An empirical comparison of the harmful effects for randomized controlled trials and non-randomized studies of interventions

Minhan Dai[1], Luis Furuya-Kanamori[2], Asma Syed[3], Lifeng Lin[4] and Qiang Wang[1]*

[1]Mental Health Center, West China Hospital of Sichuan University, Chengdu, China, [2]School of Public Health, Faculty of Medicine, The University of Queensland, Herston, QL, Australia, [3]Department of Population Medicine, College of Medicine, Qatar University, Doha, Qatar, [4]Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ, United States

Introduction: Randomized controlled trials (RCTs) are the gold standard to evaluate the efficacy of interventions (e.g., drugs and vaccines), yet the sample size of RCTs is often limited for safety assessment. Non-randomized studies of interventions (NRSIs) had been proposed as an important alternative source for safety assessment. In this study, we aimed to investigate whether there is any difference between RCTs and NRSIs in the evaluation of adverse events.

Methods: We used the dataset of systematic reviews with at least one meta-analysis including both RCTs and NRSIs and collected the 2 × 2 table information (i.e., numbers of cases and sample sizes in intervention and control groups) of each study in the meta-analysis. We matched RCTs and NRSIs by their sample sizes (ratio: 0.85/1 to 1/0.85) within a meta-analysis. We estimated the ratio of the odds ratios (RORs) of an NRSI against an RCT in each pair and used the inverse variance as the weight to combine the natural logarithm of ROR (lnROR).

Results: We included systematic reviews with 178 meta analyses, from which we confirmed 119 pairs of RCTs and NRSIs. The pooled ROR of NRSIs compared to that of RCTs was estimated to be 0.96 (95% confidence interval: 0.87 and 1.07). Similar results were obtained with different sample size subgroups and treatment subgroups. With the increase in sample size, the difference in ROR between RCTs and NRSIs decreased, although not significantly.

Discussion: There was no substantial difference in the effects between RCTs and NRSIs in safety assessment when they have similar sample sizes. Evidence from NRSIs might be considered a supplement to RCTs for safety assessment.

1 Introduction

Randomized controlled trials (RCTs) are considered the most unbiased study design and represent the current gold standard for assessment of efficacy of interventions (Guyatt et al., 2008). Through the randomization process, RCTs would mostly avoid the bias of confounding factors by indicating the intervention effect (Shrier et al., 2007). However, RCTs are expensive, and thus most RCTs only cover a small number of patients with a short follow-up period (Van Spall et al., 2007; Kennedy-Martin et al., 2015). In addition, sample size estimates for RCTs are usually based on the main outcome, that is, efficacy, rather than adverse events. This makes it challenging to assess safety outcomes since many outcomes occur at a low frequency—the observed events would be rare and even zero for certain outcomes. Therefore, statistical inference faces significant uncertainty caused by random errors (Bhaumik et al., 2012; Efthimiou, 2018). In addition, recruiting subjects usually involves strict inclusion criteria, and researchers tend to exclude high-risk patients, such as children, elderly people, pregnant women, patients with multiple complications, and those with potential drug interactions. These restrictions limit the representativeness of the findings of RCTs (Chou and Helfand, 2005; Golder et al., 2011).

Non-randomized studies of interventions (NRSIs) are an alternative to overcome the aforementioned issues for assessing safety. It is widely known that a case-control study is designed for when the cases of events are rare (Vandenbroucke and Pearce, 2012). There are two sources of error that could impact the estimates of NRSIs, namely, systematic error (bias) and random error. For effectiveness of intervention, the bias of NRSIs is deemed to be the main effect modifier on the results, and the random error may have limited impacts due to the large sample size and sufficient outcomes (Higgins et al., 2011). Methods such as stratification, matching, and regression analysis have been proposed to address the confounding bias for NRSIs (McNamee, 2005; Austin, 2011). Simulation studies have verified that these methods work well to control the impact of confounders on the effects (Jreich and Sebastien, 2021). However, for rare adverse events, such methods may not be feasible due to the limited number of cases. For example, when the event risk is 1/1000, even for an NRSI with a sample size of 2000, the expected number of cases would only be two, which is insufficient for the aforementioned methods. In such a case, in safety assessment, the random error may have a larger impact than the systematic error (bias), which dominates the results.

One increasingly popular method was to pool all available RCTs of the same topic together, i.e., *via* a meta-analysis, to increase the statistical power, and it has the ability to increase the power in testing whether the true effect actually exists. Nevertheless, the statistical power of these meta-analyses was still seriously insufficient (Jia et al., 2021). Researchers then proposed to include NRSIs in the meta-analysis because, for safety outcomes, the primary aim is to capture any signal of harm (Reeves et al., 2013; Valentine and Thompson, 2013). This is somewhat reasonable as we mentioned previously that for safety outcomes of rare events, systematic error may have a limited impact on the results. Even so, this has raised wide controversy as the concerns about the confounding bias still exist for NRSIs and will be synthesized into the pooled effect (Benson and Hartz, 2000; Concato et al.,

2000; Ioannidis et al., 2001; Abraham et al., 2010; Hemkens et al., 2016; Soni et al., 2019).

To address this concern, we designed an empirical study based on a database of systematic reviews of safety that compared the effects of RCTs and NRSIs to see whether there was any difference in the evaluation of adverse events between them.

2 Materials and methods

The current study findings are reported according to the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) checklist for case-control studies (von Elm et al., 2008). A brief description of the study is as follows. First, we searched for the published systematic reviews of safety and screened for those with safety as exclusive outcomes. Then, we checked the eligible systematic review for those including both RCTs and NRSIs in the meta-analyses. The RCTs and NRSIs were further matched by sample size (1:1) within each meta-analysis. Finally, the effects of each pair of RCT and NRSI were compared.

2.1 Sample size estimation

To ensure a sufficient sample size (pairs) for the statistical test, we used the following formula to estimate the minimum sample size for the current study: $n = (z_{\alpha/2} \times d/E)^2$ (Donner, 1984). Here, E indicates the margin of error and d represents the expected standard deviation of the difference of the effects (i.e., ln odds ratio, lnOR) across the pairs. For the margin of error, it is a concept similar to the bias in a simulation study, namely, how close the estimated effect is to the true effect (Donner, 1984). For the standard deviation, it is a concept similar to the between-study heterogeneity in a meta-analysis (Pateras et al., 2018). Therefore, we took 25% as the tolerable margin of error and 1 as standard deviation, indicating that there would be substantial-to-large heterogeneity across pairs (Ju et al., 2020; Xu et al., 2021a). Based on these parameters, the estimated sample size of the current study is 96.04; that is, we need at least 97 pairs of RCTs and NRSIs to ensure the statistical power to test whether the difference of the effects across the pairs was significant.

2.2 Data source

We used a dataset collected in 2020, which was primarily established to improve the evidence-based practice for safety assessment and has been documented elsewhere (Xu et al., 2021b). The dataset consists of 640 systematic reviews of healthcare interventions published in two time periods (2008–2011 and 2015–2020), with adverse events as exclusive outcomes and at least one meta-analysis. The two different periods were primarily designed for comparing how double-zero studies were dealt with by systematic review authors over time (Xu et al., 2021b). For each time period, a comprehensive literature search was performed to ensure the representativeness of the sample (systematic reviews of safety). A detailed description of the dataset can be found in our previous works (Zorzela et al., 2014; Xu et al., 2021b).

2.3 Eligibility criteria

We screened 640 systematic reviews for those with at least one outcome (each outcome referred to a separate meta-analysis) that included both RCTs and NRSIs in order to compare the effects of NRSI *vs.* RCT. In addition, considering that data extraction error is commonly seen in published meta-analyses, we only considered those providing summarized 2 x 2 table data for each study in the meta-analysis; a further double-checking process for such data through original studies is possible. Based on the same consideration, those reviews directly reporting the effect size (e.g., OR) and standard error for the meta-analysis were not considered; for such systematic reviews, it is impossible to check whether the effect sizes they used were correctly estimated or extracted, especially for NRSIs. We collected RCTs and NRSIs in systematic reviews under the condition that each pair of the RCT and NRSI has the same topic. Thus, the potential impact of different topics on the results was eliminated. In addition, only pairwise meta-analyses were considered to ensure the interventions were homogeneous.

2.4 Data collection

The meta-analytic data of each outcome from each eligible systematic review were extracted by two review authors independently. Any disagreements were solved by discussing with the lead author. These include the 2 x 2 table information (i.e., numbers of cases and sample sizes in intervention and control groups) of each study in the meta-analysis, type of design of each study (i.e., RCT or NRSI), first author of the systematic reviews, and first author and year of publication of included studies. During data extraction, any disagreements were solved by discussion. The primary data were collected from the systematic reviews, and to ensure the quality of the data, we further double-checked the data of matched pairs from the original studies included in the corresponding systematic reviews.

2.5 Data analysis

Previous studies pooled the effects of NRSIs and RCTs by treating them as subgroups in a meta-analysis and compared the pooled effects across each meta-analysis (Mathes et al., 2021). However, this method has a big disadvantage in that it requires a sufficient number of studies (i.e., 10) in each subgroup to ensure the robustness of the pooled effects. Under such a limitation, there would be very few meta-analyses that would meet the requirement and may further impact the generalizability of the findings.

In the current study, in order to compare the potential difference of the effects, we matched RCTs and NRSIs within the same meta-analysis by their sample sizes to control the impact of random error on the effects. In brief, we first calculated the sample size of each study in each meta-analysis and ranked the sample sizes within the meta-analysis. Then, those RCTs and NRSIs with similar sample sizes were matched as a pair, using the "nearest neighbor matching" method (Austin, 2011). To ensure the matched RCT and NRIS have almost the same sample size, we calculated the ratio of their sample size; only those with a ratio from 0.85/1 to 1/0.85 were considered to

avoid the potential influence of sample size on the results (Xu et al., 2021c).

In each pair, the OR and its standard error of the RCTs and NRSIs were estimated as it has been considered one of the optimal effect estimators (Doi et al., 2020; Doi et al., 2021). For those studies with zero events in single or double groups, the continuity correction was applied by adding 0.5 to each cell to produce an approximate evaluation of the OR and its standard error (Xu et al., 2021d). Furthermore, the ratio of the ORs (ROR) of NRSI against RCT was calculated to reflect the deviation of the effects; the ROR is the primary outcome of the current study (Dechartres et al., 2018). This statistics allows us to further test whether there is a difference in the effect of RCTs and NRSIs. When the weighted mean value of the ROR across the pairs is 1, there would be no difference between the effect of RCTs and NRSIs. In order to obtain the weighted mean value of the ROR, we calculated the natural logarithm of ROR (lnROR) and its standard error and then used the inverse variance heterogeneous model to combine these lnRORs (Doi et al., 2015; Doi and Furuya-Kanamori, 2020). The standard error of the lnROR of each pair can be estimated using the SEs for the RCT and NRSI estimates (Golder et al., 2011).

$$SE\left(\ln ROR\right) = \sqrt{\left(SE\ln OR_{rct}\right)^2 + \left(SE\ln OR_{nrsi}\right)^2}.$$

The pooled effect is the weighted mean value. A statistical null hypothesis would be then the pooled lnROR = 0. We used the two-sided t-test with the significant level of alpha = 0.05. Sensitivity analysis was employed by cluster robust error meta-regression to consider the potential correlation of lnRORs for the pairs within each systematic review (Xu and Doi, 2018). Further subgroup analysis by the maximum sample size of each pair was employed to see if the potential difference of the effects varies by sample size. The following five groups were prespecified: 1–50, 51–100, 101–200, 201–500, and >501. Statistical analyses were conducted in MetaXL 5.3 software (EpiGear International, Australia) and Stata 14/SE (Stata, College Station, TX).

3 Results

3.1 Basic characteristics

Of the 640 systematic reviews of adverse events, 87 included both RCTs and NRSIs. We further excluded 12 with the NRSIs only used for incidence of adverse events or did not include both RCTs and NRSIs within a meta-analysis. Of the remaining 75 systematic reviews, 31 were eligible, which had at least one outcome, contained both RCTs and NRSIs, and provided summarized 2 x 2 table data for each study in the meta-analysis (Grootscholten et al., 2008; Sun et al., 2008; Torloni et al., 2009; Touzé et al., 2009; Slobogean et al., 2010; Yaghoobi et al., 2010; Aires et al., 2015; Geng et al., 2015; Ghayoumi et al., 2015; Inokuchi et al., 2015; Wang et al., 2015; Yoon et al., 2015; Zhang and Ma, 2015; Keir et al., 2016; Peng et al., 2016; Vavken et al., 2016; Balasubramanian et al., 2017; Geminiani et al., 2017; Pecorelli et al., 2017; Cheng et al., 2018; Shah et al., 2018; Zhao et al., 2018; Ceresoli et al., 2019; Craveiro et al., 2019; Jiang et al., 2019; Menne et al., 2019; Nagy et al., 2019; Shah et al., 2019; Vaos et al., 2019; Winberg et al., 2019; Yang et al., 2019). The selection process is reported in Supplementary Figure S1, and the characteristics of the included systematic reviews are shown in Supplementary Table S1.
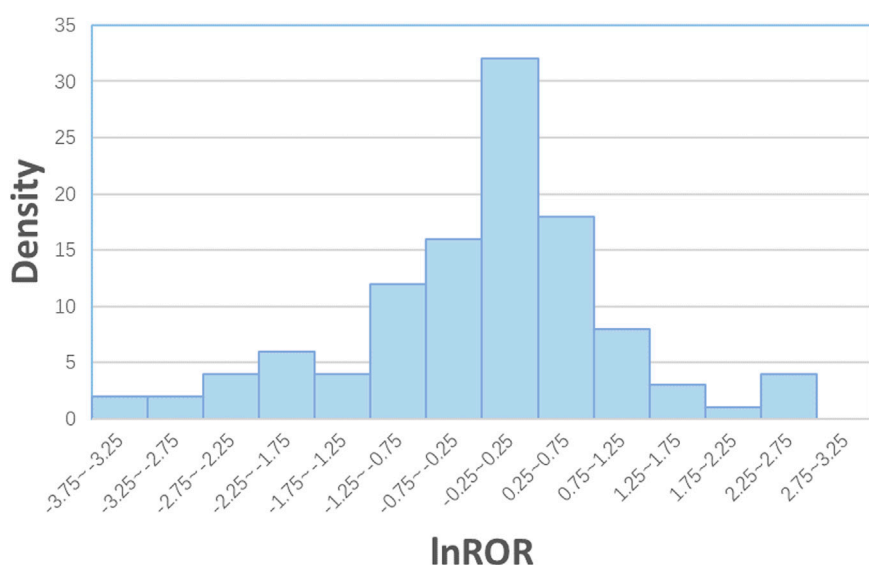
FIGURE 1
Distribution of lnRORs.

From the 31 systematic reviews, 178 meta-analyses contained both RCTs and NRSIs with a total of 1,404 studies. 119 pairs of RCTs and NRSIs were successfully matched for the analysis (Supplementary Figure S1). In a further analysis of the 238 studies from 119 pairs, we recorded two (0.84%) had data extraction errors, which were further addressed by correcting these errors. The sample size of the current study is bigger than the minimum requirement (*Sample size estimation*). Among these 119 pairs, there were 19 (15.97%) with the sample size ranging from 1 to 50, 41 (34.45%) pairs ranging from 51 to 100, 19 (15.97%) ranging from 101–200, 17 (14.29%) ranging from 201–500, and 23 (19.33%) with the sample size >500.

3.2 RCTs *vs.* NRSIs on the effects

Figure 1 shows the distribution of the lnRORs, which has an approximately normal distribution ($p = 0.446$ for skewness and $p = 0.13$ for kurtosis). The unweighted mean value of the lnROR was − 0.14 with a standard deviation of 1.23, and the single-sample t -test showed no substantial difference of lnROR over zero ($t = -1.25$, $p = 0.21$).

Supplementary Figure S2 shows the forest plot of the weighted average lnRORs. Again, no difference was observed between the effects of NRSIs against RCTs. The pooled ROR across the 119 pairs was 0.96 (95% confidence interval [CI]: 0.87, 1.07; $p = 0.49$), with no obvious between-study heterogeneity ($I^2 = 0\%$). A robust meta-regression model that considers the correlation between the pairs within a systematic review showed a similar result, with the pooled ROR as 0.96 (95% CI: 0.90, 1.03; $p = 0.27$).

3.3 Subgroup analysis

Similar conclusions were obtained from the analysis of different sample size subgroups. There was no significant difference between the weighted mean value of lnROR and 0 in each subgroup, that is, there was no significant difference in the effects between RCTs and NRSIs, regardless of sample size. The forest plots of the subgroup analyses are shown in Figure 2. However, there was a slight difference in the absolute value of the weighted mean of lnROR for each sample size subgroup, which decreased lnROR with increasing sample size (Figure 3). With the increase in sample size, the difference between RCTs and NRSIs diminished.

In addition, the treatment used in the original study had no significant effect on the results. We compared the weighted mean of lnROR in the treatment subgroup, and the results of either surgical treatment or drug therapy were close to 0, and there was no significant difference (Supplementary Figure S3).

4 Discussion

In this study, we compared the effects of RCTs and NRSIs on safety assessment based on empirical evidence. Our results showed that there was no significant difference between RCTs and NRSIs in the evaluation of adverse events of the same topic, and there was no significant difference in sample size or treatment subgroups.

In our research, although different sample size subgroups yielded similar results, there was still a slight difference in the weighted average RORs of different sample size subgroups. As shown in Figure 3, with the increase in the sample size, the value of lnROR decreases gradually; that is, the difference between RCTs and NRSIs gradually decreases. This is likely because the random error decreased as the sample size increased, and the estimated effect is therefore closer to the true effect (i.e., lnROR = 0) (Moher et al., 1994; Wang and Ji, 2020). This also indicates that small studies may lead to biased estimation of the effects and should be addressed and interpreted appropriately in further original studies as well as meta-analyses.
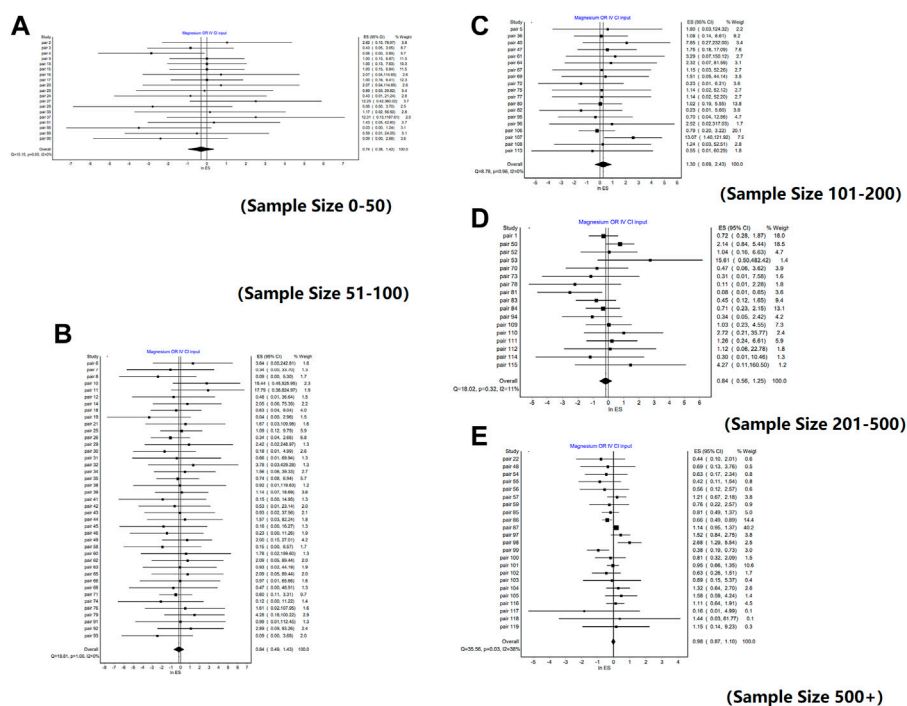
FIGURE 2
Forest plots of lnROR by sample size. [(A) Forest plot of lnROR for pairs with sample sizes between 0 and 50; (B) Forest plot of lnROR for pairs with sample sizes between 51 and 100; (C) Forest plot of lnROR for pairs with sample sizes between 101 and 200; (D) Forest plot of lnROR for pairs with sample sizes between 201 and 500; (E) Forest plot of lnROR for pairs with sample sizes, ore than 500].
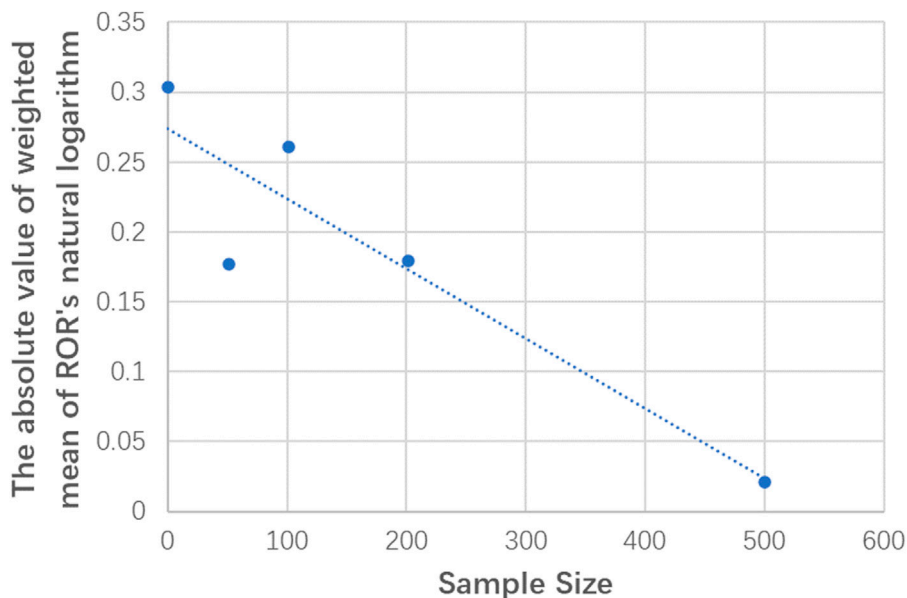

FIGURE 3
Scatter plot between the sample size and the absolute value of the weighted mean lnRORs.

Several previous studies have systematically evaluated the differences in the effects of adverse events between RCTs and NRSIs. One study included 19 systematic reviews, and the pooled ROR of RCTs compared to observational studies was estimated to be 1.03 (95% confidence interval 0.93–1.15) (Golder et al., 2011). The other two studies showed similar results (Grodstein et al., 2003; Edwards et al., 2012). These results are similar to our results and further confirm that there is no difference in the average risk

estimates of intervention adverse events between RCTs and NRSIs. One possible explanation for the findings is that for safety outcomes, the events are rare, and the sample sizes are also limited, which makes the random error the predominant error impact the effect over the systematic errors (e.g., error from confounding), and therefore under the same sample size with almost the same amount of random error, the effects are similar for RCTs and NRSIs.

However, some minor differences in the effects were observed. A study of postmenopausal hormone therapy on breast cancer survivors found that the results of observational studies were inconsistent with those of randomized trials (Col et al., 2005). This may be due to inconsistencies among the study population that they excluded people with a high incidence of adverse events. In Papanikolaou et al. (2006) study, the authors compared risks of 13 major harms of medical interventions using data from both RCTs and observational studies, and the non-randomized studies were often more conservative in their estimates of risk than the randomized trials. The study attributed these differences to the higher rate of adverse reactions reported by the RCTs because adverse events are recorded more thoroughly in RCTs, owing to regulatory requirements. It may also be caused by the different study populations. Further research on measuring the amount of random error and systematic error on NRSIs for rare events could be useful for the community to better understand the mechanism and deserves more attention.

4.1 Strengths and limitations

To the best of our knowledge, our study is currently the largest empirical study that compared the difference of the effects between RCTs and NRSIs for safety outcomes. The sample is representative, and the findings could provide indications for further evidence-based practice for assessing adverse events. In addition, we attempted to source the primary studies contained in each meta-analysis. This can avoid the errors that may exist in the extraction of data by the authors of meta-analyses. Moreover, we matched RCTs or NRSIs with the same outcome in the same systematic review according to their sample sizes, which can avoid the influence of different sample sizes on the results.

The current study has several limitations. First, we did not analyze and evaluate the bias of the included systematic review and possible confounding factors in the original study, such as drug dose, treatment duration, or study population. These confounding factors may affect the outcome of adverse events. In addition, even for the same adverse event, there are differences in how these events were defined or recorded, especially in composite outcomes. The absence of such methodological information increases the potential heterogeneity of the results and even biases the conclusion. Therefore, in the original study, detailed information on outcome collection should be sufficiently provided. Second, selection bias may occur in the current study. It has been estimated that only about 43% of the published studies reported adverse events, while the proportion is 88% in unpublished studies (Golder et al., 2016). This means in the current study, the studies included were those with better reporting on safety outcomes; thus, our results may not be representative of those with poor reporting. Third, we used the matching method for comparison; during the matching process, only 17% were matched among 1,405 studies from the 178 meta-analyses. This means the majority of RCTs and NRSIs have different sample sizes, and therefore whether the effects of them were similar or not is unclear. This is hard to be estimated as the sample size itself is a source of bias. In addition, systematic reviews of adverse events potentially have serious issues in data extraction, and these errors can mislead the conclusions (Xu et al., 2022). Even if data extraction is checked and corrected in this study, there may still be some errors. Further studies are warranted to address these issues.

5 Conclusion

In conclusion, the current study identified that there was no significant difference between RCTs and NRSIs in the evaluation of the effect of adverse events for the same topic when they have similar sample sizes. It is of great significance to the systematic reviews of adverse events that well-conducted NRSIs may provide valid results, which is similar to RCTs. Evidence from NRSIs might be considered a supplement to RCTs to improve the generalizability and comprehensiveness of the review.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

MD conceived and designed the study; MD analyzed the data and drafted the manuscript; AS collected the data, assessed the methodological quality, and edited the manuscript; LF-K, QW, and LL screened the literature; LL and LF-K provided methodological comments and revised the manuscript. All authors approved the final version for publication.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2023.1064567/full#supplementary-material

References

Abraham, N., Byrne, C., Young, J., and Solomon, M. (2010). Meta-analysis of well-designed nonrandomized comparative studies of surgical procedures is as good as randomized controlled trials. *J. Clin. Epidemiol.* 63 (3), 238–245. doi:10.1016/j.jclinepi.2009.04.005

Aires, F. T., Dedivitis, R. A., Petrarolha, S. M., Bernardo, W. M., Cernea, C. R., and Brandão, L. G. (2015). Early oral feeding after total laryngectomy: A systematic review. *Head neck* 37 (10), 1532–1535. doi:10.1002/hed.23755

Austin, P. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* 46 (3), 399–424. doi:10.1080/00273171.2011.568786

Balasubramanian, I., Fleming, C., Mohan, H. M., Schmidt, K., Haglind, E., and Winter, D. C. (2017). Out-patient management of mild or uncomplicated diverticulitis: A systematic review. *Dig. Surg.* 34 (2), 151–160. doi:10.1159/000450865

Benson, K., and Hartz, A. (2000). A comparison of observational studies and randomized, controlled trials. *N. Engl. J. Med.* 342 (25), 1878–1886. doi:10.1056/NEJM200006223422506

Bhaumik, D., Amatya, A., Normand, S., Greenhouse, J., Kaizar, E., Neelon, B., et al. (2012). Meta-analysis of rare binary adverse event data. *J. Am. Stat. Assoc.* 107 (498), 555–567. doi:10.1080/01621459.2012.664484

Ceresoli, M., Tamini, N., Gianotti, L., Braga, M., and Nespoli, L. (2019). Are endoscopic loop ties safe even in complicated acute appendicitis? A systematic review and meta-analysis. *Int. J. Surg. Lond. Engl.* 68, 40–47. doi:10.1016/j.ijsu.2019.06.011

Cheng, D., Gao, H., and Li, W. (2018). Long-term risk of rosiglitazone on cardiovascular events - a systematic review and meta-analysis. *Endokrynol. Pol.* 69 (4), 381–394. doi:10.5603/EP.a2018.0036

Chou, R., and Helfand, M. (2005). Challenges in systematic reviews that assess treatment harms. *Ann. Intern. Med.* 142, 1090–1099. doi:10.7326/0003-4819-142-12_part_2-200506211-00009

Col, N., Kim, J., and Chlebowski, R. (2005). Menopausal hormone therapy after breast cancer: A meta-analysis and critical appraisal of the evidence. *Breast cancer Res. BCR* 7 (4), R535–R540. doi:10.1186/bcr1035

Concato, J., Shah, N., and Horwitz, R. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* 342 (25), 1887–1892. doi:10.1056/NEJM200006223422507

Craveiro, N. S., Silva Lopes, B., Tomás, L., Fraga Almeida, S., Palma, H., Afreixo, V., et al. (2019). L-TRUST: Long-term risk of cancer in patients under statins therapy. A systematic review and meta-analysis. *Pharmacoepidemiol. drug Saf.* 28 (11), 1431–1439. doi:10.1002/pds.4895

Dechartres, A., Atal, I., Riveros, C., Meerpohl, J., and Ravaud, P. (2018). Association between publication characteristics and treatment effect estimates: A meta-epidemiologic study. *Ann. Intern. Med.* 169 (6), 385–393. doi:10.7326/M18-1517

Doi, S., Barendregt, J., Khan, S., Thalib, L., and Williams, G. (2015). Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemp. Clin. trials* 45, 130–138. doi:10.1016/j.cct.2015.05.009

Doi, S., and Furuya-Kanamori, L. (2020). Selecting the best meta-analytic estimator for evidence-based practice: A simulation study. *Int. J. evidence-based Healthc.* 18 (1), 86–94. doi:10.1097/XEB.0000000000000207

Doi, S., Furuya-Kanamori, L., Xu, C., Chivese, T., Lin, L., Musa, O., et al. (2021). The OR is "portable" but not the RR: Time to do away with the log link in binomial regression. *J. Clin. Epidemiol.* 142, 288–293. doi:10.1016/j.jclinepi.2021.08.003

Doi, S., Furuya-Kanamori, L., Xu, C., Lin, L., Chivese, T., and Thalib, L. (2020). Questionable utility of the relative risk in clinical research: A call for change to practice. *J. Clin. Epidemiol.* 142, 271–279. doi:10.1016/j.jclinepi.2020.08.019

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials--a review. *Statistics Med.* 3 (3), 199–214. doi:10.1002/sim.4780030302

Edwards, J., Kelly, E., Lin, Y., Lenders, T., Ghali, W., and Graham, A. (2012). Meta-analytic comparison of randomized and nonrandomized studies of breast cancer surgery. *Can. J. Surg. J. Can. de Chir.* 55 (3), 155–162. doi:10.1503/cjs.023410

Efthimiou, O. (2018). Practical guide to the meta-analysis of rare events. *Evidence-based Ment. health* 21 (2), 72–76. doi:10.1136/eb-2018-102911

Geminiani, A., Tsigarida, A., Chochlidakis, K., Papaspyridakos, P. V., Feng, C., and Ercoli, C. A meta-analysis of complications during sinus augmentation procedure. *Quintessence Int.* 2017;48(3):231–2. doi:10.3290/j.qi.a37644

Geng, H. Z., Nasier, D., Liu, B., Gao, H., and Xu, Y. K. (2015). Meta-analysis of elective surgical complications related to defunctioning loop ileostomy compared with loop colostomy after low anterior resection for rectal carcinoma. *Ann. R. Coll. Surg. Engl.* 97 (7), 494–501. doi:10.1308/003588415X14181254789240

Ghayoumi, P., Kandemir, U., and Morshed, S. (2015). Evidence based update: Open versus closed reduction. *Injury* 46 (3), 467–473. doi:10.1016/j.injury.2014.10.011

Golder, S., Loke, Y., and Bland, M. (2011). Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: Methodological overview. *PLoS Med.* 8 (5), e1001026. doi:10.1371/journal.pmed.1001026

Golder, S., Loke, Y. K., Wright, K., and Norman, G. (2016). Reporting of adverse events in published and unpublished studies of Health care interventions: A systematic review. *PLoS Med.* 13 (9), e1002127. doi:10.1371/journal.pmed.1002127

Grodstein, F., Clarkson, T., and Manson, J. (2003). Understanding the divergent data on postmenopausal hormone therapy. *N. Engl. J. Med.* 348 (7), 645–650. doi:10.1056/NEJMsb022365

Grootscholten, K., Kok, M., Oei, S. G., Mol, B. W., and van der Post, J. A. (2008). External cephalic version-related risks: A meta-analysis. *Obstetrics Gynecol.* 112 (5), 1143–1151. doi:10.1097/AOG.0b013e31818b4ade

Guyatt, G., Oxman, A., Vist, G., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., et al. (2008). Grade: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ Clin. Res. ed)* 336 (7650), 924–926. doi:10.1136/bmj.39489.470347.AD

Hemkens, L., Contopoulos-Ioannidis, D., and Ioannidis, J. (2016). Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ Clin. Res. ed)* 352, i493. doi:10.1136/bmj.i493

Higgins, J., Altman, D., Gøtzsche, P., Jüni, P., Moher, D., Oxman, A., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ Clin. Res. ed)* 343, d5928. doi:10.1136/bmj.d5928

Inokuchi, M., Sugita, H., Otsuki, S., Sato, Y., Nakagawa, M., and Kojima, K. (2015). Laparoscopic distal gastrectomy reduced surgical site infection as compared with open distal gastrectomy for gastric cancer in a meta-analysis of both randomized controlled and case-controlled studies. *Int. J. Surg. Lond. Engl.* 15, 61–67. doi:10.1016/j.ijsu.2015.01.030

Ioannidis, J., Haidich, A., Pappa, M., Pantazis, N., Kokori, S., Tektonidou, M., et al. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 286 (7), 821–830. doi:10.1001/jama.286.7.821

Jia, P., Lin, L., Kwong, J., and Xu, C. (2021). Many meta-analyses of rare events in the Cochrane Database of Systematic Reviews were underpowered. *J. Clin. Epidemiol.* 131, 113–122. doi:10.1016/j.jclinepi.2020.11.017

Jiang, Z., Xiao, H., Zhang, H., Liu, S., and Meng, J. (2019). Comparison of adverse events between cluster and conventional immunotherapy for allergic rhinitis patients with or without asthma: A systematic review and meta-analysis. *Am. J. otolaryngology* 40 (6), 102269. doi:10.1016/j.amjoto.2019.07.013

Jreich, R., and Sebastien, B. (2021). Comparison of statistical methodologies used to estimate the treatment effect on time-to-event outcomes in observational studies. *J. Biopharm. statistics* 31, 469–489. doi:10.1080/10543406.2021.1918140

Ju, K., Lin, L., Chu, H., Cheng, L., and Xu, C. (2020). Laplace approximation, penalized quasi-likelihood, and adaptive gauss-hermite quadrature for generalized linear mixed models: Towards meta-analysis of binary outcome with sparse data. *BMC Med. Res. Methodol.* 20 (1), 152. doi:10.1186/s12874-020-01035-6

Keir, A., Pal, S., Trivella, M., Lieberman, L., Callum, J., Shehata, N., et al. (2016). Adverse effects of red blood cell transfusions in neonates: A systematic review and meta-analysis. *Transfusion* 56 (11), 2773–2780. doi:10.1111/trf.13785

Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S., and Johnston, J. (2015). A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 16, 495. doi:10.1186/s13063-015-1023-4

Mathes, T., Rombey, T., Kuss, O., and Pieper, D. (2021). No inexplicable disagreements between real-world data-based nonrandomized controlled studies and randomized controlled trials were found. *J. Clin. Epidemiol.* 133, 1–13. doi:10.1016/j.jclinepi.2020.12.019

McNamee, R. Regression modelling and other methods to control confounding. *Occup. Environ. Med.* 2005;62(7):500–50. doi:10.1136/oem.2002.001115

Menne, J., Dumann, E., Haller, H., and Schmidt, B. M. W. (2019). Acute kidney injury and adverse renal events in patients receiving SGLT2-inhibitors: A systematic review and meta-analysis. *PLoS Med.* 16 (12), e1002983. doi:10.1371/journal.pmed.1002983

Moher, D., Dulberg, C., and Wells, G. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 272 (2), 122–124. doi:10.1001/jama.1994.03520020048013

Nagy, A., Mátrai, P., Hegyi, P., Alizadeh, H., Bajor, J., Czopf, L., et al. (2019). The effects of TNF-alpha inhibitor therapy on the incidence of infection in JIA children: A meta-analysis. *Pediatr. rheumatology online J.* 17 (1), 4. doi:10.1186/s12969-019-0305-x

Papanikolaou, P., Christidi, G., and Ioannidis, J. (2006). Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ Can. Med. Assoc. J. = J. de l'Association medicale Can.* 174 (5), 635–641. doi:10.1503/cmaj.050873

Pateras, K., Nikolakopoulos, S., and Roes, K. (2018). Data-generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis. *Statistics Med.* 37 (7), 1115–1124. doi:10.1002/sim.7569

Pecorelli, N., Greco, M., Amodeo, S., and Braga, M. (2017). Small bowel obstruction and incisional hernia after laparoscopic and open colorectal surgery: A meta-analysis of comparative trials. *Surg. Endosc.* 31 (1), 85–99. doi:10.1007/s00464-016-4995-6

Peng, C., Ling, Y., Ma, C., Ma, X., Fan, W., Niu, W., et al. (2016). Safety outcomes of notes cholecystectomy versus laparoscopic cholecystectomy: A systematic review and meta-analysis. *Surg. Laparosc. Endosc. percutaneous Tech.* 26 (5), 347–353. doi:10.1097/SLE.0000000000000284

Reeves, B., Higgins, J., Ramsay, C., Shea, B., Tugwell, P., and Wells, G. (2013). An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Res. synthesis methods* 4 (1), 1–11. doi:10.1002/jrsm.1068

Shah, K., Chaker, Z., Busu, T., Badhwar, V., Alqahtani, F., Alvi, M., et al. (2018). Meta-analysis comparing the frequency of stroke after transcatheter versus surgical aortic valve replacement. *Am. J. Cardiol.* 122 (7), 1215–1221. doi:10.1016/j.amjcard.2018.06.032

Shah, K., Chaker, Z., Busu, T., Shah, R., Osman, M., Alqahtani, F., et al. (2019). Meta-analysis comparing renal outcomes after transcatheter versus surgical aortic valve replacement. *J. interventional Cardiol.* 2019, 3537256. doi:10.1155/2019/3537256

Shrier, I., Boivin, J. F., Steele, R. J., Platt, R. W., Furlan, A., Kakuma, R., et al. (2007). Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *Am. J. Epidemiol.* 166 (10), 1203–1209. doi:10.1093/aje/kwm189

Slobogean, B. L., Jackman, H., Tennant, S., Slobogean, G. P., and Mulpuri, K. (2010). Iatrogenic ulnar nerve injury after the surgical treatment of displaced supracondylar fractures of the humerus: Number needed to harm, a systematic review. *J. Pediatr. Orthop.* 30 (5), 430–436. doi:10.1097/BPO.0b013e3181e00c0d

Soni, P., Hartman, H., Dess, R., Abugharib, A., Allen, S., Feng, F., et al. (2019). Comparison of population-based observational studies with randomized trials in oncology. *J. Clin. Oncol. official J. Am. Soc. Clin. Oncol.* 37 (14), 1209–1216. doi:10.1200/JCO.18.01074

Sun, J. C., Whitlock, R., Cheng, J., Eikelboom, J. W., Thabane, L., Crowther, M. A., et al. (2008). The effect of pre-operative aspirin on bleeding, transfusion, myocardial infarction, and mortality in coronary artery bypass surgery: A systematic review of randomized and observational studies. *Eur. heart J.* 29 (8), 1057–1071. doi:10.1093/eurheartj/ehn104

Torloni, M. R., Vedmedovska, N., Merialdi, M., Betrán, A. P., Allen, T., González, R., et al. (2009). Safety of ultrasonography in pregnancy: WHO systematic review of the literature and meta-analysis. *Ultrasound obstetrics Gynecol. official J. Int. Soc. Ultrasound Obstetrics Gynecol.* 33 (5), 599–608. doi:10.1002/uog.6328

Touzé, E., Trinquart, L., Chatellier, G., and Mas, J. L. (2009). Systematic review of the perioperative risks of stroke or death after carotid angioplasty and stenting. *Stroke* 40 (12), e683–e693. doi:10.1161/STROKEAHA.109.562041

Valentine, J., and Thompson, S. (2013). Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. *Res. synthesis methods* 4 (1), 26–35. doi:10.1002/jrsm.1064

Van Spall, H., Toren, A., Kiss, A., and Fowler, R. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *JAMA* 297 (11), 1233–1240. doi:10.1001/jama.297.11.1233

Vandenbroucke, J., and Pearce, N. (2012). Case-control studies: Basic concepts. *Int. J. Epidemiol.* 41 (5), 1480–1489. doi:10.1093/ije/dys147

Vaos, G., Dimopoulou, A., Gkioka, E., and Zavras, N. (2019). Immediate surgery or conservative treatment for complicated acute appendicitis in children? A meta-analysis. *J. Pediatr. Surg.* 54 (7), 1365–1371. doi:10.1016/j.jpedsurg.2018.07.017

Vavken, J., Mameghani, A., Vavken, P., and Schaeren, S. (2016). Complications and cancer rates in spine fusion with recombinant human bone morphogenetic protein-2 (rhBMP-2). *Eur. spine J.* 25 (12), 3979–3989. doi:10.1007/s00586-015-3870-9

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., et al. (2008). The strengthening the reporting of observational studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *J. Clin. Epidemiol.* 61 (4), 344–349. doi:10.1016/j.jclinepi.2007.11.008

Wang, F. B., Pu, Y. W., Zhong, F. Y., Lv, X. D., Yang, Z. X., and Xing, C. G. (2015). Laparoscopic permanent sigmoid stoma creation through the extraperitoneal route versus transperitoneal route. A meta-analysis of stoma-related complications. *Saudi Med. J.* 36 (2), 159–163. doi:10.15537/smj.2015.2.10203

Wang, X., and Ji, X. (2020). Sample size estimation in clinical research: From randomized controlled trials to observational studies. *Chest* 158, S12–S20. doi:10.1016/j.chest.2020.03.010

Winberg, H., Arnbjörnsson, E., Anderberg, M., and Stenström, P. (2019). Postoperative outcomes in distal hypospadias: A meta-analysis of the mathieu and tubularized incised plate repair methods for development of urethrocutaneous fistula and urethral stricture. *Pediatr. Surg. Int.* 35 (11), 1301–1308. doi:10.1007/s00383-019-04523-z

Xu, C., and Doi, S. (2018). The robust error meta-regression method for dose-response meta-analysis. *Int. J. evidence-based Healthc.* 16 (3), 138–144. doi:10.1097/XEB.0000000000000132

Xu, C., Furuya-Kanamori, L., and Lin, L. (2021). Synthesis of evidence from zero-events studies: A comparison of one-stage framework methods. *Res. synthesis methods* 13, 176–189. doi:10.1002/jrsm.1521

Xu, C., Furuya-Kanamori, L., Zorzela, L., Lin, L., and Vohra, S. (2021). A proposed framework to guide evidence synthesis practice for meta-analysis with zero-events studies. *J. Clin. Epidemiol.* 135, 70–78. doi:10.1016/j.jclinepi.2021.02.012

Xu, C., Ju, K., Lin, L., Jia, P., Kwong, J., Syed, A., et al. (2021). Rapid evidence synthesis approach for limits on the search date: How rapid could it be? *Res. synthesis methods* 13, 68–76. doi:10.1002/jrsm.1525

Xu, C., Yu, T., Furuya-Kanamori, L., Lin, L., Zorzela, L., Zhou, X., et al. (2022). Validity of data extraction in evidence synthesis practice of adverse events: Reproducibility study. *Bmj* 377, e069155. doi:10.1136/bmj-2021-069155

Xu, C., Zhou, X., Zorzela, L., Ju, K., Furuya-Kanamori, L., Lin, L., et al. (2021). Utilization of the evidence from studies with no events in meta-analyses of adverse events: An empirical investigation. *BMC Med.* 19 (1), 141. doi:10.1186/s12916-021-02008-2

Yaghoobi, M., Farrokhyar, F., Yuan, Y., and Hunt, R. H. (2010). Is there an increased risk of GERD after *Helicobacter pylori* eradication?: A meta-analysis. *Am. J. gastroenterology* 105 (5), 1007–1013. quiz 6, 14. doi:10.1038/ajg.2009.734

Yang, C., Yi, Q., Zhang, L., Cui, H., and Mao, J. (2019). Safety of aripiprazole for tics in children and adolescents: A systematic review and meta-analysis. *Medicine* 98 (22), e15816. doi:10.1097/MD.0000000000015816

Yoon, B. H., Ha, Y. C., Lee, Y. K., and Koo, K. H. (2015). Postoperative deep infection after cemented versus cementless total hip arthroplasty: A meta-analysis. *J. arthroplasty* 30 (10), 1823–1827. doi:10.1016/j.arth.2015.04.041

Zhang, Y., and Ma, L. (2015). Effect of preoperative angiotensin-converting enzyme inhibitor on the outcome of coronary artery bypass graft surgery. *Eur. J. cardio-thoracic Surg. official J. Eur. Assoc. Cardio-thoracic Surg.* 47 (5), 788–795. doi:10.1093/ejcts/ezu298

Zhao, Y., Peng, H., Li, X., Qin, Y., Cao, F., Peng, D., et al. (2018). Dual antiplatelet therapy after coronary artery bypass surgery: Is there an increase in bleeding risk? A meta-analysis. *Interact. Cardiovasc. Thorac. Surg.* 26 (4), 573–582. doi:10.1093/icvts/ivx374

Zorzela, L., Golder, S., Liu, Y., Pilkington, K., Hartling, L., Joffe, A., et al. (2014). Quality of reporting in systematic reviews of adverse events: Systematic review. *Bmj* 348, f7668. doi:10.1136/bmj.f7668