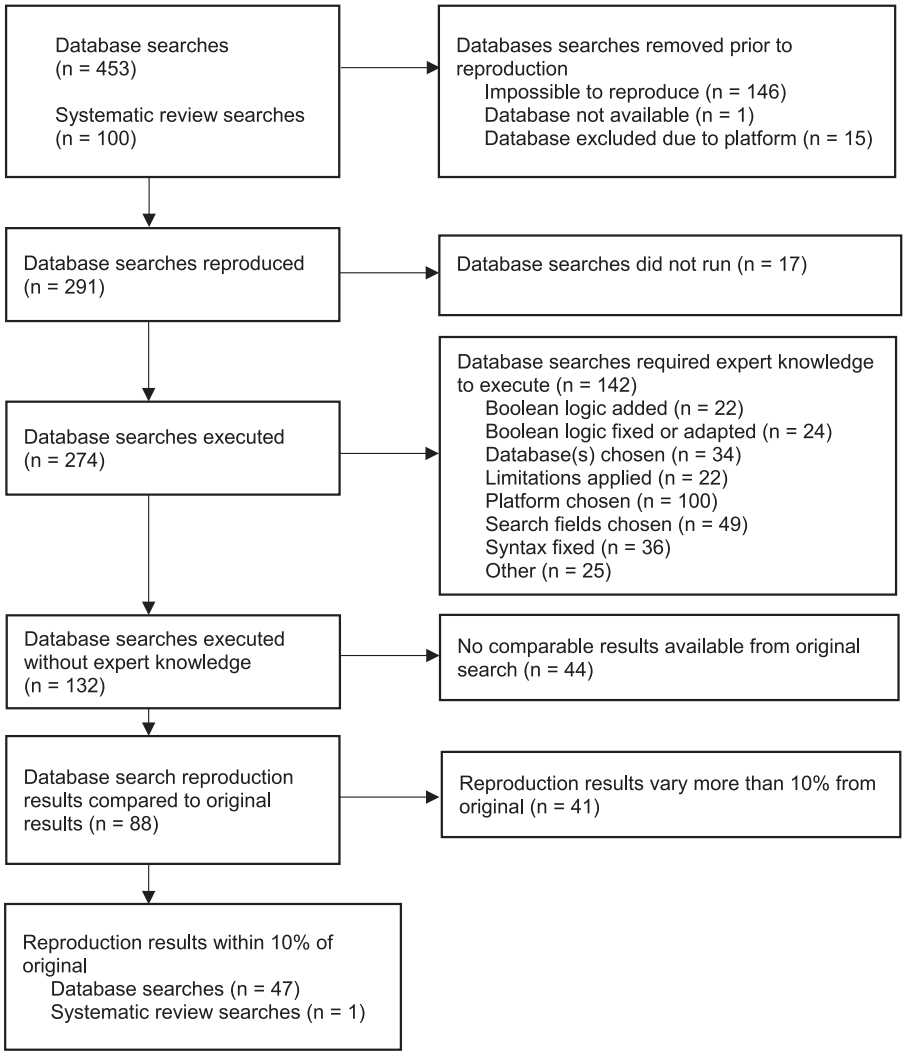

Fig. 1. Flow diagram of database search reproductions.

with a difference less than 10% between the original and the reproduction were located in the supplementary materials most of the time (68.8% [11/16] and 71.4% [35/47], respectively). With the exception of two database searches, all other searches in these categories were published in an appendix to the article. For irreproducible searches, there was more variation in where the search or search description was located. None of the systematic reviews in our sample published their search strategies in a repository (Table S5 [20]).

Searches that were specific to an individual database were more likely to be reproduced with 10% or less difference in results (Fig. 4). Only four (8.5%) of the 47 searches with 10% or less difference were generic, or designed for more than one database search. Comparatively, 48.6% (69/142) of database searches that required expert knowledge to reproduce were generic. 81.3% (13/16) of the fully reproducible searches and 63.8% (30/47) of the searches reproduced with less than 10% difference were multiline searches, meaning that terms or concepts were searched on separate lines and combined later. Single-line searches, on the other hand, were commonly used in database searches that required expert knowledge to conduct (57.7%; 82/142) and those with more than 10% difference in results (65.9%; 27/41).

3.5. Characteristics of searches: errors

Overall, 56.0% (163/291) of all database searches contained at least one error. Errors did not necessarily impact the reproducibility of the search. For example, although all 17 database searches that did not run had errors, and indeed major errors that impacted the ability to conduct the search, 62.5% (10/16) of the fully reproducible database searches had errors as well. The lowest prevalence of errors was for database searches with results variance less than 10% (36.2%; 17/47). Unintentional spelling errors were present, but uncommon, occurring in 3.8% (11/291) of the database searches for which

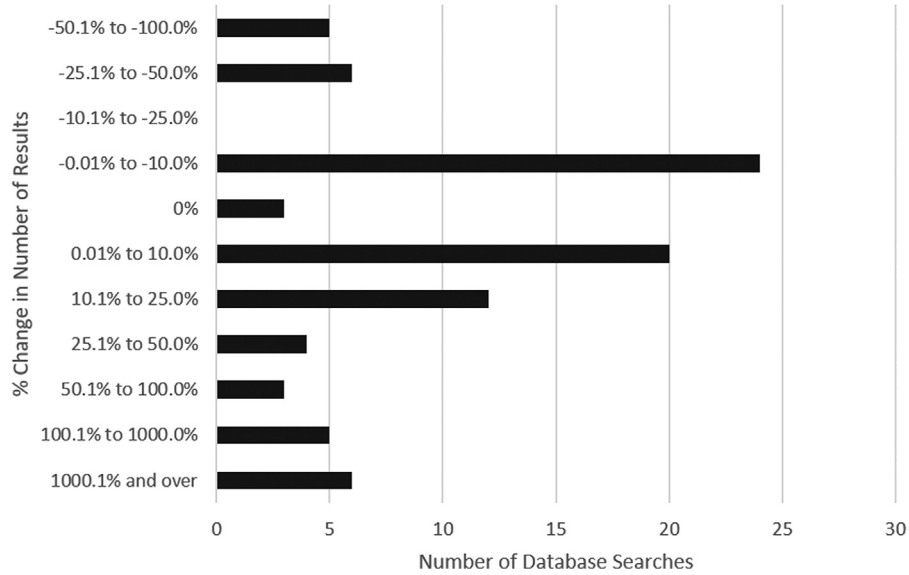

Fig. 2. Percentage change between original and reproduced database search results (n = 88).

reproduction was attempted. Most common were "other" errors, which occurred in 36.1% (105/291) of database searches (Table 2). "Other" errors largely included Boolean logic errors, such as missing parentheses or phrasing, but also included examples of duplicative lines, terms, or phrases that were not in the database's index, mislabeled line numbers, using incorrect or erroneous field codes, and not using all lines, among others.

4. Discussion

4.1. Principal findings

Only one systematic review search in our sample was fully reproducible, with all database searches for the review meeting six key PRISMA-S reporting criteria and reproduced with 10% or less difference in the number of results between the originally reported search and the reproduction. Our

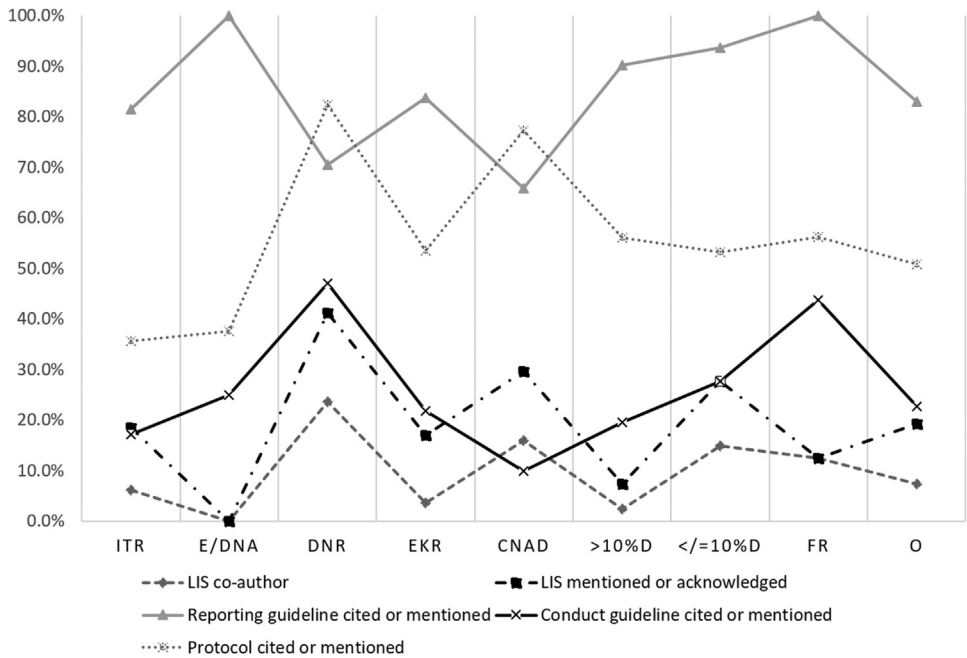

Fig. 3. Percentage of database searches with systematic review characteristics by reproduction status. Abbreviations: ITR, Impossible to reproduce (n = 146); E/DNA, Excluded or database not accessible (n = 16); DNR, Did not run (n = 17); EKR, Expert knowledge required (n = 142); CNAD, Could not assess difference (n = 44); >10%D, Greater than 10% difference (n = 41); </ = 10%D, Less than or equal to 10% difference (n = 47); FR, Fully reproducible (n = 16); O, Overall (n = 453).

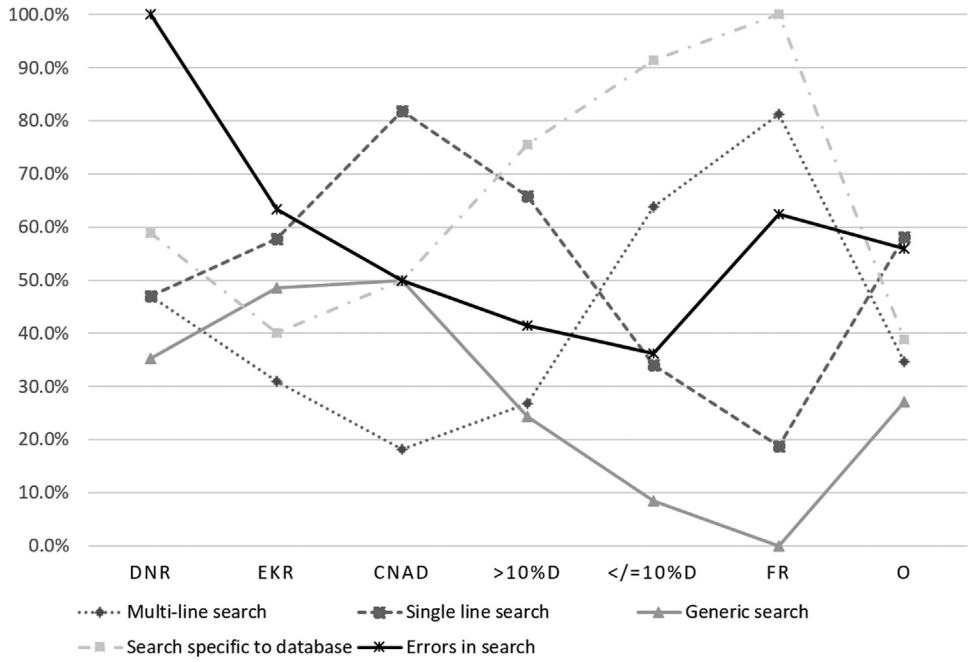

Fig. 4. Percentage of database searches with search strategy characteristics by reproduction status. Abbreviations: DNR, Did not run (n = 17); EKR, Expert knowledge required (n = 142); CNAD, Could not assess difference (n = 44); >10%D, Greater than 10% difference (n = 41); </= 10%D, Less than or equal to 10% difference (n = 47); FR, Fully reproducible (n = 16); O, Overall (n = 453).

findings show major challenges remain for systematic review search reporting, confirming prior research [4,5,10,11,13−16,31−34]. Because our study used PRISMA-S items and attempted to reproduce each search as exactly as possible, our findings are unique and showcase a far grimmer picture of search reproducibility and reporting than has been shown before [6,10,35]. Although our findings do not point to a single primary reason for lack of reproducibility, one of the main reasons for irreproducibility is that authors simply do not provide the information required about the databases/platforms they use. Furthermore, peer review and editorial oversight fails to correct the issue.

4.2. Findings in relation to other studies

This study is the first to examine both systematic review search and database search reproducibility and to attempt reproduction of all database searches within a sample of systematic reviews. At present, ours is the only study with a robust estimate of systematic review search reproducibility based not only on proxy measures but also on actual reproduction results [5,10,11,13−16,32,36−40]. In addition, only two other published studies have looked at PRISMA-S item compliance; neither of those studies used the full criteria of each PRISMA-S item to assess compliance [41,42]. In addition, both looked at compliance with all 16 PRISMA-S items, which meant that systematic reviews that did not use some types of searching methodology were penalized. Our study, on the other hand, only assessed compliance with the six PRISMA-S items that would be necessary for any systematic review containing at least one database search.

4.3. Study limitations

The small sample size (100 systematic reviews) limits generalizability and limits statistical testing of associations and differences between subgroups (e.g., Cochrane vs. non-Cochrane systematic reviews). Reproducing a larger sample of searches would be resource-intensive and, based on prior research [5,10,11,13−16,32,36−40,43], unlikely to produce significantly different results. Search reproducibility and reporting are not necessarily correlated with search quality or with systematic review outcomes. We did not study search quality, risk of bias, or the impact of searches on outcomes. We were limited by our access to databases and platforms; the reproduction team did not always have the same versions of resources, meaning our reproduction and validation numbers predictably varied [27,44]. For example, a search conducted in CINAHL, CINAHL with Full Text, and CINAHL Complete, all on the EBSCOhost platform, will produce slightly different results. There is an urgent need for database consistency and stability within and across platforms.

We did not compare whether the same records were retrieved, nor did we analyze whether studies included in the systematic review would differ between the original search and the reproduction. If the included studies would have been impacted, it is still possible that this would not affect the conclusions and pooled results of the systematic review [45,46]. Few systematic reviews in our sample

Table 2. Types of errors found in published database searches by reproduction status

	Did not run		Expert knowledge required		Could not assess variation		>10% variation		</= 10% variation		Fully reproducible database searches		Overall	
	n = 17		*n* = 142		*n* = 44		*n* = 41		*n* = 47		*n* = 16		*n* = 291	
Type of error	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Typographical	2	11.8%	26	18.3%	3	6.8%	4	9.8%	8	17.0%	7	43.8%	50	17.2%
Spelling	1	5.9%	5	3.5%	1	2.3%	0	0.0%	2	4.3%	2	12.5%	11	3.8%
Syntax	8	47.1%	34	23.9%	16	36.4%	8	19.5%	2	4.3%	2	12.5%	70	24.1%
Other	13	76.5%	56	39.4%	9	20.5%	9	22.0%	9	19.1%	9	56.3%	105	36.1%

contained the full names of the database used, even when the platform name was provided, so our reproductions and validations were based on what we had access to rather than attempting the search in all possible variations. Finally, we excluded database searches requiring Chinese or Japanese characters, although only one of those database searches could potentially have been reproduced with the data provided.

This study is intended to benchmark systematic review search reproducibility, although PRISMA 2020 and PRISMA-S were in their first year since release [3,17,47]. Therefore, we acknowledge that the systematic reviews in our samples likely would neither have been subject to the guidance in PRISMA 2020 or PRISMA-S from the journals they submitted to nor would the authors necessarily be aware of updated guidance, although their research was in progress. That being said, PRISMA 2020 and PRISMA-S both build on PRISMA 2009's guidance for search reporting [3,17,47,48]. Interestingly, the one item that achieved the most compliance in systematic reviews was completely new to the reporting guidance, namely tracking the number of results per database (PRISMA-S item 15). When PRISMA 2020 and PRISMA-S become more widely adopted, we would hope that future studies of search reproducibility will find systematic reviews and database searches with more transparent and complete reporting.

4.4. Study implications

As systematic reviews and clinical practice guidelines based upon them continue to proliferate, so does research waste. Even after Ioannidis described the "mass production of redundant, misleading, and conflicting systematic reviews," systematic reviews continue to increase in popularity [49]. Hoffmann et al. documented the growth of systematic reviews, noting that almost 80 systematic reviews were published each day by 2019 [50]. From our search to identify systematic reviews indexed in MEDLINE in a single month in 2021, we found approximately 135 systematic reviews per day, a substantial increase in 2 years. Unfortunately, the improvement in systematic review search transparency or reproducibility seems to trail behind and irreproducible, poorly reported, and poorly conducted systematic reviews continue to be published [4–8,35]. This directly impacts patients and the public through an influx of irreproducible systematic reviews and clinical practice guidelines.

One of the primary reasons for including full search strategies in systematic review publications is to reduce research waste by enabling the reuse of prior published strategies to update systematic review findings [1,51]. If searches are not reproducible or require significant expertise to re-execute, there is little point in attempting to update the systematic review. Starting a fresh search from scratch may be the only option in these cases.

Systematic reviews are not the only type of research facing a reckoning with poor reporting and lack of reproducibility [18,52]. Studies of the reproducibility of shared data and code have similarly shown that, despite journal policies on sharing, published research often remains irreproducible [53,54]. As calls for data sharing in systematic reviews increase and gain traction [6,55], we hope that systematic review searches will be acknowledged as essential data and code to preserve, document, and share.

5. Conclusion

Systematic review searches should be reproducible, but too many are not. To correct this will require a multifaceted response from searchers, systematic review teams and authors, peer reviewers, journal editors, and database providers. Using reporting guidelines as intended, particularly PRISMA-S and PRISMA 2020, can help guide authors and searchers on best practices for transparent reporting.

CRediT authorship contribution statement

Melissa L. Rethlefsen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tara**

J. Brigham: Investigation, Methodology, Validation, Writing – review & editing. **Carrie Price:** Investigation, Methodology, Validation, Writing – review & editing. **David Moher:** Methodology, Supervision, Writing – review & editing. **Lex M. Bouter:** Methodology, Supervision, Writing – review & editing. **Jamie J. Kirkham:** Formal analysis, Methodology, Supervision, Writing – review & editing. **Sara Schroter:** Methodology, Supervision, Writing – review & editing. **Maurice P. Zeegers:** Formal analysis, Methodology, Supervision, Writing – review & editing.

Data availability

All data in this study are publicly available on an OSF Project site [20], including raw and summarized data. This includes the results of all data extraction, citations to included and not included studies, and screenshots and documentation on the process of reproducing and validating search strategies. All data are shared via Rethlefsen ML, Brigham TJ, Price C, Moher D, Bouter LM, Kirkham JJ, Schroter S, and Zeegers MP. Reproducibility of systematic review searches. OSF; 2023. https://doi.org/10.17605/OSF.IO/UGNCT.

Declaration of competing interest

No specific funding was received for this study. It is part of M.L.R.'s self-funded PhD project registered at Maastricht University, the Netherlands, in collaboration with the BMJ, United Kingdom. M.L.R. and D.M. are co-authors of PRISMA-S, which is used in this study. T.J.B., C.P., L.M.B., and M.P.Z. have no competing interests. S.S. is a full-time employee of BMJ. J.J.K. is a statistical reviewer for BMJ. D.M. is an editorial board member of Journal of Clinical Epidemiology.

Acknowledgments

We give special thanks to Patti McCall (Health Science Center Libraries, George A. Smathers Libraries, University of Florida) for her contributions to screening, data extraction, and PRISMA-S adherence evaluation. We also are grateful to John Reynolds (Louis Calder Memorial Library, University of Miami) for his work validating search reproductions. We thank Tom Roper (formerly Brighton and Sussex University Hospitals NHS) for alerting us to the now defunct Healthcare Databases Advanced Search (HDAS) platform that was used in one of the systematic reviews. Finally, we acknowledge the contributions of Nikki Dettmar (University of Washington), Kjell Johnson (Østfold University College), Marte Ødegaard (University of Oslo), and Judy Wright (University of Leeds), each of whom conducted a search reproduction.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.111229.

References

[1] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.0. 2019. Available at https://www.training.cochrane.org/handbook. Accessed May 4, 2023.
[2] Institute of Medicine (U.S.), Committee on Standards for Systematic Reviews of Comparative Effectiveness Research, Eden J. Finding what works in health care : standards for systematic reviews. Washington, D.C.: National Academies Press; 2011.
[3] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
[4] Abbott R, Bethel A, Rogers M, Whear R, Orr N, Shaw L, et al. Characteristics, quality and volume of the first 5 months of the COVID-19 evidence synthesis infodemic: a meta-research study. BMJ Evid Based Med 2021;27:169–77.
[5] de Kock S, Stirk L, Ross J, Duffy S, Noake C, Misso K. Systematic review search methods evaluated using the Preferred reporting of items for systematic reviews and meta-analyses and the risk of bias in systematic reviews tool. Int J Technol Assess Health Care 2020;37:e18.
[6] Nguyen PY, Kanukula R, McKenzie JE, Alqaidoom Z, Brennan SE, Haddaway NR, 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;379:e072428.
[7] Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med 2016;13(5):e1002028.
[8] Uttley L, Quintana DS, Montgomery P, Carroll C, Page MJ, Falzon L, et al. The problems with systematic reviews: a living systematic review. J Clin Epidemiol 2023;156:30–41.
[9] Koffel JB. Use of recommended search strategies in systematic reviews and the impact of librarian involvement: a cross-sectional survey of recent authors. PLoS One 2015;10:e0125931.
[10] Koffel JB, Rethlefsen ML. Reproducibility of search strategies is poor in systematic reviews published in high-impact pediatrics, cardiology and surgery journals: a cross-sectional study. PLoS One 2016;11:e0163309.
[11] Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, Brigham TJ. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. J Clin Epidemiol 2015;68:617–26.
[12] Sampson M, McGowan J, Tetzlaff J, Cogo E, Moher D. No consensus exists on search reporting methods for systematic reviews. J Clin Epidemiol 2008;61:748–54.
[13] Yoshii A, Plaut DA, McGraw KA, Anderson MJ, Wellik KE. Analysis of the reporting of search strategies in Cochrane systematic reviews. J Med Libr Assoc 2009;97:21–9.

[14] Golder S, Loke Y, McIntosh HM. Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. J Clin Epidemiol 2008;61:440−8.
[15] Maggio LA, Tannery NH, Kanter SL. Reproducibility of literature search reporting in medical education reviews. Acad Med 2011;86:1049−54.
[16] Schellinger J, Sewell K, Bloss JE, Ebron T, Forbes C. The effect of librarian involvement on the quality of systematic reviews in dental medicine. PLoS One 2021;16:e0256833.
[17] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39.
[18] National Academies of Sciences Engineering and Medicine (U.S.), Committee on Reproducibility and Replicability in Science, National Academies of Sciences Engineering and Medicine (U.S.), Nuclear and Radiation Studies Board, National Academies of Sciences Engineering and Medicine (U.S.), Board on Research Data and Information, National Academies of Sciences Engineering and Medicine (U.S.), Board on Mathematical Sciences and Analytics. Reproducibility and replicability in science. Washington, DC: National Academies Press; 2019.
[19] Rethlefsen ML, Moher D, Bouter LM, Kirkham JJ, Schroter S, McCall-Wright P, et al. Reproducibility of systematic review search strategies: a descriptive study. OSF Registries 2021. https://doi.org/10.17605/OSF.IO/KBVSR.
[20] Rethlefsen ML, Brigham TJ, Price C, Moher D, Kirkham JJ, Bouter LM, et al. Reproducibility of systematic review searches. OSF Framework 2023. https://dx.doi.org/10.17605/OSF.IO/UGNCT.
[21] Covidence. Melbourne, Australia: Veritas Health Innovation; 2022. Available at http://www.covidence.org. Accessed May 4, 2023.
[22] Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4(1):1.
[23] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. Table 2 glossary. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Available at https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-020-01542-z/tables/2. Accessed May 4, 2023.
[24] Chan J. PubMed updates and retirement of the legacy site. NLM Tech Bull 2020;436:e6.
[25] Canese K, Chan J, Collins M, Trawick B, Weis S. The new and improved PubMed is here. NLM Tech Bull 2020;434:e2.
[26] RAND function. Microsoft. Available at https://support.microsoft.com/en-gb/office/rand-function-4cbfa695-8869-4788-8d90-021ea9f5be73. Accessed October 19, 2023.
[27] Burns CS, Nix T, Shapiro RM 2nd, Huber JT. MEDLINE search retrieval issues: a longitudinal query analysis of five vendor platforms. PLoS One 2021;16:e0234221.
[28] Garcia-Puente M, Pastor-Ramon E, Agirre O, Morán JM, Herrera-Peco I. Research note. Open letter to the users of the new PubMed: a critical appraisal. Prof Inf 2020;29(3):e290336.
[29] Sisson A, Ouellette L. It's worse than you think: significant search function unreliability in the major medical databases. In: SCC/MLA Annual Meeting. vMeeting: South Central Chapter of the Medical Library Assocation; 2021.
[30] Nguyen PY, Astell-Burt T, Rahimi-Ardabili H, Feng X. Green space quality and health: a systematic review. Int J Environ Res Public Health 2021;18(21):11028.
[31] Lopez-Nicolas R, Lopez-Lopez JA, Rubio-Aparicio M, Sánchez-Meca J. A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000-2020). Behav Res Methods 2021;54:334−49.
[32] Sampson M, McGowan J. Errors in search strategies were identified by type and frequency. J Clin Epidemiol 2006;59:1057−63.
[33] Brunskill A, Hanneke R. The case of the disappearing librarians: analyzing documentation of librarians' contributions to systematic reviews. J Med Libr Assoc 2022;110:409−18.
[34] Helbach J, Pieper D, Mathes T, Rombey T, Zeeb H, Allers K, et al. Restrictions and their reporting in systematic reviews of effectiveness: an observational study. BMC Med Res Methodol 2022;22:230.
[35] Hameed I, Demetres M, Tam DY, Rahouma M, Khan FM, Wright DN, et al. An assessment of the quality of current clinical meta-analyses. BMC Med Res Methodol 2020;20:105.
[36] Meert D, Torabi N, Costella J. Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. J Med Libr Assoc 2016;104:267−77.
[37] Mullins MM, DeLuca JB, Crepaz N, Lyles CM. Reporting quality of search methods in systematic reviews of HIV behavioral interventions (2000-2010): are the searches clearly explained, systematic and reproducible? Res Synth Methods 2014;5(2):116−30.
[38] Salvador-Oliván JA, Marco-Cuenca G, Arquero-Avilés R. Errors in search strategies used in systematic reviews and their effects on information retrieval. J Med Libr Assoc 2019;107:210−21.
[39] Toews LC. Compliance of systematic reviews in veterinary journals with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) literature search reporting guidelines. J Med Libr Assoc 2017;105:233−9.
[40] Faggion CM Jr, Huivin R, Aranda L, Pandis N, Alarcon M. The search and selection for primary studies in systematic reviews published in dental journals indexed in MEDLINE was not fully reproducible. J Clin Epidemiol 2018;98:53−61.
[41] Norling B, Edgerton Z, Bakker C, Dahm P. The quality of literature search reporting in systematic reviews published in the urological literature (1998-2021). J Urol 2023;209:837−43.
[42] Sadeghi-Ghyassi F, Zarea Gavgani V, Fathifar Z, Makani N, Vaez R, Montazeri M. Quality of reporting of literature search strategies in systematic reviews published on the role of telehealth during COVID-19. J Inf Sci 2022. https://doi.org/10.1177/01655515221087649.
[43] Pieper D, Heß S, Faggion CM Jr. A new method for testing reproducibility in systematic reviews was developed, but needs more testing. BMC Med Res Methodol 2021;21:157.
[44] Barrick K, Riegelman A. Phrasing in reproducible search methodology: the consequences of straight and curly quotation marks. Coll Res Libr 2020;82(7):978.
[45] Alqaidoom Z, Nguyen PY, Awadh M, Page MJ. Impact of searching clinical trials registers in systematic reviews of pharmaceutical and non-pharmaceutical interventions: Reanalysis of meta-analyses. Res Synth Methods 2023;14(1):52−67.
[46] Djulbegovic B, Ahmed MM, Hozo I, Koletsi D, Hemkens L, Price A, et al. High quality (certainty) evidence changes less often than low-quality evidence, but the magnitude of effect size does not systematically differ between studies with low versus high-quality evidence. J Eval Clin Pract 2022;28:353−62.
[47] Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160.
[48] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6(7):e1000100.

[49] Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016; 94(3):485−514.
[50] Hoffmann F, Allers K, Rombey T, Helbach J, Hoffmann A, Mathes T, et al. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol 2021;138:1−11.
[51] Nguyen P-Y, McKenzie JE, Hamilton DG, Moher D, Tugwell P, Fidler FM, et al. Systematic reviewers' perspectives on replication of systematic reviews: a survey. Cochrane Evid Synth Methods 2023;1(2):e12009.
[52] Cobey KD, Fehlmann CA, Christ Franco M, Ayala AP, Sikora L, Rice DB, et al. Epidemiological characteristics and prevalence rates of research reproducibility across disciplines: a scoping review of articles published in 2018-2019. Elife 2023;12:e78518.
[53] Stodden V, Guo P, Ma Z. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. PLoS One 2013;8:e67111.
[54] Crüwell S, Apthorp D, Baker BJ, Colling L, Elson M, Geiger SJ, et al. What's in a badge? A computational reproducibility investigation of the open data badge policy in one issue of psychological science. Psychol Sci 2023;34:512−22.
[55] Page MJ, Nguyen PY, Hamilton DG, Haddaway NR, Kanukula R, Moher D, 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.