Journal of Clinical and Translational Research

Journal homepage: http://www.jctres.com/en/home

ORIGINAL ARTICLE

Meta-analysis of clinical trials in the 2020s and beyond: a paradigm shift needed

Jonathan J. Shuster*

Department of Health Outcomes and Bioinformatics, College of Medicine, University of Florida, Gainesville, Florida 32605, United States of America

ARTICLE INFO

Article history:
Received: February 15, 2022
Revised: April 11, 2023
Accepted: June 10, 2023
Published online: July 12, 2023

Keywords:
clinical trial
meta-analysis
random effects

**Corresponding authors:*
Jonathan J. Shuster
Department of Health Outcomes and
Bioinformatics, College of Medicine,
University of Florida, 2026 NW 34 Ter,
Gainesville, FL 32605, United States of
America.
Tel: +1(352)682-0893
Email: shusterj@ufl.edu

ABSTRACT

Background: A peer-reviewed meta-analysis methods article mathematically proved that mainstream random-effects methods, "weights inversely proportional to the estimated variance," are flawed and can lead to faulty public health recommendations. Because the arguments causing this off-label (unproven) use of mainstream practices were subtle, changing these practices will require much clearer explanations that can be grasped by clinical and translational scientists. There are five assumptions underlying the mainstream's derivation of its statistical properties. This paper will demonstrate that if the first is true, it follows that the last two are false. Ratio estimation, borrowed from classical survey sampling, provides a rigorous alternative. Papers reporting results rarely fully disclose these assumptions. This is analogous to watching TV ads with the sound muted. You see high quality of life and do not hear about the complications. This article is a poster child for translational science, as it takes a theoretical discovery from the biostatistical world, translates it into language clinical scientists can understand, and thereby can change their research practice.

Aim: This article is aimed at future applications of meta-analysis of complete collections of randomized clinical trials. It leaves it to past authors as to whether to reanalyze their data. No blame for past use is assessed.

Methods: By treating the individual completed studies in the meta-analysis as a random sample from a conceptual universe of completed studies, we use ratio estimation to obtain estimates of relative risk (ratio of failure rates treatment: control) and mean differences, projecting our sample value to estimate the universe's value.

Results: Two examples demonstrate that the mainstream methods likely adversely impacted major treatment options. A third example shows that the key mainstream presumption of independence between the study weights and study estimates cannot be supported.

Conclusion: There is no rationale for ever using the mainstream for meta-analysis of randomized clinical trials.

Relevance for Patients: Future meta-analysis of clinical trials should never employ mainstream methods. Doing so could lead to potentially harmful public health policy recommendations. Clinical researchers need to play a primary role to assure good research practices in meta-analysis.

1. Introduction

As hard as this is to believe, the recent paper, Shuster [1], mathematically proved beyond any doubt that despite being in common use for over four decades, the mainstream methods of conducting random effects meta-analyses (true individual study-by-study effect sizes can differ) are unsound and are likely to produce misleading results that could be threats to public health. It is commendable if you, as a clinical investigator or reader, would be skeptical of this statement. Needless to say, because Shuster [1] was controversial, it was

DOI: http://dx.doi.org/10.18053/jctres.09.202304.22-00019

one of the most heavily peer-reviewed biostatistical papers ever, with review material from nine of nine sources (including three world-renowned meta-analysts) agreeing with his conclusions. However, in this article, in a non-technical way, you will be shown clearly that the mainstream methods rely on five incompatible assumptions that underlie their validity. This makes the evidence basis of the mainstream no different than claiming an evidence basis for off-label usage of a drug. You will be shown that if the first assumption is true, the last two cannot be true. In one highly cited example, we show that the mainstream-based claim of efficacy for an invasive treatment has no scientific basis. In another, the mainstream methods failed to detect a highly significant outcome. Had these methods been available and used, the use of a cardiotoxic type II diabetes drug could have been discontinued at least 3 years earlier than what actually occurred. We provide free access to well-documented user-friendly Excel templates to conduct rigorous analyses of the main research questions. This paper places no blame on the well-intentioned researchers who developed these mainstream methods. However, if meta-analysis is to remain at the apex of "Evidence Pyramids," it is imperative that statistical practice should be changed. This paper is needed for two reasons. First, with the availability of user-friendly software for the mainstream methods, a high proportion of these analyses is done without input from biostatisticians or epidemiologists. Second, changes to statistical practice will not happen overnight. Readers should be concerned when they read papers using meta-analysis in the biomedical literature. In short, this is about proven science, not opinions.

2. Assumptions Underlying the Validity of Current Mainstream Methods

When methodologists derive analytic procedures, they make working distributional assumptions that enable them to complete their work. Every time the procedure is used in practice, these should be fully disclosed. In a specific application, if any of the assumptions are wrong, the evidentiary basis of the results is in jeopardy. Unfortunately, few analytic procedures have adequate diagnostic tests for their assumptions. In meta-analysis, there has been little vetting of the robustness of procedures when their assumptions fail.

"Weighting inversely proportional to the estimated variance estimation" (aka the mainstream method) is by far the most common method used in combining data from a complete set of randomized clinical trials of a research question. These methods were derived under five assumptions (A1-A5) below, which must be true up to strong approximations. These are rarely disclosed in full, and current software does not provide adequate warnings. Assumptions A1 and A3 are reasonable in most applications. Assumption A2 is questionable (no adequately powered diagnostic test exists for it). Unfortunately, even if Assumption A1 is true, it follows that Assumptions A4 and A5 are false. This leaves open the strong likelihood that past meta-analyses may have reached unsupportable conclusions, possibly contributing to inappropriate public health recommendations.

A1: The true primary effect sizes for each study are drawn independently from a single large "urn" of primary effect sizes. This assumption tells us we are targeting the unweighted mean of all studies in the urn.

A2: The true primary effect sizes in the urn follow a normal (bell-shaped) distribution whose unweighted mean is the target parameter of interest.

A3: The individual study provides an unbiased estimate of its study-specific true primary effect size and has an approximate normal distribution about its true primary effect size.

A4: Up to a strong approximation, the weights are "constants" rather than seriously random variables. In other words, if you repeat the total experiment under the same Assumptions A1-A3 and the same urn, this assumption presumes you obtain identical weights up to a strong approximation. This assumption is mandatory to use the formulas for the mean and standard error in the mainstream methods but will be shown to be false under Assumptions A1-A3. More on this below.

A5: There is no association between-study weight and study true effect size. For example, if big studies tend to have higher (lower) effect sizes than smaller studies, the method will tend to overestimate (underestimate), respectively, the overall effect size. This could lead to unacceptable bias.

3. How Mainstream Weighted Random Effects Methods Work

The "variance" for the estimates of effect size for each study consists of two components, the reasons why its individual study estimate of effect size differs from the true mean of the effect sizes in the urn: (a) Within-study variance, which is estimated under Assumption A3 and (b) Between-study variance, which is the variance of the true effect sizes in the urn, per Assumptions A1 and A2. The first component (a) depends on the accuracy of the within-study estimate and varies from study to study. The second component (b) is the same for all studies. The overall estimate is the weighted average of the individual study estimates with weights inversely proportional to the study's estimated variance, which is the sum of the estimated within-study variance and the estimated between-study variance. If all five assumptions were true, these weights would minimize the standard error of the estimate of the overall effect size, over all choices of weights that sum to one (a requirement for unbiasedness). Note that all other things being equal, larger between-study variance pushes the weights closer to equal weights and smaller between-study variance pushes the weights closer to fixed effects.

4. Why Assumption A4 is False

What this assumption requires is that if we repeat the experiment under Assumptions A1-A3, the resulting weights will be the same up to a strong approximation.

Imagine a meta-analysis where we independently generate the data twice under Assumptions A1-A3. Clearly, the true study effect sizes for these two repetitions are sure to differ. It follows that the diversity (sample variance) of these true effect sizes will

DOI: http://dx.doi.org/10.18053/jctres.09.202304.22-00019

differ. All things being equal, the one with the greater sample variance in true effect sizes will have weights closer to equality than the other, thanks to a larger between-study variance. As a concrete example, when the number of studies combined is eight, there is a 61% probability that one sample variance for these true effect sizes will be at least 50% higher than the other. Assumption A4 requires these to be the same to a near certainty. The derivation of the 61% figure is in the Appendix for those with biostatistical expertise. The between-study variance is a major determinant of the weights and clearly differs between repetitions of obtaining the meta-analysis data under Assumptions A1-A3.

Support for the fact that weights are seriously random variables comes from an unlikely source, lead developer of perhaps the most popular software product for this subject, Comprehensive meta-analysis (CMA), Borenstein [2], who states this assumption in Section 7.4.3, "The studies that were performed are a random sample from the universe." This concedes the point that mainstream weights, which are functions of the studies, are seriously random variables, not constants. This potentially invalidates the claims of no bias in the overall effect size estimate and legitimacy of confidence intervals and P-values.

In short, the mainstream relies on theory that was never intended for this type of application and as such, the distribution theory is used off label.

5. Why Assumption A5 is False

This one should be clear from the fact that the weights are determined by the variances (diversity) of the effect sizes. The more diverse the true study-specific effect sizes are (Assumptions A1-A3), the closer the weights are to being equal. In short, the mainstream weights are in part determined by the effect size estimates rendering the claim of independence untrue.

6. Why Assumption A2 Should Not be Trusted

Assumption A2 presumes that the true effect size for each study is drawn from the same urn and has a normal distribution. This implies that on average, the true study-specific effect sizes are the same regardless of study design. There is no adequately powered diagnostic test that can prove with reasonable certainty that this is true. For example, as shown by Shuster [1], any non-zero correlation between weight and effect size will bias the overall estimate of effect size and invalidate its standard error formula. Further, there is no adequately powered diagnostic test that can prove with reasonable certainty that the individual true study-specific effect size follows a normal distribution.

7. How Ratio Estimation Works

Our inferential framework is identical to that of randomized clinical trials. The role of patient in the clinical trial is played by study in the meta-analysis. The following is a quotation from Shuster [1], "A meta-analysis (clinical trial) inference is based on the sample of studies (patients) in the meta-analysis (clinical trial) as a conceptual random sample of past, present, and future studies (patients), drawn from a large target population of studies (patients) with the same eligibility criteria. The inference is to this target population."

Our universe is a large conceptual population of completed studies and the actual studies are a conceptual random sample from this universe. Our inference is to the target parameter in the entire conceptual population. Our estimate is the corresponding value in the sample of studies in the analysis. The target metric simply projects what the relative risk (or difference in means or difference in proportions) would be if all patients received the experimental therapy versus that if all patients received the control therapy. This framework is different from the mainstream, and hence, it is important to note that the ratio method targets a different population parameter than the mainstream.

Note that this setup can accommodate any distribution of means or proportions for the two treatment arms, making it a model-free random effects framework for meta-analysis. The mainstream imposes severe restrictions through its five Assumptions A1-A5.

7.1. Illustration for relative risk (risk ratio)

For each study in the universe, if we had the number of failures on each treatment (experimental and control), we could project the number of "failures" that would occur if every subject was in the experimental group (control group), respectively. For each individual study, this would be the total sample size (treatment + control) for the study multiplied by the proportion failing in the experimental group (control group), respectively. For example, in the first study in Table 1, we see that the experimental group had two failures in 26 patients, while the control group had one failure in 26 patients. We project that if all 52 subjects had gotten the experimental treatment, we would project that we would have had 52(2/26) = 4 failures. Similarly, we would project that if all patients had received the control, we would project 52(1/26) = 2 failures. Note that projections need not be whole numbers. If, for each treatment, we added the projected number of failures for all studies in the universe and take the ratio that would yield the projected true relative risk: Projected # failing in the universe (experimental group) divided by Projected # failing in the universe (control group). The corresponding projected ratio in the actual conducted sample of completed studies gives us the estimate. Technical notes: The confidence intervals and P-values are derived using the natural logs of the ratio and back converting the confidence interval using natural antilogs. The Users' guide

Table 1. Neto *et al.*[4] example for relative risk

Study#	Deaths on RX	N (Rx)	Deaths on control	N (Control)
1	2	26	1	26
2	3	23	2	13
3	27	163	69	212
4	13	558	15	533
5	24	76	23	74
6	3	154	1	75
7	1	75	2	74
8	0	50	1	50
9	1	20	1	20

and Excel software do all calculations for you automatically if you enter the tabular data analogous to Table 1.

7.2. Illustration for a difference in means or proportions

For each study in the conceptual universe, we project the difference in its totals if all patients received treatment less than if all patients were controls as the difference in means (or proportions) multiplied by the total sample size (treatment + control). The target population projection adds these up for all studies in the universe and divides this total by the total number of patients in the universe of all conceptually completed studies. The estimate is simply the corresponding value in our sample. If you refer to the difference in means data from the second example in the users' guide, second study, you will note that the experimental group had a sample mean of −3.0 in 42 patients while the control group had a sample mean of −2.5 in 51 patients. This makes the projected mean difference of −0.5 (experimental minus control) in 93 patients for a projected total of −0.5*93 = −46.5. The Users' guide and Excel software do all calculations for you automatically if you enter the tabular data analogous to this example in the User's Guide.

8. How Equal Weighting Works

We do not advocate equal weighting, but it can give us important insight into the credibility of analyses that use mainstream weights. We use the same methods as the mainstream to calculate the estimate and standard error but use equal weights instead of mainstream weights.

9. How Statistical Inference is Done

For any form of meta-analysis, including the mainstream, to obtain point estimates, confidence intervals, and P-values, the following approximations are used: The standardized difference, the difference between the overall estimate of effect size and the true global mean effect size, divided by its standard error of the estimate is obtained.

a. The mainstream uses a standard normal approximation, although the package CMA now has an option to use a T-approximation with degrees of freedom equal to the number of studies being combined less one
b. The ratio estimation method uses a T-approximation with degrees of freedom equal to the number of studies being combined less two
c. The equally weighted method uses a T-approximation with degrees of freedom equal to the number of studies being combined less one.

More on these approximations will appear in the discussion.

10. Numerical Examples

We shall provide three illustrations, one for the primary published relative risk of an invasive intervention, one for the myocardial infarction data of Nissen and Wolski [3], and one from a submitted article that incorrectly reported one odds ratio. The correction did not affect within-study variance estimators, but dramatically impacted the weights, demonstrating that weights and effect size estimates, contrary to Assumption A5, cannot be presumed to be independent.

10.1. Example 1: Relative risk

Neto *et al*. [4] in a highly cited meta-analysis of randomized trials found a benefit in their invasive intervention over the control for their primary outcome, total mortality. Table 1 provides the published numerators and denominators for each of the contributing studies, while Table 2 provides the results (i) as published, (ii) doubling all numerators and denominators, (iii) equally weighted, and (iv) by the method of Shuster [1]. As of 11/2022, this Journal of the American Medical Association paper has been cited 877 times.

Table 2 yields surprising results. Intuitively, doubling all numerators and denominators which keep the study-by-study estimates (signals) the same, but would diminish the noise (standard errors) within each study by a factor of about 30%, should yield a more significant result. Why would the confidence interval for the overall estimate of effect size grow by 15% while losing the significant finding, with the *P*-value becoming 0.15 instead of 0.013? This is indeed a red flag that will be clarified in the discussion. Neither the equally weighted nor the Ratio estimate produces definitive results on efficacy. In this case, this published result affected public health policy based on an off-label use of statistical methodology.

10.2. Example 2: Rosiglitazone and increased myocardial infarction risk

In their publication, Nissen and Wolski [3] used a fixed-effects method, even though the combined trials were highly diverse in terms of control groups, eligibility, duration and dose of treatment, and duration of follow-up. They used odds ratios instead of relative risk, the preferred metric. When event rates are low, the distinction is minor. Table 3 contrasts the results of mainstream methods, the published result of Nissen and Wolski [3], with those of Shuster [1], for relative risk. The Nissen and Wolski published that confidence interval excludes the neutral value of 1.00 but includes clinically insignificant values close to 1.00. Had ratio methods been available, a full ban on rosiglitazone might have occurred in 2007, thanks to the fact that the confidence interval includes only clinically significant increased risk for rosiglitazone. Although sales dropped from over $2 billion in 2007 and beyond, a large volume of sales continued for years afterward. As late as 2010, annual sales totaled almost $700 million. Several other nations did not ban the drug until 2010 or 2011. To further confuse the situation in 2007, Diamond and Kaul [5] published a non-significant mainstream analysis which may have slowed the decline at the additional human cost of cardiac events. The Nissen and Wolski [3] New England Journal of Medicine publication is one of the most cited meta-analysis reports, with 5908 citations as of 11/2022.

10.3. Example 3: From a peer-review of a submission to a major medical journal

The crux of this six-study observational example is that a peer-reviewer discovered that the odds ratio estimate in one of the

Table 2. Results for data in Table 1

Method	Estimated relative risk RX: control (95% CI)	P-value: two-sided	Ratio of 95% confidence lengths method: Inv Var
Mainstream weights (Published)	0.71 (0.55, 0.93)	0.013	1.00
Double numerators and denominators	0.78 (0.56, 1.09)	0.15	1.15
Equally weighted	0.82 (0.54–1.26)	0.33	1.38
Ratio (Survey sampling)	0.70 (0.44, 1.11)	0.11	1.49

Table 3. Nissen and Wolski re-analysis for myocardial infarction relative risk for rosiglitazone

Method	Estimated relative risk RX: control (95% CI)	*P*-value: two-sided	Ratio of 95% confidence lengths method: Inv Var
Mainstream Weights (RR)	1.28 (0.94, 1.75)	0.12	1.00
Ratio (Survey Sampling) (RR)	1.41 (1.14, 1.75)	0.0026	0.82
Nissen and Wolski (OR)	1.43 (1.03–1.98)	0.032	1.03

studies was wrong, and the actual odds ratio was 1/reported odds ratio. The generic data are given in Table 4 below. The reported estimated odds ratio of Study 4 was 0.78 when in fact it was 1.28. This occurred in the largest study in the meta-analysis (62% of the subjects) and pushed its estimated odds ratio from near the center of the original meta-analysis to close to being the largest estimated odds ratio. This resulted in a substantial increase in the between-study variance estimate. According to Assumption A1, this came from a single "draw" from the urn that affected the between-study variance estimation. Contrary to Assumption A5, the impact of the effect size change upon the weights was dramatic: Under the original scenario, the weight for this study was 23.3%. Under the corrected data, it dropped to 19.7%, and weights for the other five studies also changed. Note that equal weighting would assign 16.7% weight to each of the six studies. The change of one effect size estimate altered its weight by 3.6% or about half of the way from its original weight to equal weights. Therefore, the value of the study mean effect sizes drawn from the urn (A1) impacts the between-study variance estimate, and hence, Assumption A5 cannot be trusted. Note also that sample size weights can be vastly different from mainstream weights (study 4 had 62% of patients, but 19.7% weight for the mainstream).

Table 4. Generic data from submitted article to a major journal

Study	Group A events #Yes/#No (Odds)	Group B events #Yes/#No (Odds)	Estimated odds ratio (calculated)
1	14/225 (0.062)	245/1599 (0.153)	0.41
2	46/489 (0.094)	453/2570 (0.176)	0.53
3	90/551 (0.163)	625/2355 (0.265)	0.62
4	594/2204 (0.270)	3198/15218 (0.210)	1.28
5	42/342 (0.123)	97/806 (0.120)	1.02
6	22/277 (0.079)	107/1872 (0.057)	1.38

11. Discussion

Despite 48 years of practice, the mainstream method for weighted random effects meta-analysis should not be used in the future. "Bayes" methods also have some of the same issues (sample sizes are random variables not constants, and associations between sample size and effect size will produce bias).

11.1. Assumptions underlying inferences for the three methods

(a) For the standardized difference, mainstream methods rely on a "normal distribution" that in addition to Assumptions A1-A5, presumes that the number of studies is large enough to utilize the standard normal distribution. (b) The ratio method relies on the single assumption that the number of studies is large enough to apply its large sample T-distribution, with degrees of freedom equal to the number of studies less two, to its standardized difference. Within-study approximations are not relevant. Studies with zero events on one or both arms are included. Continuity corrections are unnecessary and never made. (c) The equally weighted method relies on the single assumption that the number of studies is large enough to apply its T-approximation, with degrees of freedom equal to the number of studies less one, to its standardized difference. Within-study approximations are not relevant.

11.2. Assumptions behind the ratio method

There are no assumptions except for (b) above. Shuster *et al*. [6] vetted the approximation for relative risks, when the number of studies ranged from 5 to 20, with nearly 40,000 diverse scenarios, each replicated 100,000 times. The coverage of the 95% confidence intervals was consistently close to 95%. However, the corresponding coverage using the less conservative normal approximation was generally well below 95%. This should be a warning that the mainstream coverage of their purported 95% confidence intervals is suspect when the number of studies being combined is in the 5–20 range. The vetting of differences in means and proportions is more difficult and needs independent funding with supercomputers to properly vet. For these studies, a limitation is needed in any paper with fewer than 20 studies.

The first two numerical examples demonstrate the dangers of relying upon the mainstream methods. The first is counterintuitive while the second illustrates that estimation bias in the mainstream is a real threat to getting a conclusive result.

Shuster [1] reported on a small sample of 32 highly cited past meta-analyses that used mainstream methods for relative risk and found major disparities in eight (25%). It is fortunate that this is not higher, but this is not good enough for trust in mainstream methods.

An analysis of the 31 of these studies reported in Shuster *et al*. [6] and Shuster and Walker [7] (the 32nd study's reanalysis had a few studies added but trended as the 31 we analyzed) also dispel the one remaining scientific as opposed to traditional reasons for using the mainstream: The mainstream might produce on average narrower confidence intervals. If you analyze the natural logs of the ratios of the lengths of the confidence intervals (the traditional way to analyze non-negative ratios) and treat the studies as a random sample of highly cited meta-analyses, we obtain an estimate of the population ratio of widths (Mainstream: Shuster) of 1.10 (Mainstream wider) with 95% confidence interval from 0.93 to 1.29. The mainstream plausible mean in the total population of such potential reanalysis ranges from slightly shorter to substantially longer.

Further, due to mainstream proponents' concerns about its normal approximation, newer versions of CMA have added a t-option (degrees of freedom number of studies less one) that can be used instead of the normal approximation. When we replaced the normal with the t, the new mainstream methods were significantly less accurate than the survey-based method. For our sample of 31 meta-analyses, in the log scale, the mainstream averaged 30% wider than the survey-based methods, with 95% confidence interval from 9% wider to 54% wider. This is yet another strong reason not to use the mainstream.

Note that diagnostic test information, such as Cochran's Q, I^2, and Egger's test for selection bias, as described in Borenstein *et al*. [8] is not relevant to the validity of Shuster [1].

Shuster *et al*. [6], with a substantial number of simulations, found that when the target population relative risks in the two universes were the same (ratio and equal weighting), the ratio method had an average confidence interval length reduction of about 10% compared to equal weighting.

11.3. The mainstream's moving target

Suppose we have a sequence of meta-analyses where the urns described for obtaining the true study-specific effect sizes (Assumption A1) are identical, but each member of the sequence has within-study variances of 90% of the previous member of the sequence. Under the mainstream model, all of these meta-analyses have the same true effect sizes, namely the unweighted mean effect size in the urn. The true mainstream variance of the global effect size estimate is the sum of its between-study variance (Same for all members in the urn) and the within-study variance (which will shrink toward zero as you get later into the sequence). Thus, the mainstream estimates will become closer and closer to the unweighted estimator as we get further into the sequence. You therefore cannot rule out an artifact of where you might be in the sequence for any meta-analysis where the qualitative conclusions of the mainstream differ from the unweighted (one significant and one not). The Neto *et al*. [4] example is one case of this, but this is a very common occurrence. Note that if Assumption A4 is false, every weighted combination is estimating a different overall target population mean, and the mainstream analysis of the sequence will push the overall estimate toward targeting the unweighted mean in the urn. The key question when looking at the difference between the mainstream point estimate and the unweighted point estimate is whether it is simply sampling error or is it bias induced by failure of Assumption A5 (that the presumption that weights and effect sizes are uncorrelated is false). There is no way to be sufficiently certain, as there is no adequately powered statistical test that can prove the lack of such a correlation.

Note that the phenomenon of seeing two sets of data with the same signals but noise level of the second reduced by a common factor and turning the result from significant to not significant cannot occur in the common statistical methods: t-tests, analysis of variance or covariance, regression, logistic regression, frequency tables, or Cox regression (survival analysis).

11.4. Recommendations for meta-analysis of clinical trials with tabular data

(1) Use random effects rather than fixed effects; (2) With fewer than five studies, do a Systematic Review, not a formal meta-analysis, since the large sample distribution of the estimates should not be trusted; (3) with 5-20 studies, issue a limitation that the number of studies is small with a caveat on successful vetting for relative risk; (4) use Shuster [1] until new methods become available; (5) if you have individual patient data, note that the off-label implications for the mainstream for tabular data may or may not apply to individual patient data. Shuster [1] can still work if any of Assumptions A2, A4, and A5 are used in the individual patient analysis, potentially endangering its evidence base; (6) a biostatistics group with access to supercomputers should conduct large simulations along the lines of Shuster *et al*. [6] for the robustness of the T-approximation of the survey sampling method for differences in means or proportions; (7) a parallel width of confidence interval comparison on mainstream versus survey sampling should be done, and (8) if you have published a meta-analysis that had a substantive influence on public health policy, consider conducting an equal weighted analysis on the log of the relative risks or differences in means or proportions to see if this new analysis supports your original conclusions. If they do not, consider writing a report, using the recommended methods of Shuster [1].

12. Conclusion

Based on a reasonable fear gleaned from examples 1 and 2, the continued use of the mainstream methods is threatening to public health interests. In example 1, despite the published mainstream inference, there is no evidence that a widely used invasive intervention is effective. Example 2 had the survey sampling method been available and utilized, and the use of rosiglitazone in Type II diabetes would likely have been eliminated far earlier than what occurred, saving a substantial number of cardiac events from happening. Other similar misjudgments stemming from the mainstream methods are all but certain to occur in the future.

Biostatisticians accept the fact that an unlucky dataset can yield misleading results, but they cannot accept misleading results caused by the use of off-label statistical methodology.

Acknowledgments

The author wishes to thank Editor, Dr. Michal Heger for his critically important assistance in the write-up.

Funding

This paper was entirely self-funded. The author has personal licenses for all software used.

Conflicts of Interest

None.

References

[1] Shuster JJ. Meta-Analysis 2020: A Dire Alert and a Fix. Biostat Biom Open Access J 2021;10:73-8.

[2] Borenstein M. Common Mistakes in Meta-Analysis and How to Avoid Them. Englewood, NJ: Biostat Inc.; 2019.

[3] Nissen SE, Wolski K. Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes. N Engl J Med 2007;356:2457-71.

[4] Neto AS, Cardosa SO, Manetta JA, Pereira VG, Esposito DC, de Oliveira Prado Pasqualucci M, *et al*. Association Between Use of Lung-Protective Ventilation with Lower Tidal Volumes and Clinical Outcomes Among Patients Without Acute Respiratory Distress Syndrome: A Meta-Analysis. JAMA 2012;308:1651-9.

[5] Diamond GA, Kaul S. Rosiglitazone and Cardiovascular Risk. N Engl J Med 2007;357:938-9.

[6] Shuster JJ, Guo JD, Skyler JS. Meta-Analysis of Safety for Low Event-Rate Binomial Trials. Res Synth Methods 2012;3:30-50.

[7] Shuster JJ, Walker MA. Low-Event-Rate Meta-Analyses of Clinical Trials: Implementing Good Practices. Stat Med 2016;35:2467-78.

[8] Borenstein M, Hedges LV, Rothstein HR, Higgins JP. Introduction to Meta-Analysis. New York, NY: John Wiley and Sons; 2009.

[9] Shuster JJ. Nonparametric Optimality of the Sample Mean and Sample Variance. Am Stat 1982;36:176-8.

APPENDIX

(A) How to obtain free software:

To obtain free quality assured software to help you analyze a collection of randomized clinical trials, download, and save the three files from the Website

https://biostat.ufl.edu/research/faculty-developed-software/

There is a user-friendly User's Guide and two Excel templates, one for relative risk and one for differences in means and proportions. There are two real worked examples that guide you through the data input and interpretation of the results.

(B) This part of the appendix is for readers with some statistical training:

Quantification of the randomness of the between-study variance, a key component of the weights

Suppose we look at the true between-study variance in the urn (Assumption A1) denoted by σ^2 and suppose "an informer" had the true study-specific effect sizes for the completed studies in the meta-analysis. Suppose she denotes the unweighted sample variance of two independent repetitions of these true study-specific effect sizes by S_j^2 (j = 1,2) and provides us only with this value. Absent additional extraneous information, S_j^2 (j = 1,2) are each optimal estimates of its between-study variance (minimum variance unbiased) per Shuster [9]. It is superior to (less random) than the mainstream estimate of σ^2. It follows from Assumptions A1 and A2 that (M-1) S_j^2/σ^2 are independent and have Chi-square distributions with degrees of freedom (M-1) where M is the number of studies being combined. The ratio $F = S_2^2/S_1^2$ has a central F-distribution with degrees of freedom M-1 for both the numerator and denominator.

A meta-analysis of eight studies would have a 61% chance that the larger of the two sample variances would be at least 50% larger than the smaller. The between-study variance is therefore a seriously random variable making the mainstream weights, which rely heavily on the between-study variance, seriously random variables. The mainstream reliance on the weights being near constants is not supportable.

Why the ratio estimates are expected to perform well in their target population

The relative risk and difference of means and proportions are calculated as the ratio of sample means. Both the numerator and denominator are optimal (i.e., nonparametric minimum variance unbiased estimators per Shuster [9]) for their corresponding population parameter.

Publisher's note

AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

DOI: http://dx.doi.org/10.18053/jctres.09.202304.22-00019