Sleep Medicine Reviews 66 (2022) 101708

Contents lists available at ScienceDirect

Sleep Medicine Reviews

journal homepage: www.elsevier.com/locate/smrv

Data reproducibility issues and their potential impact on conclusions from evidence syntheses of randomized controlled trials in sleep medicine

Chang Xu [a, b], Suhail A.R. Doi [c], Xiaoqin Zhou [d], Lifeng Lin [e], Luis Furuya-Kanamori [f], Fangbiao Tao [a, b, *]

[a] Ministry of Education Key Laboratory for Population Health Across-life Cycle & Anhui Provincial Key Laboratory of Population Health and Aristogenics, Anhui Medical University, Anhui, China
[b] School of Public Health, Anhui Medical University, Anhui, China
[c] Department of Population Medicine, College of Medicine, QU Health, Qatar University, Doha, Qatar
[d] Department of Clinical Research Management, West China Hospital, Sichuan University, China
[e] Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, USA
[f] UQ Centre for Clinical Research, Faculty of Medicine, The University of Queensland, Brisbane, Australia

ARTICLE INFO

Article history:
Received 17 May 2022
Received in revised form
6 October 2022
Accepted 11 October 2022
Available online 19 October 2022

Keywords:
Reproducibility
Data extraction errors
Evidence synthesis practice
Sleep medicine

ABSTRACT

In this study, we examined the data reproducibility issues in systematic reviews in sleep medicine. We searched for systematic reviews of randomized controlled trials published in sleep medicine journals. The metadata in meta-analyses among the eligible systematic reviews were collected. The original sources of the data were reviewed to see if the components used in the meta-analyses were correctly extracted or estimated. The impacts of the data reproducibility issues were investigated. We identified 48 systematic reviews with 244 meta-analyses of continuous outcomes and 54 of binary outcomes. Our results suggest that for continuous outcomes, 20.03% of the data used in meta-analyses cannot be reproduced at the trial level, and 43.44% of the data cannot be reproduced at the meta-analysis level. For binary outcomes, the proportions were 14.14% and 40.74%. In total, 83.33% of the data cannot be reproduced at the systematic review level. Our further analysis suggested that these reproducibility issues would lead to as much as 6.52% of the available meta-analyses changing the direction of the effects, and 9.78% changing the significance of the P-values. Sleep medicine systematic reviews and meta-analyses face serious issues in terms of data reproducibility, and further efforts are urgently needed to improve this situation.

1. Introduction

Systematic reviews and meta-analyses consist of the most important source of high-quality evidence to support decision-making in modern healthcare practice [1]. To ensure the quality of systematic reviews and meta-analyses, a standard and rigorous process is required to be undertaken by review authors [2]. Amongst these processes, data extraction is one of the most important steps; it transfers data from the original source (e.g., clinical trials) to the systematic reviews for further synthesis, quantitatively or qualitatively. Any reproducibility issues in the data extraction would ultimately impact the results and could even produce misleading conclusions, regardless of the robustness of the other steps.

Given the importance of data extraction, current guidelines for systematic reviews have recommended the application of good practice strategies (e.g., double extraction) to ensure the reliability of the data [3]. Even so, in evidence synthesis practice, errors frequently occur during the data extraction process and pose a serious source of bias that threatens the validity of the results. A replication study on 34 Cochrane systematic reviews found that

Abbreviations: CI, Confidence interval; IQR, Interquartile range; MD, Mean difference; OR, Odds ratio; PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses; RCT, Randomized controlled trial; RR, Risk ratio; SD, Standard deviation; SMD, Standardized mean difference: SMD.

* Corresponding author. Ministry of Education Key Laboratory for Population Health Across-life Cycle, Anhui Medical University, Anhui, China.
E-mail address: fbtao@ahmu.edu.cn (F. Tao).

https://doi.org/10.1016/j.smrv.2022.101708
1087-0792/

almost 58.5% of them had data extraction errors [4]. A more recent reproducibility study analyzed data from 829 meta-analyses included in 201 systematic reviews and found that 66.8% of the meta-analyses and 85.1% of these reviews had erroneous data. These errors led to changes in the direction of the effects in 3.5% and the significance of the P-value in 6.6% of the meta-analyses, respectively [5].

There has been an increasing number of systematic reviews in the field of sleep medicine during the past decades. Technical development of these systematic reviews mainly focused on design, methods, and reporting [6,7], while researchers seldom paid attention to the reproducibility of the data, especially for those undertaking quantitative syntheses (meta-analysis). Understanding data reproducibility issues and the potential impact on the results will have implications for evidence-based practice and decision-making. Therefore, in this study, based on our previous work on methodological weaknesses [7], we went a step further and examined data reproducibility issues of the systematic reviews in this area.

2. Methods

2.1. Protocol development

A protocol for this study was developed prior to formulating the implementation (See Appendix file). The amendment includes an additional subgroup analysis that was added by effect estimate and type of error on the impact of the pooled result during the data collection, according to our findings in another work [5]. The reporting of the current study was based on the PRISMA 2020 statement whenever possible [8].

2.2. Inclusion criteria

Systematic reviews of randomized controlled trials (RCTs) on the topic of sleep medicine that contained one or more pairwise meta-analyses were eligible for inclusion [9]. To ensure the feasibility of the reproducibility process, only those that provided mean, standard deviation (SD), and sample size for continuous outcomes and those that provided event counts and sample size for binary outcomes in both comparison groups were considered. Overviews, scoping reviews, and narrative reviews were not considered. Pooled analyses that did not use a comprehensive (at least one database) literature search were also not considered. In some situations, authors may present their paper as an original study combined with a meta-analysis; again, such publications were not considered in the current study.

2.3. Literature search

We searched for systematic reviews published in academic journals of sleep medicine that were indexed in PubMed, Medline, or Embase databases from inception to 22nd October 2019. This was done in three steps: 1) we first identified 23 journals of sleep medicine from SCImago Journal & Country Rank (https://www.scimagojr.com/); 2) then, we checked Beall's list and identified four predatory journals, which were excluded; 3) finally, we searched the above three databases for related publications from the remaining 19 journals.

The literature search was done by the lead author, a senior methodologist of evidence-based medicine. We did not update the literature search, did not consider grey literature, and did not employ hand searching on the reference list of eligible reviews. This was because a representative set of studies is sufficient for reporting on the characteristics of data reproducibility issues in systematic reviews. A detailed description of the literature search, along with the search strategy, has been recorded in our previous work [7].

2.4. Literature screen

Considering that different databases could overlap for some records, before the literature screen, we used the EndNote X7 software to find out potential duplicates. Then, one review author screened the records in terms of the titles and abstracts; those obviously out of the scope were removed in this step. The full texts of the remaining records were screened by two review authors independently. Any disagreements were discussed until a consensus was reached. Cohen's kappa statistic was used to estimate inter-rater agreement [10].

2.5. Data extraction

Data extraction was conducted by one author (Z.XQ) with a self-checking process and further double-checked by the lead author (C.X.). A recent RCT has shown that a single extraction with verification was comparable to the double extraction method [11]. The systematic review level, meta-analysis level, and study level information were collected, which included: year of publication of the systematic review, outcome of each meta-analysis, intervention types (pharmaceutical vs. non-pharmaceutical) of each outcome, synthesis methods used in the meta-analysis (e.g., random-effects model), effect estimates used in the meta-analysis (e.g., mean difference [MD], standardized mean difference [SMD], odds ratio [OR], risk ratio [RR]), the data used in the meta-analysis (i.e., mean, SD, and group sample size for continuous outcomes and event counts and sample size for binary outcomes in the two arms) of each included study, publication year and citation of each included study within a meta-analysis. For continuous outcomes, since they generally involve missing data estimation (i.e., missing mean or SD), the estimation method was also collected if reported. The above data were collected based on the reporting of the systematic reviews.

2.6. Data reproduction

The original sources (i.e., full texts of the RCT, supplementary files, public registry) of each included study were examined to see if the components used in the meta-analysis were correctly extracted. For continuous outcomes, missing data estimation was also reproduced due to the nature of the data structure; this was done by using the same estimation method reported by the review authors, or using the methods recommended by the Cochrane Handbook (versions 4.2.6 to 6.3) [12] if the review authors failed to report such information.

Two types of data reproducibility issues can be defined: *data extraction errors* and *data estimation errors*. When the data presented in the meta-analysis was not the same as the data reported in the original sources, we treated it as having a *data extraction error* [5]. When our estimate for missing statistical information was different from the review authors' estimation, we treated it as having a *data estimation error*. The former happens in both binary and continuous outcomes, while the latter only in continuous outcomes.

According to our recent work [5], errors during data acquisition could be due to the following five mechanisms: numerical error, ambiguous data error, mismatching error, zero assumption error, and misidentification error. Here, we focused on the first four, which belong to *data extraction errors*. The last one is not a type of

extraction error; instead, it is a type of identification error that refers to whether a study should be included or not [5].

For *data estimation errors*, one special case is when researchers used standard deviation as standard error, which further led to an estimation error. To highlight this type of error, we listed it separately as a *misconception error*.

In case the review authors claimed they had contacted the principal author and successfully obtained the original data, even though there was discordance, we treated the review as having no errors. Data reproduction was conducted by the same two authors, again, first extracted by one author and double-checked by the lead author. Table 1 presents the definition of the two types of errors.

2.7. Missing data

Missing data on error identification occurs when the original sources of the studies cannot be obtained. We expected the missing data rate would be small and would not impact the results. To examine this assumption, we conducted a sensitivity analysis by removing those meta-analyses with missing data and re-estimating the main outcomes.

2.8. Statistical analysis

Basic characteristics of the meta-analyses were summarized descriptively using proportions or median value and interquartile range (IQR). The primary outcome of interest was the prevalence of errors at the study level, meta-analysis level, and systematic review level. For the meta-analysis level, the prevalence of errors was estimated as the number of meta-analyses with at least one study having errors against the total number of meta-analyses. Similarly, for the systematic review level, the prevalence of errors was calculated as the number of systematic reviews with at least one meta-analysis having errors against the total number of systematic reviews. The percentages of each type of error defined above among the total errors were also of interest.

The secondary outcome of interest was the potential impact of various errors on the results of the meta-analyses. This was done by comparing the synthesized results by using the error-addressed data to the results reported by the reviews, using the same effect estimates and methods. For those meta-analyses with ambiguous data errors, the potential impact could not be investigated as the real data was unclear. To measure the degree of the impact, we predefined a tolerable bias limits on the effect estimate of 20% and 50% as cut-offs of a moderate impact and large impact, which was calculated as $\left|\frac{\widehat{\theta}_{\text{corrected}}-\widehat{\theta}_{\text{orginal}}}{\widehat{\theta}_{\text{orginal}}}\right| \times 100\%$, where $\widehat{\theta}$ are the estimated pooled effects [13]. The direction of the effect and the significance of the P-value were also compared.

Subgroup analyses were pre-specified and first performed for intervention types of the trials, namely, pharmaceutical vs. non-pharmaceutical, with the relative difference in the error rate measured by OR and its 95% confidence interval (CI) owing to its "portability" property [14]. This is because our previous study suggested that data extraction errors may differ by intervention types [5]. Subgroup analysis on the impact of the pooled results was examined based on effect estimator type (e.g., MD vs. SMD, OR vs. RR) and error type (e.g., single error vs. mixed errors), again, according to our previous findings [5]. Sensitivity analysis was done for missing data which has been described earlier.

Considering the substantial difference in the data structure and mechanism for errors (e.g., missing data estimation) for binary and continuous data, we summarized the above results separately for these two types of data. Data analysis was undertaken using Stata/SE 16.0 (Stata Crop LCC, College Station, TX), with alpha = 0.05 as the significance level.

3. Results

From 1,630 records through the initial search, 353 systematic reviews were identified, with 161 involving healthcare interventions [7], and 87 that were based on RCTs that were further screened. After exclusions of those without complete data (n = 39), 48 were ultimately eligible for inclusion, for which 43 contained continuous outcomes and 22 contained binary outcomes. Within the 48 systematic reviews, there were 244 meta-analyses of continuous outcomes with 1,448 trials in total and 54 meta-analyses of binary outcomes with 403 trials in total that were eligible for the analysis (Figure S1). The inclusion list and exclusion list (with reasons) are presented in the Appendix.

3.1. Basic characteristics

Tables 2 and 3 present the basic data characteristics. About two-thirds of the systematic reviews (n = 29, 60.41%) were published after 2015. Of the 244 meta-analyses of continuous outcomes, 97 (39.75%) focused on pharmaceutical interventions, and 147 (60.25%) were on non-pharmaceutical interventions. For effect estimates, MD was used in most of the meta-analyses (n = 173, 70.90%), while SMD was only used in 29.10% (n = 71). The median number of RCTs included in each meta-analysis was 3 (IQR: 2 to 6),

Table 1
Descriptions of the different types of errors during the data extraction (adapted from *BMJ. 2022; 377: e069155*).

Type of errors	Description
Data extraction error	Errors that happened during the data extraction process
1. Numerical error	Extracted numerical values were incorrect. This may be due to typo, calculation error, or extraction of data of another outcome [5].
2. Ambiguous data error	Extracted data could not be reproduced from all available sources (unknown whether it is correct or not) due to ambiguous definitions of the outcomes, while the review authors did not specify how the data were obtained/calculated. In some situations, the outcome(s) even could not be found in the original study and related materials (e.g., supplementary file, ClinicalTrials.gov). [5]
3. Zero-assumption error	This was a special case of ambiguous error, and generally occurs in safety outcomes. The outcome was not reported in the original study and related materials (e.g., supplementary file, ClinicalTrials.gov), while the review authors assumed that no event occurred. [5]
4. Mismatching error	The extracted data were incorrectly matched to the intervention/exposure groups, but the numerical values were correct. This could occur in any cells of the summarized table. [5]
Data estimation error	Errors that happened during the data estimation process
1. Estimation error	The estimation of missing mean value and(or) standard deviation that the estimated values cannot be reproduced using the same estimation method or the standard methods with the same data by the authors.
2. Misconception error	A special type of under data estimation error where researchers confused standard deviation as standard error, which further led to estimation error.

Table 2
Basic characteristics of 244 meta-analyses of continuous outcomes.

Basic characteristics	No. of MA (n = 244)
Type of intervention (MA)	
Pharmaceutical	97 (39.75%)
Non-pharmaceutical	147 (60.25%)
Effect estimators (MA)	
Mean Difference	173 (70.90%)
Standard Mean Difference	71 (29.10%)
Analysis model (MA)	
Random effects	166 (68.03%)
Fixed effects	78 (31.97%)
Other	0 (0.00%)
Number of studies per MA (Median, IQR)	3 (2–6)
2–5	171 (70.08%)
6–10	44 (18.03%)
11 or more	29 (11.89%)
Number of continuous MAs per SR (Median, IQR)	4 (2–7)
1–5	29 (67.44%)
6–10	8 (18.60%)
11 or more	6 (13.95%)
Needs missing data estimation (SR)	
Yes	40 (93.02%)
No	3 (6.98%)
Reported missing data estimation methods (SR, n = 40)	
Yes	12 (30.00%)
No	28 (70.00%)

Note: SR, systematic review; MA, meta-analysis; IQR, interquartile range.

Table 3
Baseline characteristics of 54 meta-analyses of binary outcomes.

Baseline characteristics	No. of MA (n = 54)
Type of Intervention (MA)	
Pharmaceutical	35 (64.81%)
Non-pharmaceutical	19 (35.19%)
Effect estimators (MA)	
Odds ratio	15 (27.78%)
Risk ratio	39 (72.22%)
Analysis model (MA)	
Random effects	40 (74.07%)
Fixed effects	14 (25.93%)
Other	0 (0.00%)
Number of studies per MA (Median, IQR)	5 (3–7)
2–5	28 (51.85%)
6–10	18 (33.33%)
11 or more	8 (14.81%)
Total sample size per MA (Median, IQR)	654.5 (295–1857)
< 500	24 (44.44%)
≥ 500	30 (55.56%)
Number of binary MAs per SR (Median, IQR)	2 (1–3)
1–5	54 (100.00%)
6–10	0 (0.00%)
11 or more	0 (0.00%)

Note: SR, systematic review; MA, meta-analysis; IQR, interquartile range.

and the majority (n = 171, 70.08%) had 5 or fewer RCTs. For the 43 systematic reviews, 40 (93.02%) needed an estimation of the missing mean value or missing SD, or both, while only 12 (30.00%) had reported some details of the estimation methods.

Amongst the 54 meta-analyses of binary outcomes, 27.78% (15/54) utilized the OR as the effect estimator, while the majority (72.22%, 39/54) utilized RR as the effect estimator. Similar to meta-analyses of continuous outcomes, the random-effects model was widely employed for binary meta-analyses (74.07%, 40/54). The median number of studies per meta-analysis was 5 (IQR: 3 to 7); that is, half had 5 or more studies included. However, only 14.81% (8/54) had 11 or more studies. The median sample size per meta-analysis was 654.5 (IQR: 295 to 1857), and about 55.56% (30/54) had sample sizes larger than 500.

The distribution of some of the variables substantially differed from those of continuous outcomes. Specifically, for meta-analyses of binary outcomes, 64.81% (35/54) referred to pharmaceutical interventions, and only 35.19% (19/54) referred to non-pharmaceutical interventions. In contrast, for those of continuous outcomes, more referred to non-pharmaceutical interventions (60.25%). There was also a large difference based on word cloud analysis for the binary and continuous outcomes (Figure S2).

3.2. Reproducibility of the data

3.2.1. Continuous outcomes

For the 1,448 trials, the data on 298 (20.03%) could not be reproduced. Among them, 29.87% (89/298) were due to data extraction errors, 57.05% (170/298) were due to data estimation errors, and 13.09% (39/298) were mixed with both errors. For the error site, 9.73% (29/298) occurred on the mean value, 56.71% (169/298) occurred on the SD, 6.04% (18/298) occurred on the sample size, 16.78% (50/298) occurred on both mean value and SD, 0.34% (1/298) occurred on both mean and sample size, 2.35% (7/298) occurred on the mean value, SD and sample size, and 8.05% (24/298) were not classifiable due to ambiguous or missing information (Fig. 1a).

At the systematic review level, 37 (86.05%) of the 43 systematic reviews had at least one meta-analysis of continuous outcomes with data errors. Of the 244 meta-analyses, 106 (43.44%) had errors on at least one included trial. The median proportion of studies with errors within a meta-analysis was 33.33% (IQR: 20.00%–50.00%).

Subgroup analysis suggested that, for studies with pharmaceutical interventions, there was a higher proportion (22.10% [99/448] vs. 16.54% [134/669]) of errors than those with non-pharmaceutical interventions (OR = 1.43, 95% CI: 1.07 to 1.91; P = 0.015).

3.2.2. Binary outcomes

For the 403 trials, 14.14% (57/403) of the data used in the meta-analysis could not be reproduced, all due to *data extraction errors*. Amongst the errors, numerical error accounted for the most (59.65%, 34/57), followed by ambiguous data error (36.84%, 21/57). Two (3.51%, 2/57) mismatching errors were recorded, and no (0.00%, 0/57) zero-assumption error was recorded. In terms of the components, the majority (82.46%, 47/57) of the errors referred to the erroneous extraction of event counts, 15.79% (9/57) referred to both group sample size and the event counts extraction, and only 1.75% (1/57) referred to the group sample size extraction (Fig. 1b).

For the 54 eligible meta-analyses, there were 22 (40.74%) with data extraction errors in at least one trial. For those with errors, the proportion of studies with data extraction errors within a meta-analysis ranged from 5.88% to 100.00%, with a median of 22.50% (IQR: 14.29%–50.00%). For the 22 systematic reviews, 59.09% (13/22) had data extraction errors in at least one meta-analysis of binary outcomes. When combining the meta-analyses of binary outcomes and continuous outcomes together for a systematic review, the proportion of error at the systematic review level in total was 83.33% (40/48).

Subgroup analysis in terms of the type of interventions suggested that those studies with pharmaceutical interventions were more likely to have the data incorrectly extracted by review authors (15.41% [47/305] vs. 10.20% [10/98]), although the effect was not statistically significant (OR = 0.62, 95% CI: 0.30 to 1.29, P = 0.20).

3.3. Reproducibility of the results

3.3.1. Continuous outcomes

Of the 106 meta-analyses of continuous outcomes with data errors, only 92 could be used to investigate the impacts of the errors on the results with exclusions due to ambiguous data error or

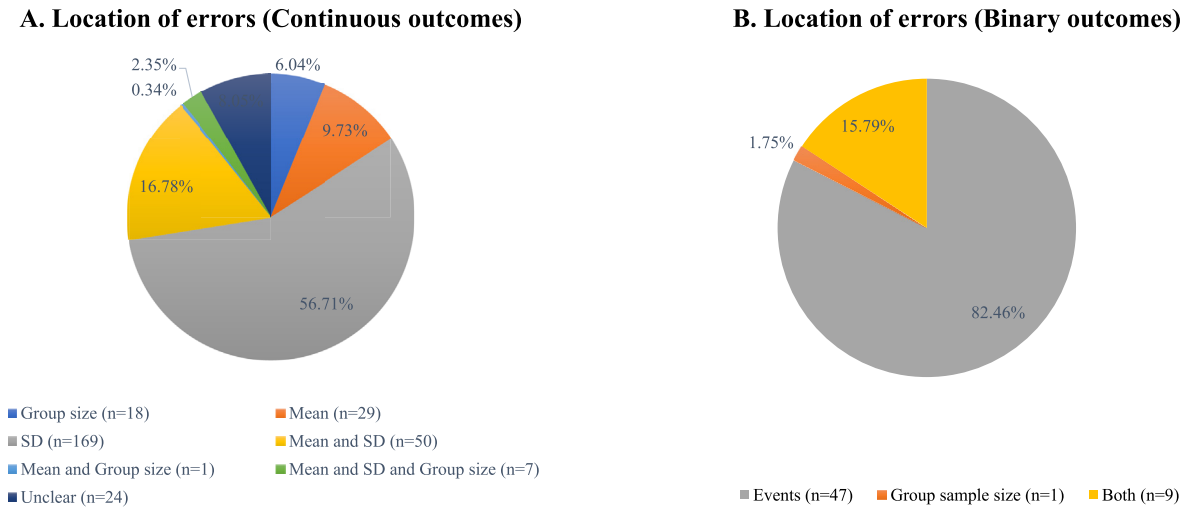

Fig. 1. Location of the errors A) continuous outcomes and B) binary outcomes.

inestimable missing SD in 14 meta-analyses. Our results suggested that, when using our corrected (and reproducible) data, the magnitude of the effects was moderately or largely impacted in 36.96% (34/92) of the meta-analyses, where an absolute bias was larger than 20%; the magnitude of the effects was largely impacted in 17.39% (16/92) of the meta-analyses, where an absolute bias was larger than 50%. In addition, 6.52% (6/92) of the meta-analyses would have the direction of the effects changed to the other side, and 9.78% (9/92) of the meta-analysis would have the significance of the P-value changed at the 0.05 threshold.

For the 67 meta-analyses that used the MD as effect estimates, the proportions with moderate and above impacts, large impacts, changes in the effect direction, and changes in the significance of P-value were 35.82%, 16.42%, 5.97%, and 13.43%, respectively. For the 25 that used the SMD as the effect estimate, the proportions were 40.00%, 20.00%, 8.00%, and 0.00%. These results suggested that the SMD was more susceptible to the impacts on the point estimation (40.00% vs. 35.82%, P = 0.71; 20.00% vs. 16.42%, P = 0.69; 8.00% vs. 5.97%, P = 0.73), while the MD was more susceptible to the impacts on the P-value (13.43% vs 0.00%, P = 0.17); see Fig. 2A.

For each single type of error, we were able to investigate the impacts of *data extraction error* (n = 25), *estimation error* (n = 20), and *misconception error* (n = 13) on the pooled results. For those with *data estimation error*, the proportions with moderate and above impact, large impact, changes in the effect direction, and changes in the significance of P-value were 15.00%, 5.00%, 0.00%, and 5.00%. For those with *data extraction error*, the proportions were 20.00%, 12.00%, 0.00%, and 4.00%; For those with *data misconception error*, the proportions were 84.62%, 46.15%, 15.38%, and 7.69%, which were substantially larger than for other types of errors (Fig. 3).

For the 60 meta-analyses with only one type of aforementioned errors, the proportions with moderate and above impacts, large impacts, changes in the effect direction, and changes in the significance of P-value were 31.67%, 16.67%, 3.33%, and 5.00%. For the 32 with two or more aforementioned errors, the proportions were 46.88%, 18.75%, 12.50%, and 18.75%, which were substantially higher than those with one type of errors (46.88% vs. 31.67%, P = 0.14; 18.75% vs. 16.67%, P = 0.80; 12.50% vs. 3.33%, P = 0.12; 18.75% vs. 5.00%, P = 0.049); see Fig. 3.

3.3.2. *Binary outcomes*

The impacts of data extraction errors on the pooled results were assessed on 16 of the 22 meta-analyses of binary outcomes deemed erroneous, because 6 of them contained ambiguous errors and the true value of the data was unclear to us. All of the 16 meta-analyses involved a single error; 15 contained numerical errors, and 1 contained a mismatching error. Therefore, a subgroup analysis by the type of errors was not feasible. Fig. 2B presents the impacts of the errors on the results.

In total, when using error-corrected data, 25.00% (4/16) of the effects would be moderately or substantially impacted, and 18.75% (3/16) would be substantially impacted. In addition, 6.25% (1/16) of the meta-analyses would have the direction of the effects changed,

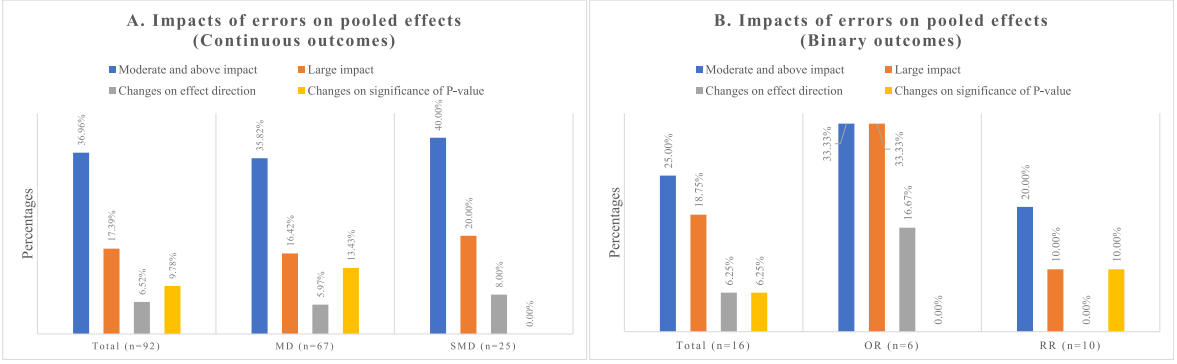

Fig. 2. Impacts of data reproducibility issues on the results by A) continuous outcomes and B) binary outcomes.

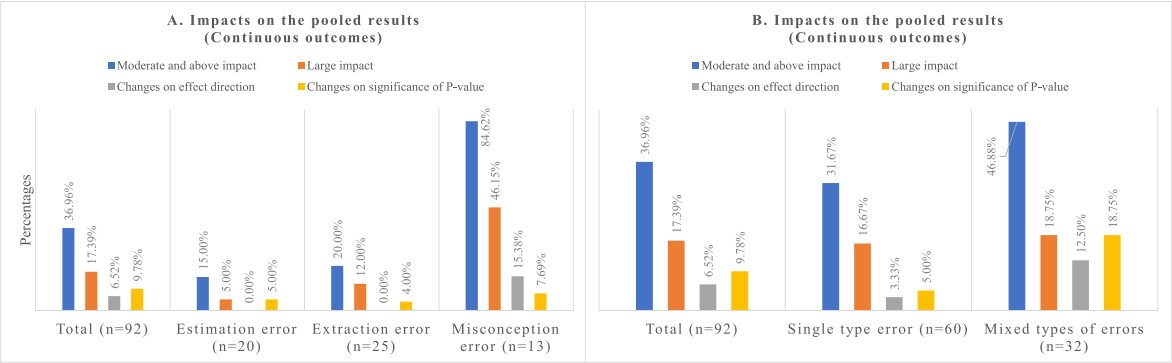

Fig. 3. Impacts of data reproducibility issues on the results by type of errors. A) classified by data extraction error and data estimation error; B) classified by single type of error and mixed type of errors.

and 6.25% (1/16) would have the significance of P-value altered. Subgroup analysis was conducted by the effect measures used. For those using the OR as the effect measure (n = 6), one in three (33.33%, 2/6) of the effects would be substantially impacted, 16.67% (1/6) would have the direction of the effects changed, and there was no altered the significance of the P-value. For those using the RR as the effect estimator (n = 10), one in five (20.00%, 2/10) of the effects would be moderately or substantially impacted, none (0.00%, 0/10) had the direction of the effects changed, and 10.00% (1/10) had an altered the significance of the P-value.

3.4. Sensitivity analysis

Missing information on data errors occurred in 16 (1.08%) trials for meta-analyses of continuous outcomes due to the limited access to the original source. By removing the meta-analyses with missing information, the proportion of errors did not show substantial changes: 15.69% (179/1,141) for the trial level, 41.13% (95/231) for the meta-analysis level, and 82.5% (33/40) for the systematic review level. Missing information on data errors occurred in 5 (1.24%) trials for meta-analyses of binary outcomes, and again, there had no substantial changes by removing those with missing information: 11.02% (40/363) for the trial level, 38.46% (20/52) for the meta-analysis level, and 55.00% (11/20) at the systematic review level.

4. Discussion

In this study, we have investigated data reproducibility issues in published systematic reviews and the potential impacts on the pooled results of the meta-analyses. Our results suggest that for continuous outcomes, one-fifth of the data used in meta-analyses in sleep medicine cannot be reproduced at the trial level, almost half (43%) of the data cannot be reproduced at the meta-analysis level, and more than three-quarters (86%) of the data cannot be reproduced at the systematic review level. These reproducibility issues led to just under 1 in 10 (6.5%) of the meta-analyses to present changes in the direction of the effects, and 1 in 10 (9.8%) of the meta-analysis had the significance of the P-value changed, and those meta-analyses with two or more types of errors are more susceptive to the changes. Similar findings were observed for meta-analyses of binary outcomes.

The proportion of errors at the study level for meta-analysis of binary outcomes was smaller than those of continuous outcomes (14.14% vs. 20.03%). This could be expected, as in our parallel study of continuous outcomes, errors not only occurred during data extraction but also occurred in data estimation (missing value estimation). Actually, at the meta-analysis level, the proportion of meta-analyses with errors was similar in these two different data types (40.74% vs. 43.44%). There was a much higher proportion of errors for continuous outcomes with regard to the systematic review level; again, this can be partly explained by the additional type of errors. On all accounts, ultimately, the impacts of these errors on the conclusions were comparable for meta-analyses of binary and continuous outcomes.

The findings of the current study largely concur with our previous investigation on systematic reviews of safety [5]. The proportions at the trial level and systematic review level were similar, although the current study had a lower proportion than our previous study (43.4% vs. 66.8%) at the meta-analysis level. The proportion at the systematic review level of this study was also higher than the study by Jones et al. [4] (83.3% vs. 58.5%) and the study by Gøtzsche et al. [15] (83.3% vs. 63.0%), where continuous outcomes were examined. This may be due to the inherent variation among different topics. The subgroup analysis of the current study on the type of interventions reinforced our previous finding that data on trials with pharmaceutical interventions are more likely to be erroneously reproduced than non-pharmaceutical interventions.

In terms of the impacts, the results of our study were discordant with the previous conclusion: in a methodological review, Tim et al. summarized studies on data extraction errors and concluded that such errors had a minor impact on the results [16]. One important mechanism could be that the impact of data reproducibility issues largely depends on the number of studies, total sample size, as well as the proportion of such issues within a meta-analysis. For example, in the current study, the median number of studies per meta-analysis was 3 for continuous outcomes and 5 for binary outcomes, so a difference in the data in one of the studies may lead to a substantial change in the results.

In the subgroup analysis, studies of pharmaceutical interventions were more likely to have the data incorrectly extracted, regardless of binary outcomes or continuous outcomes. In another investigation of ours on systematic reviews of safety, the same phenomenon was observed [5]. The mechanism behind it was unclear; it could be due to the definition and reporting of the outcomes varying among pharmaceutical and non-pharmaceutical interventions. We hypothesize that outcomes of pharmaceutical interventions are more diverse and complex. However, there is currently no direct evidence to support our hypothesis. Further in-depth investigations are worthwhile to be implemented.

For meta-analyses of continuous outcomes, we noticed that data extraction error had a larger impact than data estimation error, while data misconception error had the largest impact on the former two. This is beyond our expectation that estimation error did not matter much, as it is the most frequent type of error and can be largely influenced by different estimation methods. One possible explanation could be that data estimation error often occurs on

standard deviation rather than the mean value, and therefore it may have a larger impact on the significance. This hypothesis can be reinforced by our results that both data estimation error and data misconception error showed a higher proportion of meta-analyses that changed the significance of the P value compared with data extraction error. Unfortunately, due to the small sample size, we are unable to investigate the impact of each single error on the results for binary outcomes.

In our study, we found that the SMD was more susceptible to the impacts of data reproducibility issues on the point estimation, while the MD was more susceptible to the impacts on the P-value. This could be partly explained by the properties of the two estimates — the magnitude of SMD could be largely influenced by the SD while the MD does not [12,17,18]. This is also the reason that the Cochrane Handbook cautioned the interpretation of SMD [12]. Another reason might be due to the totally different samples and weighting schemes used in the two such that those meta-analyses which use the MD tend to have a higher proportion of pharmaceutical interventions (42.2% vs. 338%) and a smaller proportion that use the random-effects model (61.9% vs. 83.1%). Such differences were also observed for OR and RR, while the sample size was too small to support a conclusive finding. Due to the above reasons, the findings of the current study were unable to determine which of the two estimates is better.

The implications of the current findings are: 1) it is recommended that at least one statistician to be included as co-authors on further systematic reviews where studies are synthesized mathematically. 2) authors should check the data carefully and ensure no errors of the information and provide the details the extracted data (e.g., location of the information, formulars of the estimation of missing data) to allow for spot checking of the data and results.3) A pilot training for data extractors should be employed and a duplicate extraction strategy should be strictly implemented. 4) peer reviewers should check the reproducibility of the data when reviewing systematic reviews. 5) a statistical review of the study results should be considered by journals to assess reproducibility before publication. 6) An implementation guideline should be developed to formulate the data extraction and estimation of systematic reviews.

To the best of our knowledge, this is the first study that investigates data reproducibility issues and the impacts on the conclusions from systematic reviews in the field of sleep medicine. The findings of the current study highlight the importance of improving the validity of data extraction and missing data estimation in meta-analyses of binary as well as continuous outcomes. Our findings are expected to provide an in-depth insight into how these reproducibility issues impact the results and the need for further strategies to avoid such issues.

This study should be read in light of the limitations. First is the eligibility criteria since, to ensure the feasibility of the implementation of the current study, we set the eligibility criteria of systematic reviews to be based on RCTs and provision of full summarized data for each included trial. Such limitations would no doubt lead to the loss of representativeness of the findings of this study. However, there might be no better solution for this as the reporting of published systematic reviews is still poor [19,20]. Second, the collection of the systematic reviews was limited to 19 academic journals in the sleep medicine area; some systematic reviews related to sleep medicine might be published in general journals that were not included in this study. As the policy of publication and requirements on the quality differs a lot across journals, the findings of this study may not be well suited to those systematic reviews published in general journals. However, based on available empirical evidence, we believe that the non-reproducibility of the data is a broad and serious issue that applies to the majority of systematic reviews in general. Third, the current study targeted pairwise meta-analyses; we did not consider network meta-analyses or other types of meta-analysis. This is because the mechanisms of data reproducibility issues may differ under various synthesis methodologies and assumptions. Further methodological studies should focus on other types of meta-analysis. Moreover, for the evaluation of data estimation error, it is well-known that different software could generate some differences in the effect estimates and variances; Although the differences would be minor, this may have some mild impact on our results.

5. Conclusions

In conclusion, sleep medicine systematic reviews and meta-analyses face serious issues in terms of the reproducibility of their data extraction and estimation. These reproducibility issues can lead to changes in the conclusions. The impact on the conclusions differs by effect estimates and types of errors. Finally, our findings have implications for current evidence-based guidelines, which may benefit from incorporating guidance specific to these types of data reproducibility issues in systematic reviews and meta-analyses that are reported here.

Practice Points

1. One-fifth of the data used in meta-analyses of continuous outcomes in sleep medicine cannot be reproduced at the trial level, almost half (43%) of the data cannot be reproduced at the meta-analysis level, and more than three-quarters (86%) of the data cannot be reproduced at the systematic review level.
2. In meta-analyses of binary outcomes, data reproducibility issues occurred more than one in ten (14.14%) at the study level, four in ten (40.74%) at the meta-analysis level, and almost six in ten (59.09%) at the systematic review level.
3. These reproducibility issues led to changes in the conclusions of meta-analyses for both binary and continuous outcomes; those meta-analyses with two or more types of errors are more susceptive to the changes.

Research Agenda

1. Future systematic review authors should pay special attention to the validity of the data extraction for meta-analysis to ensure the quality of the evidence synthesis.
2. Considering the high frequencies and the serious impacts, healthcare personnel should pay great attention to potential data reproducibility issues before decision-making, and may rate down the level of the evidence if reproducibility issues are considerable.
3. Current evidence-based guidelines may benefit from incorporating guidance specific to these types of data reproducibility issues in systematic reviews and meta-analyses.
4. Implementation guidelines on data extraction and estimation for meta-analysis are urgently needed.

Data availability statement

The original data can be found at https://osf.io/vzfk7. A copy of the data can also be found at https://www.researchgate.net/publication/364198334_Data_reproducibility_issues_and_their_potential_impact_on_conclusions_from_evidence_syntheses_of_randomized_controlled_trials_in_sleep_medicine?_sg=tsIkiLRbG_jXVr9jaYN3N7EGA1TqMTnhIoL_X3udHPxpNPNiAl77SSSzigo8RR6l1ST7457e5kqPZM6z7jIyUzwkJ1NBRdtVfBA5IWS6.nBlq1oO7dzRmy_xA20ggOFlZ7bPbu0yl8j2feivvvPoiDCqp41x0OAEy93-uAdltcD4xrOhviR2Kqp8ez4zHJg.

Funding

This work was made possible by the National Natural Science Foundation of China (72204003) and a seed fund for talented earlier researchers from Anhui Medical University (9021783201) and program grant #NPRP-BSRA01-0406-210030 from the Qatar National Research Fund (a member of Qatar Foundation). Luis Furuya-Kanamori was supported by an Australian National Health and Medical Research Council Fellowship (APP1158469). The findings herein reflect the work, and are solely the responsibility of the authors. The funding bodies had no role in any process of the study (i.e., study design, analysis, interpretation of data, writing of the report, and decision to submit the article for publication).

Author's contributions

Conception and design: CX, TFB, and SD; Manuscript drafting: CX; Data collection: CX, ZXQ; Data analysis and result interpretation: CX, ZXQ; Statistical guidance: SD and LL; Methodology guidance: SD and LFK; Manuscript editing: CX, LL, SD, LFK. All authors have read and approved the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.smrv.2022.101708.

References

[1] Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature 2018;555(7695):175–82.
[2] Chandler J, Cumpston M, Thomas J, Cumpston M, Li T, Page MJ, et al. Chapter I: introduction (updated February 2021). Cochrane. In: Higgins JPT, Thomas J, Chandler J, et al., editors. Cochrane Handbook for systematic reviews of interventions version 6.2; 2021. Available from, www.training.cochrane.org/handbook.
[3] Li T, Higgins JPT, Deeks JJ. Chapter 5: collecting data (updated February 2022). Cochrane. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for systematic reviews of interventions version 6.3; 2022. Available from, www.training.cochrane.org/handbook.
[4] 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(7):741–2.
[5] Xu C, Yu TQ, Furuya-Kanamori L, Lin L, Zorzela L, Zhou XQ, et al. Validity of data extraction in evidence synthesis practice of adverse events: reproducibility study. BMJ 2022;377:e069155.
[6] Condon HE, Maurer LF, Kyle SD. Reporting of adverse events in cognitive behavioural therapy for insomnia: a systematic examination of randomised controlled trials. Sleep Med Rev 2021;56:101412.
[7] Xu C, Furuya-Kanamori L, Kwong JSW, Li S, Liu Y, Doi SA. Methodological issues of systematic reviews and meta-analyses in the field of sleep medicine: a meta-epidemiological study. Sleep Med Rev 2021;57:101434.
[8] 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.
[9] Xu C, Zhou X, Zorzela L, Ju K, Furuya-Kanamori L, Lin L, et al. Utilization of the evidence from studies with no events in meta-analyses of adverse events: an empirical investigation. BMC Med 2021;19(1):141.
[10] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
[11] Li T, Saldanha IJ, Jap J, Smith BT, Canner J, Hutfless SM, et al. A randomized trial provided new evidence on the accuracy and efficiency of traditional vs. electronically annotated abstraction approaches in systematic reviews. J Clin Epidemiol 2019;115:77–89.
[12] Deeks JJ, Higgins JPT, Altman DG. Chapter 10: analysing data and undertaking meta-analyses (updated February 2022). Cochrane. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for systematic reviews of interventions version 6.3; 2022. Available from, www.training.cochrane.org/handbook.
[13] Xu C, Ju K, Lin L, Jia P, Kwong JSW, Syed A, Furuya-Kanamori L. Rapid evidence synthesis approach for limits on the search date: how rapid could it be? Res Synth Methods 2022;13(1):68–76.
[14] Doi SA, Furuya-Kanamori L, Xu C, Lin L, Chivese T, Thalib L. Controversy and Debate: questionable utility of the relative risk in clinical research: paper 1: a call for change to practice. J Clin Epidemiol 2022;142:271–9.
[15] Gøtzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. JAMA 2007;298(4):430–7.
[16] Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol 2017;17(1):152.
[17] Lin L, Aloe AM. Evaluation of various estimators for standardized mean difference in meta-analysis. Stat Med 2021;40(2):403–26.
[18] Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials (Technical Support Document in Evidence Synthesis; No. TSD2). National Institute for Health and Clinical Excellence; 2011. Available from, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.662.2162&rep=rep1&type=pdf.
[19] Kahale LA, Diab B, Brignardello-Petersen R, Agarwal A, Mustafa RA, Kwong J, et al. Systematic reviews do not adequately report or address missing outcome data in their analyses: a methodological survey. J Clin Epidemiol 2018;99:14–23.
[20] Spineli LM, Pandis N, Salanti G. Reporting and handling missing outcome data in mental health: a systematic review of Cochrane systematic reviews and meta-analyses. Res Synth Methods 2015;6(2):175–87.