Published in final edited form as:

Res Synth Methods. 2012 March ; 3(1): . doi:10.1002/jrsm.1039.

Meta-analysis of safety for low event-rate binomial trials

Jonathan J Shuster, Ph.D.^{1,*}, Jennifer D. Guo², and Jay S. Skyler, MD³

¹Department of Health Outcomes and Policy, College of Medicine, University of Florida, Gainesville, FL 32610

²Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, FL 32610

³Diabetes Research Institute and Division of Endocrinology Diabetes & Metabolism, University of Miami Miller School of Medicine, Miami, FL 33136

SUMMARY

This article focuses on meta-analysis of low event-rate binomial trials. We introduce two forms of random effects: (1) “Studies at Random” (SR), where we assume no more than independence between studies; and (2) “Effects at random” (ER), which forces the effect size distribution to be independent of the study design. Based upon summary estimates of proportions, we present both unweighted and study size weighted methods, which under SR target different population parameters. We demonstrate mechanistically that the popular DerSimonian-Laird (DL) method, as DL actually warned in their paper, should never be used in this setting. We conducted a survey of the major cardiovascular literature on low event rate studies and found that DL using odds ratios or relative risks to be the clear method of choice. We looked at two high profile examples from diabetes and cancer, respectively, where the choice of weighted vs. unweighted methods makes a large difference. A large simulation study supports the accuracy of coverage of our approximate confidence intervals. We recommend that before looking at their data, users should pre-specify which target parameter they intend to estimate (weighted vs. unweighted), but estimate the other as a secondary analysis.

Keywords

meta-analysis; random effects; odds ratio; relative risk

1. INTRODUCTION

This article is aimed at several potential audiences. First, it can be read by biostatisticians and epidemiologists with limited background in meta-analysis. In addition, clinical practitioners, who are either consumers of the information or conduct these types of studies themselves, can get a good feel for the concepts even if they do not fully grasp the formulas. Biostatisticians and epidemiologists, familiar with meta-analysis, can gain an unprecedented mechanistic viewpoint on how empirically derived weights (i.e. derived from estimates of variance) can fail in the low event-rate binomial trials scenario. Finally, and perhaps most importantly, biostatistical reviewers for journals need to be aware of the issues surrounding statistical practices for the most commonly used methods they will see for combining low event rate binomial trials. For readers who seek to understand the major points at a non-technical level, we have provided Appendix 1.

*Correspondence to: Jonathan J. Shuster, PO Box 100177, Gainesville, FL 32610-0177. jshuster@biostat.ufl.edu.

Meta-analysis (quantitative systematic reviews) of clinical trials deals with putting together the totality of experience about trials on an intervention or family of interventions, to make an inference about the “average” effect of that intervention. If one Googles “Evidence Pyramid”, this entry is at the very top, ahead of randomized clinical trials in most versions. Hence, it is extremely important for readers to understand that studies of safety issues, where events are relatively rare, have been inappropriately analyzed in many, if not most cases. “Fixed effects” methods presume that the true effect sizes in every trial are identical. One can think of a multicenter drug trial with well defined eligibility, follow-up, and primary endpoint as a good example where a fixed effects method is reasonable. In most scenarios, “random effects” methods should be employed. These methods allow for differing effect sizes from study to study, and attempt to estimate a global effect size for the intervention. If done correctly, random effects methods are valid whether or not fixed effects methods are valid, but they will tend to have less precision when indeed fixed effects are true. However, when combining trials, there is no diagnostic test that can prove that fixed effects are valid. For example, the Cochran Q-test for heterogeneity vs. homogeneity of effect sizes, per Borenstein et. al. (2009), can lead to two functionally equivalent conclusions: homogeneity is rejected, or the results are inconclusive as to homogeneity. This test never has sufficient power to conclusively determine whether or not homogeneity exists. In fact, Borenstein et. al. (2010) state, “The strategy of starting with a fixed-effect model and then moving to a random-effects model if the test for heterogeneity is significant relies on a flawed logic and should be strongly discouraged.” When combining a moderate to large number of trials, they are almost never true replications of each other, but rather the individual trials will have differences in eligibility, dosage, follow-up, control medications, concomitant medications, and geographical locale. With this diversity of designs, it stands to reason that fixed effects will almost certainly not be true. When it comes to conclusions about subject safety, we recommend minimizing the risk of reaching conclusions that may be artifacts of erroneous assumptions, even if this reduces study power when those assumptions are indeed correct. A reviewer questioned this conservative approach on the grounds that when it comes to safety it might be better to err on the side of caution. While we respect this point of view, the implications of significant findings, especially when they relate to side effects, can be enormous and subject the intervention under investigation to multiple jeopardy in both time and in the number of different side effects that might be monitored. Our view is that meta-analysts must never compromise on the soundness of their science. When the event rates are low, Section 16.9.5 of the Cochran Handbook [Higgins and Green (2011)], suggests that there are no satisfactory random effects methods and users should employ fixed effects methods to detect the signal (hypothesis testing only). While fixed effects do provide a valid test of the narrow null hypothesis that all of the effect sizes are simultaneously zero, as Shuster (2011) noted, the methods cannot be relied upon for correct point and interval estimates of the overall effect size when effects are random. This statement on lack of methods is not entirely true, as Shuster et. al. (2007) published a method for large numbers of low event trials being combined, Stijnen et. al. (2010) published a Bayes-likelihood method, and Rucker et. al. published a method based upon the angle transformation. Earlier, Emerson et. al. (1993, 1996) presented an excellent method based upon risk differences.

In Section 2, we shall define the typical statistical model used in meta-analysis, and the typical methods used in the medical literature to analyze these. It needs to be noted that the central parameter in this model is an unweighted mean, and therefore methods that target this parameter seek to estimate an unweighted mean of all of the effect sizes in the target population. If there is no association between the effect sizes and the study design parameters (most notably sample size) then weighted and unweighted methods target the same parameter, making the distinction irrelevant. However, in the likely event that such an association does exist [See Shuster (2011) for scenarios where such associations are

expected], then weighted and unweighted methods both have counterintuitive elements. It seems to be logical that more weight should be placed upon larger studies. Yet the literally targeted parameter of the model is the unweighted mean, meaning that the weighted method is estimating something other than the theoretically targeted one. In Sections 3 and 4, we shall demonstrate how one can recognize the distinction, and by creating an alternate parameter to target in Section 4, we can resolve any counterintuitive issues. Furthermore, the targeted metric (for example relative risk), whether weighted or unweighted, can be given a very clear physical interpretation. Section 3, presents a modification of methods that Shuster et. al. (2007) proposed, covering unweighted methods for relative risk. Section 4 presents a weighted approach, largely based upon the weighted approach of Shuster (2010a, Section 3). These sections will emphasize relative risk estimation, but we cover odds ratios and risk differences in Appendix 2. Both our unweighted and weighted methods will rely upon functions of summary estimates of proportions, rather than study level estimates of effect size. As such, they avert several difficult issues when compared to study-specific effect estimation. Section 5 presents a mechanistic illustration where weighted methods that rely on study level estimates of relative risks, including the popular DerSimonian-Laird (DL) method (1986), break down. In fairness to DerSimonian and Laird, we do not believe they ever intended that their methods would be used in the low event rate scenario. See the last paragraph of the discussion in that paper, and the quotation from Laird et. al. (2010), “They conclude that DL does indeed have bias, especially when within sample sizes are small and event rates are low.” In Section 6, we present a random survey of low event rate binomial meta-analyses from cardiovascular disease, published between 2005 and 2010. The survey demonstrates that the DL method is the overwhelmingly most common approach, that the metric used is nearly always either the Odds Ratio (OR) or the Relative Risk (RR), and that the DL calculated confidence intervals in the log scale tend to be considerably shorter than those from the unweighted methods. Although DL tracks fairly well with our weighted approach of Section 4, there are notable differences. As seen in this survey, users also seem to be totally unaware that the DL method needs both large samples and large numbers of studies being combined in the low event situation. Section 7 presents two high profile medical applications where unweighted methods lead to substantially different conclusions than weighted methods. From these, we see that how one defines a target parameter can lead to drastically different conclusions about an important public health question. In Section 8, via simulations, we study the properties of the estimates and confidence intervals when the number of studies being combined is small (5–20). These simulations support the accuracy of the coverage of the methods under these circumstances. The final section is devoted to a discussion. This paper does not make a general recommendation as to whether to use weighted or unweighted methods. However, the choice should be based upon which targeted parameter makes more sense in a particular application, and for objectivity, that choice should be declared before looking at any data.

2. MODEL FOR META-ANALYSIS

This section will be useful in general for meta-analysis including, but not limited to binomial applications.

The classical model for meta-analysis, which can be multivariate (See for example DerSimonian and Laird (1986), van Houwelingen et. al. (1993), Emerson et. al. (1993, 1996), Burr and Das (2005), Borenstein et. al. (2009, 2010), Higgins et. al. (2009), Senn(2010), Shuster (2010a), or Stijnen et. al. (2010) is

$$\hat{\theta}_j = \theta_j + \varepsilon_j \quad (1)$$

where $j=1,2,\dots,M$ identifies the study (total observed studies= M), the estimated study effect sizes θ_j are independent, the true study specific target population effect sizes θ_j form a random sample from a large population whose mean is θ , and conditional on the j -th study, the target population mean of the ε_j is zero $\{E(\varepsilon_j)=0\}$. Under this representation that the $\{\theta_j\}$ indeed form a random sample from this large population, the parameter θ is the true **unweighted population mean** of the effect sizes in this population. Note that there is an implicit “prior distribution” of the θ_j , and if this distribution has a variance of zero, $\text{Var}(\theta_j)=0$, we have **fixed effects**.

To keep assumptions to a minimum, we shall not impose any form on the prior distribution under our methods. Bayes advocates will make some assumptions on this prior distribution, but in some situations, these assumptions become unimportant as the number of studies in the meta-analysis, M , becomes large.

Most papers and texts do not discuss how the model (1) might be thought of as arising, but there are two and only two conceptual models we shall consider.

Studies at Random (SR)

All that is assumed under SR is that the studies are independent. However, operationally, it helps to visualize the studies as being drawn from a conceptual urn containing a large number of studies. This is the model that Shuster et. al. (2007) work under.

Effects at Random (ER)

Under ER, the θ_j are drawn at random from a conceptual urn, but the studies are fixed. In other words for each fixed study, we visualize that a true effect size is drawn at random, and that the study is conducted under this effect size. This is the operational model that Emerson et. al. [1993, 1996] work under. This model makes assumptions over and above SR, namely that the distribution of the random effects, θ_j is independent of the study design, including error properties of the ε_j .

Unweighted Estimation

We use

$$\hat{\theta}_U = \sum_j \hat{\theta}_j / M \quad (2)$$

with variance-covariance matrix (variance if univariate) based on the first two moments of the $\hat{\theta}_j$.

Note that since the estimate and the first two sample moments together are invariant under a random permutation of the indices $1,2,\dots,M$, and that after such a random permutation of these indices, the $\hat{\theta}_j$ become identically distributed, we can apply SR to the unweighted method, even when in actuality the assumed model is ER. Higgins et. al. (2009) have also utilized the idea of a random permutation of the indices to ensure this exchangeability in their approach to ER.

Note further that the unweighted estimate has as its variance-covariance matrix the variance covariance matrix of $\hat{\theta}_j$ divided by M . The sample variance-covariance estimate of the $\hat{\theta}_j$ is a consistent estimate for the actual variance-covariance matrix of the θ_j .

Bonett (2009) has an interesting application of unweighted methods, where he uses $\hat{\theta}_U$ from (2) to estimate a conditional parameter $\theta = \sum_j \theta_j / M$. From (1), the only source of variation in

estimating $\hat{\theta}$ is the within study error, making this approach a fixed effects analysis, with no accounting for a difference between θ (the true sample mean of within study means actually sampled) and θ_j (the population mean of the true study level sample means in the universe of studies). In reality this is a compromise between fixed and random effects, which works well in some scenarios, but as noted in Section 5, should be avoided when low event rate binomial data are analyzed.

Weighted Estimation

In this subsection, we shall consider two weight formulations, one empirical (based on estimated variances and therefore weights are random variables) and one sample size based (and therefore weights are non-random in the ER model).

Here, we use

$$\hat{\theta}_w = \sum_j W_j \hat{\theta}_j \quad \text{with } \sum_j W_j = 1. \quad (3)$$

For univariate applications, we use

$$\text{Var}(\hat{\theta}_w) = \sum_j W_j^2 \text{Var}(\hat{\theta}_j) \quad (4)$$

Notes: (A) For empirically based weights, that is, weights based upon the estimated variance within and between studies, the weights, W_j are random variables, and not fixed entities. The variance formula (4) only takes into account the variability due to the effect size estimates and not upon the weights. These variances in the weights will be seen in Section 5 as substantial in low event rate binomial estimation of the relative risk (and similarly the odds ratio). For sample size based weights, per Emerson et. al. (1993, 1996), $\hat{\theta}_w$ is indeed unbiased, and the variance formula (4) is valid for ER.

(B) Things can break down for both empirically based weights and for sample size based weights when in fact the more general SR model holds. As pointed out in Shuster (2010a), for ER to hold, the further assumption, that there is no association between the random or non-random weights, W_j , and the estimates, $\hat{\theta}_j$ must be true. Again, to get a clear view of the issue, imagine we randomly reshuffle the indices $1, 2, \dots, M$ in (3). Now, even in the sample sized based weight scenario, we have made the bivariate $(W_j, \hat{\theta}_j)$ identically distributed vectors. The permutation makes the W_j random. The estimate $\hat{\theta}_w$ is unbiased if and only if after a random permutation of the indices $1, 2, \dots, N$

$$E(W_j \hat{\theta}_j) = E(W_j)E(\hat{\theta}_j).$$

That is, if and only if the weights and effect size estimates, after this permutation, are uncorrelated. The concern under SR, even when the weights are sample size based, is that there may be a systematic association between effect size and study size. In other words, if the inherent assumption that study size and true study effect size are unrelated is false, ER-based methods break down even when sample size based weighting is used. An interesting point made in Shuster (2010a), is that the unweighted estimate is a biased-corrected weighted estimate (irrespective of the weights), under a random permutation of the indices $1, 2, \dots, M$.

(C) One interesting approach for binomial applications that makes empirical weighting and sample size weighting almost identical is to employ the angle transformation as a metric:

$$\sin^{-1} \sqrt{P_2} - \sin^{-1} \sqrt{P_1} \quad (5)$$

This variance stabilization transformation unifies empirical weighting and sample size weighting to some extent. Since this metric is hard to interpret in safety studies, and has not enjoyed wide use, we shall not discuss this metric further. Nonetheless, this metric is worthy of further study. See Rucker et. al. (2009) for more detail about this approach and a graphical interpretation for this metric.

3. UNWEIGHTED METHODS USING TREATMENT SUMMARY PROPORTIONS

In this section, we shall derive unweighted (Study Level) methods for estimating a population relative risk. (Methods for estimating odds ratios and risk differences are deferred to Appendix 2).

Unweighted Relative Risk Interpretation

The unweighted method targets the study as the sampling unit, rather than the patient. For Studies at Random (SR), imagine a conceptual urn of studies in the population. The targeted overall relative risk can be thought of as follows: Pick a study at random from the urn (each study having the same chance of being drawn), and a yet to be assigned patient randomly from the selected study's subjects. What is the ratio of failure probabilities if the patient is randomly assigned to Treatment #2 to that if randomly assigned to Treatment #1?

For effects at random (ER), the interpretation is similar. The hypothetical experiment for interpreting relative risk is to pick a study at random from the fixed collection of studies (each study having the same chance of being drawn), and draw an effect size from the hypothetical urn of effect sizes. Again we randomly select a yet to be assigned patient from the selected study, and ask what is the ratio of failure probabilities if randomly assigned to Treatment #2 to that if randomly assigned to Treatment #1.

The first step is to obtain summary measures of the two target population proportions, and then estimate the relative risk as the ratio of these two proportions. As we shall see in Section 5, using the summary proportions, rather than the study level logs of the relative risks, prevents some difficult issues from arising from low event rate binomial studies.

$$\text{Let } \hat{P}_{ij} = F_{ij}/N_{ij} \quad \text{where}$$

F_{ij} and N_{ij} represent the number of failures and sample size for Treatment i and trial j .

We shall write a model for $\hat{P}_{ij}$ and compare it to that given in (1).

$$\hat{P}_{ij} = P_{ij} + \varepsilon_{ij} \quad (6)$$

$$E(P_{ij}) = \pi_i \quad (7)$$

Conditional on study j ,

$$E(\varepsilon_{ij})=0 \quad (8)$$

If one compares the above notation with the bivariate version of the model (1) above, one notes

$$\theta_j=(P_{1j}, P_{2j}), \hat{\theta}_j=(\hat{P}_{1j}, \hat{P}_{2j}), \theta=(\pi_1, \pi_2), \text{ and } \varepsilon_j=(\varepsilon_{1j}, \varepsilon_{2j}),$$

and that this model holds with no assumption beyond the SR (which as noted, includes ER as a special case).

$$\text{Point estimates: } \hat{\pi}_i = \sum_j \hat{P}_{ij} / M \quad \text{for } i=1, 2$$

Based on the central limit theorem for independent, identically distributed vectors, for a large number of studies, M , $(\hat{\pi}_1, \hat{\pi}_2)$ has an asymptotic bivariate normal distribution with mean (π_1, π_2) and variance covariance matrix

$$V = \text{Cov}(\hat{P}_{1j}, \hat{P}_{2j}) / M \quad (9)$$

A consistent estimator for this covariance matrix, V , denoted by $\hat{V}_M$, is defined as

$$\hat{V}_M = \|C_{kl}\| / M \quad \text{with } k=1, 2 \quad l=1, 2$$

$$C_{kl} = \sum_j \{ \hat{P}_{kj} - \hat{\pi}_k \} \{ \hat{P}_{lj} - \hat{\pi}_l \} / (M-1), \quad (10)$$

Relative Risk Estimation

$$RR = \pi_2 / \pi_1$$

Point Estimate:

$$\widehat{RR} = \hat{\pi}_2 / \hat{\pi}_1$$

$$Z_{LRR} = \{ \text{Log}(\widehat{RR}) - \text{Log}(RR) \} / SE$$

has an asymptotic t-distribution with $M-1$ degrees of freedom, with the standard error SE defined below. Based on a large array of empirical studies (See Section 8), we have determined that this t-distribution has much more accurate coverage for small to moderate numbers of studies being combined, than does the asymptotically equivalent normal distribution.

$$SE = \sqrt{\{(Q_1 + R_1 - 2S_1)/M\}} \quad \text{with}$$

$$Q_1 = C_{11}/[\hat{\pi}_1]^2$$

$$R_1 = C_{22}/[\hat{\pi}_2]^2$$

$$S_1 = C_{12}/[\hat{\pi}_1 \hat{\pi}_2]$$

We can form an approximate $100(1-\alpha)\%$ confidence interval for the parameter $\text{Log}(\text{RR})$ as

$$\text{Log}(\widehat{RR}) \pm T(\alpha/2, M-1)SE \quad (11)$$

where $T(\gamma, \lambda)$ is the 100γ upper percentile of the central T-distribution with λ degrees of freedom.

To obtain the point estimate and confidence interval in the original scale, we take natural antilogs (exponentiation) of the point and interval estimates in the log scale.

To compute a P-value for the test $\text{RR}=1$, i.e. $\log \text{RR}=0$. We compute

$$Z_{\text{LRR}} = \text{Log}(\widehat{RR})/SE \text{ with } P\text{-Value} = 2\text{PROBT}(-|Z_{\text{LRR}}|, M-1), \quad (12)$$

where $\text{PROBT}(\gamma, \lambda)$ is the cumulative central T-distribution with λ degrees of freedom.

4. WEIGHTED ESTIMATION OF A WEIGHTED PARAMETER

In this section, we shall derive weighted (Patient Level) methods for estimating a population relative risk. (Methods for estimating odds ratios and risk differences are deferred to Appendix 2).

Weighted Relative Risk Interpretation

The weighted method targets the patient as the sampling unit, rather than the study. For Studies at Random (SR), imagine a conceptual random selection of a yet to be assigned **patient** from the collection of all potential patients in the conceptual urn of studies, with each patient having the same chance of being selected. What is the ratio of failure probabilities if the patient is randomly assigned to Treatment #2 to that if randomly assigned to Treatment #1?

For effects at random (ER), the interpretation is similar. The hypothetical experiment for interpreting relative risk is to pick a yet to be assigned patient at random from the fixed collection of studies, draw an effect size at random from the conceptual urn of effect sizes for the selected patient's study. What is the ratio of failure probabilities if the patient is randomly assigned to Treatment #2 to that if randomly assigned to Treatment #1?

Note that the definitions for weighted and unweighted relative risks match under ER, but differ under SR.

Using the model of Section 2, the target parameter for a weighted analysis will be,

$$\theta_w = \sum_j W_j \theta_j$$

where summation is over all trials in the universe, and W_j represents the fraction of subjects in the universe that belong to Trial j . As we shall see, the θ_j will be two dimensional for relative risk. In Appendix 2, it will be four dimensional for the odds ratio and two-dimensional for the risk difference. Under SR, the parameter θ_W is completely different from the θ defined in Section 2, unless all trials were exactly the same size.

Using the notation from Section 3, if the relative weights for the sampled studies are defined as

$$U_j = (N_{1j} + N_{2j})/2, \quad (13)$$

the weighted estimator of θ_W is

$$\hat{\theta}_W = \sum_j U_j \hat{\theta}_j / \sum_j U_j \quad (14)$$

For $i=1,2$ and $j=1,2,\dots,M$ let

$$A_{ij} = U_j \hat{P}_{ij} \quad \text{be the adjusted number of events.}$$

Let us define sample means as follows:

$$\bar{A}_i = \sum_j A_{ij} / M \quad \text{and} \\ \bar{U} = \sum_j U_j / M$$

Since the estimated proportions, $\hat{P}_{ij}$ have a common denominator, the **relative risk** is estimated as the following ratio.

$$\widehat{RR} = \bar{A}_2 / \bar{A}_1. \quad (15)$$

This represents a consistent estimator of the relative risk we defined for the definition in the first paragraphs of this section. Furthermore, using the delta method and the central limit theorem for independent, identically distributed vectors,

$\log(\widehat{RR})$ is asymptotically t-distributed ($M-2$ df) with mean $\log(RR)$ and asymptotic variance.

$$SE^2 = [\{S(A_{1j})/\bar{A}_1\}^2 + \{S(A_{2j})/\bar{A}_2\}^2 - 2\{C(A_{1j}, A_{2j})/(\bar{A}_1 \bar{A}_2)\}]/M \quad (16)$$

where $S()$ represents the sample standard deviation and $C(,)$ represents the sample covariance. The standard error of $\log(\widehat{RR})$ is $\text{SQRT}(SE^2)$.

By asymptotic t, we mean $[\log(\widehat{RR}) - \log(RR)]/[SE]$ is approximately central t-distributed with $M-2$ df for large M , the number of studies in the meta-analysis. (This is asymptotically equivalent to asymptotic normality, but empirically gives more accurate approximations than those based on normality or with the T-distribution with $M-1$, $M-1.5$, $M-2.5$, or $M-3$ degrees of freedom.)

As in the previous section, an approximate $100(1-\alpha)\%$ confidence interval for $\text{Log}(\text{RR})$ is

$$\text{Log}(\widehat{RR}) \pm T(\alpha/2, M-2)\text{SE} \quad (17)$$

where $T(\gamma, \lambda)$ is the 100γ upper percentile of the central T-distribution with λ degrees of freedom.

To obtain the point estimate and confidence interval in the original scale, we take natural antilogs (exponentiation) of the point and interval estimates in the log scale.

To compute a P-value for the test $\text{RR}=1$, i.e. $\log \text{RR}=0$. We compute

$$Z_{\text{LRR}} = \text{Log}(\widehat{RR})/\text{SE} \quad \text{with P-Value} = 2\text{PROBT}(-|Z_{\text{LRR}}|, M-2), \quad (18)$$

where $\text{PROBT}(\gamma, \lambda)$ is the cumulative central T-distribution with λ degrees of freedom.

Interesting note: If within every study, the numbers of subjects on each treatment arm were exactly equal, ($N_{1j} = N_{2j}$) the point estimates would be exactly the same as if one collapsed all of the data into a single 2 by 2 table. Furthermore, the estimate of relative risk would be the ratio of the number of failures Treatment #2:Treatment #1. However, the sampling properties are evaluated under random effects by the methods of this section.

5. CRITICAL EVALUATION OF THE CONTRIBUTION OF A SINGLE STUDY AND ITS IMPLICATIONS

This section deals strictly with properties within a single study, in terms of study-specific relative risk estimation. To make things very concrete, we shall specify the study true parameters for a low event-rate situation, and investigate how elements critically important to the validity of the use of these estimators are seriously violated. Moreover, we shall see a strong relationship can exist between the purported weight and effect size rendering inverse variance based estimates especially problematic when incorporated into a meta-analysis. Specifically, we shall examine the contribution of a single randomized trial, where $N=100$ patients are randomly assigned to each of two treatments, and we shall estimate the relative risk as (a) $\widehat{RR} = \widehat{P}_2/\widehat{P}_1$ the ratio of the observed failure rates (Treatment #2:Treatment #1) when there are no zero event results, or (b) this same ratio after adding 0.5 to each cell in the 2 by 2 table when one or both study arms have no events. The performance of the estimate of $\log(\text{RR})$, $\log(\widehat{RR})$, and its estimated variance will be investigated under three modes of operation. This continuity correction is the default in the two major software packages (Comprehensive Meta- Analysis and REVMan 5).

Mode 1: No exclusions

Mode 2: Exclude studies with no events on both treatment arms (most common method)

Mode 3: Exclude studies with no events on at least one treatment arm (This will illustrate that zero event treatment arms are not the only impediments to getting proper estimates and their standard errors when event rates are low. This also will indicate that no continuity correction strategy can ameliorate the issues of low events.)

Parameters of interest include (A) Bias in estimated $\log(\text{RR})$; (B) Coefficient of variation in the estimated variance of the estimate (This drives the weights and it is ideally zero for empirically weighted methods.); (C) Ratio: the quotient of the expected value of the

estimated variance of $\log(\widehat{RR})$ to the true mean square error of the estimate of $\log(RR)$ (This is supposed to be one. Values above one vs. below one give one a grasp on whether there is a trend toward overestimation or underestimation respectively, of the critically important within study variability.); and (D) Correlation between the estimated $\log(RR)$ and $(1/\text{Est_var})$, the purported relative weight that would be used in either a fixed effects analysis or a random effects analysis where the Cochran Q chi-square statistic had value less than its degrees of freedom (This correlation is supposed to be zero for empirically weighted methods to work). More formally:

$$\text{Bias} = E[\log(\widehat{RR})] - \log(RR) \quad (19)$$

where $E[\]$ =expected value.

$$\text{CV_Var} = 100\text{SQRT}\{\text{VAR}(\text{Est_var})\} / E[\text{Est_var}] \quad (20)$$

where $\text{Est_var} = [(1 - P_1)/\{100 P_1\}] + [(1 - P_2)/\{100 P_2\}]$, the classical large sample variance estimate of $\log(\widehat{RR})$ where sample size=100, and CV=coefficient of variation.

$$\text{Ratio} = E[\text{Est_var}] / \text{MSE}[\log(\widehat{RR})] \quad (21)$$

with MSE=Mean Square Error, the expectation of the square of the difference between the estimate, $\log(\widehat{RR})$, and its true value, $\log(RR)$.

Table 1, compiled by exact calculations using SAS 9.2, supplies some noteworthy entries. Note that a bias of the order of 0.1 (−0.1) in estimating $\log(\widehat{RR})$ translates into tendency of about a 10% overestimation (underestimation) respectively of the relative risk. Next, note that the large coefficients of variation in the variance estimates indicate that the true variances of the individual low event rate studies cannot be viewed as fixed (non-random as required by empirically weighted methods). Third, for the first two sets of examples (2% vs. 1% and 4% vs. 2%), the large values of “ratio” tell us that the typically used estimates of variance perform poorly. Fourth, note that eliminating the zeros does not conditionally improve things substantially. In Mode 3, there are no continuity corrections, and yet, the estimated variances still have large coefficients of variation. Finally, and perhaps most striking, there is a strong correlation between $\log(\widehat{RR})$ and the reciprocal of the estimated variance (the relative weight). The larger estimates tend to go with lower weights, clearly inducing a bias in the meta-analysis. Ideally, these weights should not depend on the outcomes, given that the true target population parameters are fixed entities.

In short, methods such as DL, that rely on estimated variances to determine the weights, fail on several key ingredients when they wish to apply formulas (3) and (4), which ignore sampling variation in the weights, and plausible associations between these weights and the study specific point estimates. The violations in the Bias and Ratio columns of Table 1 mean that the classical model (1) is false for low event binomial trials with individually estimated logs of relative risks and this should dissuade users even from applying non-random weighting schemes for study level log relative risks. The method of Bonett (2009) also must rely on within study error properties and is likewise problematic in the setting of low event-rate estimation of relative risk. Note also for low event-rate trials, the identical issues will occur for odds ratios, which in any case are tightly related to relative risks in this scenario.

6. A SURVEY OF THE CARDIOVASCULAR LITERATURE: SCOPE OF THE PROBLEM

6.1. Cardiovascular Pilot study (Design)

From the 69 months of the study window (1/2005 to 9/2010 inclusive), we randomly selected six months of issues from the eleven top journals that publish cardiovascular related meta-analyses involving low event-rate binomial trials [six high impact journals (*Annals of Internal Medicine*, *Archives of Internal Medicine*, *British Medical Journal*, *Journal of the American Medical Association*, *The Lancet*, and *The New England Journal of Medicine*) and five high impact cardiovascular specialty journals (*Circulation*, *Journal of the American College of Cardiology*, *European Heart Journal*, *Heart Rhythm*, and *Heart*)]. These turned out to be Feb05, Oct06, March08, April08, March09, and May09. All cardiovascular related meta-analyses from the eleven journals involving binomial endpoints were examined, and 15, as detailed in Table 2, involved at least one low-event rate study (expected number of events in at least one cell was below 5.0). For each paper, only the first meta-analysis that met our eligibility requirement of low events was reanalyzed in the survey.

6.2. Cardiovascular Pilot study (Results)

This number (15) of low event meta-analysis articles found in our pilot study projects to an estimated 30 such articles per year in these journals.

Of the fifteen papers, fourteen used random effects for at least some of their analyses and the other, De Ferrari et. al. (2009), used fixed effects after testing for heterogeneity via Cochran's Q. Only one, Mottillo et. al. (2009), used a Bayes approach. Of the 14 non-Bayes analyses, seven Myung et. al (2009), Stripolli et. al. (2008), Piccini et. al. (2009), Abbate et. al (2008), Brar et. al. (2009), De Luca et. al. (2009), and Berger et. al. (2009) expressly used DL, but in checking the other seven non-Bayes approaches, five including de Denus et. al. (2005), Bath et. al. (2005), Verdecchia et. al. (2009), Bavry et. al. (2006), and Bellamy et. al. (2009) had results identical to DL for the first meta-analysis involving low event rates in the paper. To confirm, we used the default in Comprehensive Meta-Analysis (CMA) version 2.0, <ftp://140.131.114.243/B1%D0%A7%F7/%AD%F0%A4%E9%B7s%A6%D1%AEv/97%BD%D7%A4%E5/CMA/Meta-Analysis%20Manual.pdf>, which is DL, with continuity correction if needed per subsection 3.1 of Sweeting et. al. (2004), which adds 0.5 to each of the four cells of the 2by2 outcome table that has one zero event cell, and excludes studies with no event on both treatments. One of the others, Collet et. al. (2006) had an identical point estimate and identical lower confidence limit for the OR but had an upper limit of 0.99 (DL gives 1.01). However, the published P-value for this odds ratio was 0.055, something not consistent with excluding 1.00 from the confidence interval. We could not come close to reconciling the De Ferrari et. al. (2009) results and strongly suspect miscalculations by these authors. Note that all 15 used either relative risks (7) or odds ratios (8) as their metric. In short, we conclude that in cardiovascular medicine, the DL method, using OR or RR as the metric is nearly universal (13+/15 used or intended to use DL, whether explicitly referenced or not), and that this raises grave concerns about the evidence basis of past meta-analyses (and potentially future ones) in this field.

Of the papers that used DL, none had any warnings about the asymptotic properties, either in terms of the number of studies or low event rates within studies. As seen in Table 2, only three of the 13 papers that employed DL had 20+ studies in the meta-analysis we displayed. Four papers did discuss small numbers of studies or small sample sizes within studies, but only in the context of power.

When we contrast the DL result to the unweighted result, we note large differences between the point estimates and widths of the confidence interval in the log scale. The DL results have much shorter intervals. However, our weighted approach tracks fairly closely with DL, with notable exceptions, Abbate et. al. (2008) and Bellamy et. al. (2009).

7. APPLICATION TO HIGH PROFILE META-ANALYSES THAT AFFECTED PUBLIC POLICY

7.1 Rosiglitazone (Avandia)

This subsection is not designed to second guess any decisions made at and after the FDA hearings on rosiglitazone in July 2010, where because of concerns about increased cardiovascular side-effects of rosiglitazone, (i) the advisory committee recommended much stronger restrictions on the use of the drug (7/2010); (ii) the FDA, without consultation with the Data and Safety Monitoring Board, closed the “TIDE” trial to further accrual (7/2010); and (iii) the FDA virtually removed rosiglitazone from the market (9/2010). TIDE would have randomized a substantial number of subjects to either rosiglitazone or pioglitazone. These decisions were made on much more information than that derived in the meta-analyses. Yet the very fact that these meta-analyses were quoted in many of the presentations and as the primary evidence quoted in the news media indicates they were very important factors in these three decisions. After all, it was the meta-analysis in Nissen-Wolski (2007) that triggered FDA hearings in 2007. The aim of this subsection is to present this high profile example which demonstrates the importance of properly presenting the nature of the parameter being estimated.

Nissen-Wolski (2007) published a meta-analysis of 48 rosiglitazone trials that purportedly demonstrated a significant increase in the risk of myocardial infarction (MI) over controls, based on a fixed effects meta-analysis, using the method of Peto. They also concluded that there was a non-significant trend for increased cardiac deaths (CD). Shuster et. al. (2007) reanalyzed these trials by unweighted methods similar to those in Section 3, and found a different conclusion for both endpoints: no significant increase in risk of MI, but a significant increase in risk of CD. In the FDA hearing in 2007, the panel voted to keep rosiglitazone on the market with additional black box warnings. Nissen-Wolski (2010) updated their meta-analysis to 56 studies, without reporting either our type of analysis, Shuster et. al. (2007) or the default Comprehensive Meta-Analysis (CMA), version 2.0 as used by the Diamond et. al. (2007) re-analysis of Nissen-Wolski (2007). Nissen-Wolski (2010) in fact also used CMA, except they used Cochran Q to decide between fixed and random effects, and with $P > 0.10$, they utilized fixed effects, based on the Peto method, as they did in their 2007 paper.

Table 3 contrasts four analyses using the updated 2010 results. On the upper part of each cell, this is done with no exclusions. The FDA hearings of 2010 spent considerable time looking at the sensitivity with vs. without the RECORD trial per Home et. al. (2009). Hence, in the lower part of each cell in Table 3, we also present the analyses without the RECORD trial. This provides an example of the robustness of unweighted methods when applied to a large number, M , of trials contributing to the analysis, when large trials of questionable quality are excluded in a sensitivity analysis. Since each trial carries weight $1/M$, no single trial will dominate an unweighted analysis. Note that for Cardiac Death, the point estimate, after removal of the RECORD trial, increased by 42%, 27%, and 49% for the three weighted analyses, Nissen-Wolski [33], default CMA, and our own based upon Section 4, respectively. Of these, only CMA had its new point estimate contained in the original confidence interval, and that value of 1.23 was only slightly below the 1.28 upper limit. Perhaps most spectacular is the increase in the width of the confidence intervals in the log

scale (where the actual inferences are derived in all four analyses). Removal of RECORD, just one of the 56 studies, though the largest, has marked effects for all of the weighted methods, but much more modest impact on the unweighted estimate, indicating, at least in this instance, better stability of the unweighted estimation procedure.

It is of interest to note that although for both endpoints, the qualitative conclusions of Nissen-Wolski (2007, 2010) were internally consistent over time, they both disagreed qualitatively with the unweighted results in Shuster et. al. (2007) and in the present analysis.

7.2 Erythropoiesis Stimulating Agents

In this subsection, we shall reevaluate the results of Bohlius et. al. (2009), with respect to potentially increased mortality while on study of erythropoiesis stimulating agents vs. control. While these agents are used for indications other than supportive care in cancer, this paper only dealt with cancer patients. They are prescribed to shorten the time patients are anemic (low red blood cell counts) on chemotherapy and/or radiation therapy. This study combined 53 randomized cancer trials and found a highly significant increase in mortality while on study on these agents as compared to controls. The FDA, in part on the basis of these data, issued a recommendation that these agents should not be used as part of cancer treatment with curative intent. This action eliminated a major portion of the multi-billion dollar annual sales of these agents.

Although the analysis of Bohlius et. al. (2009) was based upon actual patient level data, and unable to obtain such data, we were forced to utilize study level data, this is not a major issue in low event-rate studies, as long as the follow-up is equally rigorous on both treatment arms. In such situations, the ratio of person-years at risk within a study (Treatment #2: Treatment #1) is closely approximated by the ratio of the sample sizes. Table 4 provides the outcomes for four analyses: the published hazard ratio (instantaneous relative risk) analysis, the analogous relative risk analysis for binomial data using the default method for CMA, an unweighted analysis of relative risk based on Section 3, and a weighted analysis of relative risk based on Section 4. We note the close agreement between the Bohlius et. al (2009) result that uses patient level data and the other two weighted results that use study level data.

In short, it seems clear that the actual published result is not strictly estimating the unweighted parameter, as presumed in the classical model (1), but something more akin to a weighted parameter. The qualitative conclusion in Bohlius et. al (2009) is probably correct, but not in the unweighted metric as their methods tacitly claimed.

8. SIMULATIONS OF COVERAGE FOR WEIGHTED AND UNWEIGHTED METHODS

In this section, we performed three sets of simulations for the assembly of 5–20 studies in a meta-analysis to help determine how the large sample approximations might work in practice. The bottom line is that (1) the t-approximations of Sections 3 and 4 performed well; (2) the approximations for the unweighted methods give somewhat more consistent coverage (closer to 95%) than the weighted, and (3) the weighted methods performed somewhat better in terms of shorter confidence intervals, under the special case that the two methods are estimating the same target parameter. The normal approximations performed poorly throughout, in terms of coverage. Except for a breakdown by sample size, we have no interest in individual parameter settings within a particular set of simulations. Rather, we consider them as typical, with diverse sample sizes, failure rates, and degree of randomness in the individual proportions. The ranges are informative as to the worst case scenarios while the means give one an idea of typical coverage for this wide range of settings.

8.1. Description of Simulation #1

The idea behind this simulation is to fix the number of studies being combined, the true mean proportions for failure on each treatment arm, a “diversity parameter” for random effects, and the range of total sample sizes. On each study, patients are randomized to the two treatments by independent coin flips, except that the actual sample size assigned to each arm is constrained to be between 30% and 70% of the total. For convenience, total sample sizes within an individual study in the meta-analysis are independent, uniformly distributed over a pre-specified sample size range from NL to NH. The diversity parameter, D , is used to create the random effects. If π_1 and π_2 are the global target population mean proportions for the treatment groups, the individual study (j) proportions P_{ij} are generated as independent uniformly distributed random variables over $\pi_i (1-.5D)$ to $\pi_i (1+.5D)$ for Treatment# $i=1,2$. The true relative risk for the unweighted analysis (Study based RR) is $RR = \pi_2/\pi_1$. Because we select the study-specific parameters independently of the sample sizes, the weighted analysis has the same target population relative risk as the unweighted analysis.

We use the notation Low(Step)High as starting at “Low” and increasing to “High” in steps of “Step”. For example 0.02(0.01)0.05 refers to the list 0.02, 0.03, 0.04, 0.05.

Parameter scenarios: $M=5(1)20$ studies in a meta-analysis; $\pi_1 = 0.02(0.02)0.10$; $RR=1.0(0.5)2.5$; $D=0.2(0.2)1.0$; $NL=100$; $NH=600(200)2000$. $\{\pi_2 = RR \pi_1\}$.

For each of the 16 values of M , there are 800 scenarios, each replicated 100,000 times. These simulations lead to very little in the way of sampling error. For estimating a proportion in the neighborhood of 95%, the standard error for the simulated coverage is approximately 0.07%. Table 5 provides the results. The first observation is that the normal approximation is associated with substantial undercoverage of the 95% confidence interval for low numbers of studies, M , being combined, but improves with increasing M . Second, the unweighted coverage, based on the T-distribution with $M-1$ degrees of freedom, is universally accurate, with no simulation having less than 94.9% or greater than 96.2% empirical coverage. Third, the weighted coverage probability (T with $M-2$ df) is not quite as accurate, but acceptable with a range from 93.7% to 96.6% empirical coverage. Lastly, the final column looks at the ratio of the average lengths of the confidence intervals in the log scale of the relative risk. As we might have expected, when both weighted and unweighted methods target the same parameter, the weighted method averages shorter lengths than the unweighted, typically by the order of 10% shorter.

8.2. Description of Simulation #2

In this experiment, all parameters were identical to Simulation #1, except that instead of generating the study individual proportions by uniform distributions, they were generated by Beta distributions keeping identical means and variances of the true within study proportions as in Simulation #1. We did Simulation #2 to see in a limited way if the results are or are not robust against the choice of diversity construction. In fact, the results expressed in Table 6 look very similar to those in Table 5. The t-coverage for the unweighted method had empirical coverage of at least 94.8% (nominal 95%) in every one of the 12,800 [(800 per sample size times 16 different sample sizes $M=5(1)20$] scenarios.

8.3 Description of Simulation #3: (Intentionally designed to have a true unweighted relative risk of 1.0.)

As an illustration where the weighted and unweighted methods estimate different parameters, we experimented as follows. We defined NL to NH as the range of total sample sizes, and the individual studies within the meta-analysis had total sample sizes

independently and uniformly distributed over this range. As in the first two simulations, individual study subjects were allocated by independent balanced coin flips subject to the constraint that between 30% and 70% are allocated to each treatment arm. We define two probability extremes PL and PU, and let the failure probability change linearly from PL to PU on Treatment #1, and from PU to PL on Treatment #2, as the sample size for study j , N_j ranges from NL to NU. The relative risk will clearly be a monotonic function of the actual sample size of the study. Specifically, for given actual sample size for study j , N_j , the true failure rate on Treatment # 1 is:

$$P_{1j} = PL + (PH - PL) \{ (N_j - NL) / (NH - NL) \}$$

and for Treatment #2 is:

$$P_{2j} = PH - (PH - PL) \{ (N_j - NL) / (NH - NL) \}.$$

Once these parameters are set, the actual study specific proportions were generated exactly as in Simulation #1, using the diversity parameter D.

To obtain the true weighted relative risk, WRR, we need to evaluate the expected number of failures on each treatment for a randomly selected study, and take the ratio (Treatment #2:Treatment #1). These quantities are

$E_{ij} = 0.5 * E[N_j E(P_{ij} | N_j)]$ where N_j =study size, P_{ij} is the failure rate on Treatment #i, $i=1,2$.

By design of our simulation, $E(P_{1j} | N_j) = \alpha_1 + \beta_1 N_j$,

where $\alpha_1 = PL - [(PH - PL)NL / (NH - NL)]$ and $\beta_1 = [(PH - PL) / (NH - NL)]$.

$$E(P_{2j} | N_j) = \alpha_2 + \beta_2 N_j,$$

where $\alpha_2 = PH - [(PH - PL)NL / (NH - NL)]$ and $\beta_2 = -\beta_1$

$$E_{ij} = 0.5 * E[\alpha_i N_j + \beta_i N_j^2] = 0.5 [\alpha_i E(N_j) + \beta_i E(N_j^2)]$$

From the discrete uniform distribution over NL to NH

$$E(N_j) = .5(NL + NH)$$

$$E(N_j^2) = [6(NH - NL + 1)]^{-1} [NH(NH + 1)(2NH + 1) - NL(NL - 1)(2NL - 1)]$$

The weighted relative risk is $WRR = E_{2j} / E_{1j}$.

Note that if and only if $PL = PH$, then the weighted and unweighted relative risks will be identical with value 1.0. Otherwise the weighted relative risk is below the unweighted RR of 1.0 (exceeds the unweighted RR of 1.0) if $PH > PL$ ($PH < PL$) respectively.

Parameters: For the simulation, we looked at $M=5(1)20$ number of studies in the meta-analysis, $PL=0.01(0.01)0.05$, $PH=PL+0.02(0.01)0.10$, $NL=100$, $NH=600(200)2000$, and $D=0.2(0.2)1.0$. This yielded 880 scenarios for each of the 16 values of M (or 14,080 simulations) of 100,000 replications each.

As can be seen in Table 7, the coverage of the 95% confidence intervals for their intended parameters is accurate via the t-distribution, although slightly less so than in Tables 5 and 6. In fact, for the unweighted analysis, only 7.0% (0.0%) of the scenarios had empirical coverage that was off by more than 0.5% (1.0%) of the nominal value of 95%. For the weighted method, 63% (19%){2.5%}[0.04%] of the scenarios had empirical coverage that was off by more than 0.5% (1.0%){1.5%}[2.0%] of the nominal value of 95%. Again as in Tables 5 and 6, the coverage of the normal approximation is well below the nominal level of 95%. As we might expect, coverage of the opposite parameter is generally low. For low M , the coverage seems quite good, but this is an artifact of the large width of the confidence intervals. It is also a bit higher for the unweighted method's coverage of the weighted parameter than vice versa because of its slightly wider confidence limits. But this table also gives a warning against confusing the weighted parameter for the unweighted parameter and vice versa. For example, if one interprets relative risk as in most published models (namely unweighted), then the value of 1.0 (the true unweighted RR throughout) is excluded from the weighted 95% confidence interval in at least one quarter of the simulations ($>25,000/100,000$) in 31% of the scenarios even using the t-approximation. Similarly, the unweighted analysis 95% confidence interval excludes the true weighted relative risk in more than one quarter of the simulations in 18% of the scenarios. One interesting observation is that there are scenarios where the confidence interval for the opposite parameter seems to be quite accurate. This is because we have some very gentle slopes in the mix of scenarios. When the slopes are small, the parameters are close in value. The extremes in the coverage for the simulations for the true overall Weighted Relative Risk were 0.9234 ($NL=100$, $NH=600$, $PL=0.05$, $PH=0.07$) and 0.6649 ($NL=100$, $NH=2000$, $PL=0.02$, $PH=0.10$). The Unweighted RR is 1.0 for all scenarios in this simulation study.

Combined, over the 39680 total scenarios, the lowest empirical coverage of the 95% confidence interval for the unweighted method (via t-distributions) was 94.1% and the highest was 96.2%. Corresponding extremes for the empirical coverage for the weighted method were 93.0% and 96.6%.

9. DISCUSSION

If one wishes to conduct meta-analysis that involves the minimum level of assumptions, then one should choose Studies at Random over Effects at Random, and employ the methods of Sections 3 and 4. Since they are based upon first and second moments of independent identically distributed vectors, the summary statistics provide minimum variance unbiased estimators for their summary population counterparts. See Shuster (1982) for more details on the method of moments. They can therefore only be improved upon by making further assumptions, or by adding auxiliary information such as a Bayes prior distribution, study level covariates, or using patient level data.

Despite warnings in DerSimonian-Laird (1986), Bradburn et. al. (2007), Shuster et. al. (2007), and Higgins and Green (2011), the DL method remains widely used for meta-analysis of low event rate binomial trials. This paper can therefore serve as a concrete reference for biostatistical associate editors of medical journals that publish low event rate meta-analysis of safety. We have clearly shown that in low event-rate situations, the within study variance estimates that contribute to the weights for DL are volatile random variables. This is exacerbated, as noted by Shuster et. al. (2010), by the additional fact that under SR

(which includes ER as a special case), after a random permutation of the study labels 1,2, ...,M, the weights, even if totally sample size based, must be regarded as random variables. Whatever properties hold before the permutation must also hold after this permutation, and hence it is not a valid counter-argument to state that you would not make such a permutation. If your statistical properties are invalid after the permutation, they are invalid before the permutation, since such a permutation does not change the estimate or standard error. Furthermore, when it comes to random effects meta-analysis of relative risks or odds ratios, all linear weighted methods per equations (3) and (4), whether weights are random or fixed, should be avoided, as the fundamental model (1) is violated. As seen in Section 5, estimation of within study error properties cannot be trusted and this causes serious validity problems for estimating between study variance, as it depends heavily on the within study variances. For risk differences, however, thanks to unbiased within study estimation of means and variances, between study variance estimation is feasible.

For weighted analysis of low event rate random effects meta-analysis, we recommend the use of the Relative Risk over the Odds Ratio. The reason, as seen in Appendix 2, is that this reduces the number of parameters from 14 {4 means, 4 variances, and 6 covariances} to 5 {2 means, 2 variances, and 1 covariance}. On comparing the odds ratio and relative risk estimates, the estimated odds ratio is in fact the estimated relative risk times the ratio of the adjusted number of non-events. For low event rate trials, this ratio will be virtually 1, and the contributing adjusted non-event estimators will be almost perfectly correlated. We used the odds ratio in Sections 6 and 7.1 simply to be consistent with the other analyses. For the studies in those sections, the weighted relative risk point estimates, confidence intervals, and P-values are very similar to their odds ratio counterparts.

An interesting observation can be seen from Sections 6 and 7. Despite the typical model seen in the texts that under Studies at Random, one is trying to estimate an unweighted parameter, the published estimates appear to be much closer to the weighted parameter than to the unweighted parameter.

One important observation centers on the rate of convergence of the asymptotic distributions for weighted vs. unweighted methods. While we cannot mathematically prove superior convergence, the central limit theorem works more rapidly when the risk of outliers is relatively low. The unweighted method uses the bivariate summary proportions, which will all be relatively low, and unlikely to be prone to large outliers. On the other hand, the weighted method uses the adjusted number of failures in its ratio estimate of relative risk. As such, these will have enormous variability, since large trials will tend to have many more adverse events than small ones. This volatility is seen in the rosiglitazone example of sensitivity to removal of the RECORD trial of Home et. al. (2009), especially for cardiac deaths. In responses to two letters to *Statistics in Medicine* by Carpenter et. al.(2008) and Rucker et. al. (2010), Shuster et. al. (2008) and Shuster (2010b) have demonstrated, at least empirically, that unweighted meta-analysis is robust to both (a) exclusion of a small number of typically sized trials and (b) to early termination of a large trial, respectively. However, as a referee pointed out, the weighted method will be advantageous in terms of robustness when a few very small trials are accidentally excluded. The advantage of the weighted approach is that it will tend to have narrower confidence limits, at the cost of slower convergence to a t-distribution as the number of studies being pooled becomes large.

Recommendation for Forest Plots

Given the low event rates, classical forest plots should not be presented. Instead, forest plots should use exact conditional methods for individual study odds ratios (point and interval estimates). The software package StatXact (Cytel Software), <http://www.cytel.com/software/StatXact.aspx> is ideal for this function.

While we make no global recommendations as to which of the weighted or unweighted methods might be more appropriate, we believe the weighted method will be the more popular. Here are a few questions one may pose in deciding. (1) Do you want your target parameter to estimate the same thing irrespective of the study designs? Unweighted methods do this in general, but weighted ones do not. As an example, weighted methods are much more affected by early termination of trials. (2) Do you want your inference to be based at the patient level (use weighted) or at the trial level (use unweighted)? (3) In the absence of all other considerations, which has a higher priority: narrow confidence limit (use weighted) or accurate coverage (use unweighted)?

We next show a situation where unweighted methods will be intuitively superior to weighted. Consider a multi-surgeon trial of an implanted device vs. best medical management. Clearly, the skill of the surgeon will be an important determinant in failure rates. The surgeon therefore plays the role of the study, and we consider them to be a sample of surgeons in a hypothetical population. If we use the SR model, an unweighted inference would target the relative risk in the general population of surgeons, beyond those who participated in the study. In that context, there is no reason to weight surgeons on the basis of how many subjects they accrued to the trial. In fact, the weighted analysis will estimate a moving target, highly dependent on the percentage of patients accrued by each surgeon.

The state of the science for these low event rate studies is that our methods, both weighted and unweighted, work as advertised even in situations where a relatively small number of trials are combined. For risk differences, the methods of Emerson et. al. (1993, 1996) seem to work well, but they need to undergo additional robustness studies to assess the coverage of the confidence intervals in far more scenarios than they covered. Likewise, the methods of Stijnen et. al. (2010) and Rucker et. al. (2009) appear promising for odds ratios and the angle transformations respectively, but need extensive robustness studies before considering these as reliable tools. In the case of the Stijnen et. al. (2010) method, a more stable algorithm is clearly needed to make this otherwise attractive method feasible in general. One attractive new Binomial-Beta method is due to Cai et. al. (2010). For each study, this involves approximating the number of failures on Treatment #2, conditional on the total failures on the study, as binomial with sample size equal to the total number of failures on the study, and with a Beta prior distribution involving the sample sizes for the proportions. This method applies only to a subset of low event trials, requiring all studies to have low events and large sample sizes for the binomial approximation to work well. Of these methods, only ours are directed at SR, as opposed to ER.

While this paper concentrates on the **analysis** of these studies, excellent guidance on how to design a meta-analysis is contained in Altman et. al (2009) and Liberati et. al. (2009), and the Cochrane Handbook [Higgins and Green, 2011)],

One final point is that although our methods were designed to tackle the low event rate random effects binomial problem, these methods are also perfectly valid when event rates are not low or are a mix of lower and higher proportions. However, we present this for low event rate studies, as there are satisfactory alternative solutions to combinations of moderate rate binomial trials. How our methods perform vs. others for higher rate events is a good subject for further research.

A SAS macro for low event rate meta-analysis is available on the “Jon Shuster’s SAS design and analysis programs” button on <http://ags.bwh.harvard.edu/>, the website for Clinical and Translational Science Statisticians.

Recommended Changes to Practice for Low Event-rate Binomial Random Effects Meta-Analysis

- A. Do not use weighted methods of study level odds ratios or relative risks, such as the DerSimonian-Laird (DL) Method. Use fixed effects only if either (i) the studies can be considered true replications of each other or (ii) with caution when the number of studies being combined is too small (e.g. 2–4) to consider the large sample distributional properties for random effects to hold.
- B. If as a manuscript reviewer, DL is used, request an alternate analysis.
- C. Make sure an unweighted metric that parallels your weighted metric or vice versa is reported, even if it is a secondary analysis.

Acknowledgments

This work was partially supported by grant 1UL1TR000064 from the National Center for Translational Sciences, National Institutes of Health. The authors would like to thank Dr. Michael Link, President of the American Society for Clinical Oncology for his discussion of the effect of the finding of apparent excess mortality in cancer on prescriptions for erythropoiesis stimulating agents. Thanks go to the editor, associate editor, and reviewers for their helpful suggestions and comments.

Financial Disclosure: JJS served on a data and safety monitoring committee for Action Pharma, Finland, 2007–2009. To the best of our knowledge, this company is not involved with diabetes drugs or supportive cancer therapy. He has no other relevant disclosures to make. JDG has no financial disclosures to report. Over the past two years, JJS reports the following: serves on the Board of Directors of Amylin Pharmaceuticals, DexCom Inc., and Moerae-Matrix Inc.; has been a consultant for BD Technologies, Cebix, Exsulin, Gilead, Sanofi, and Takeda; receives research grant support from Bayhill Therapeutics, Halozyne, Intuity, and Osiris Therapeutics; has received scientific lecture support from Sanofi; and is currently a shareholder and/or option holder of Amylin Pharmaceuticals, DexCom Inc., Ideal Life, Moerae-Matrix Inc., Opko Health, and Tandem Diabetes.

References

- Abbate A, Biondi-Zoccai GG, Appleton DL, Erne P, Schoenenberger AW, Lipinski MJ, Agostoni P, Sheiban I, Vetrovec GW. Survival and cardiac remodeling benefits in patients undergoing late percutaneous coronary intervention of the infarct-related artery: evidence from a meta-analysis of randomized controlled trials. *J Am Coll Cardiol*. 2008; 51:956–64. [PubMed: 18308165]
- Agresti A, Hartzel J. Strategies for comparing treatments on a binary response with multi-centre data. *Stat Med*. 2000; 19:1115–1139. [PubMed: 10790684]
- Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*. 2009 Jul 21.339:b2700. [PubMed: 19622552]
- Bath PM, Gray LJ. Association between hormone replacement therapy and subsequent stroke: a meta-analysis. *BMJ*. 2005; 330:342–345. [PubMed: 15640250]
- Bavry AA, Kumbhani DJ, Rassi AN, Bhatt DL, Askari AT. Benefit of early invasive therapy in acute coronary syndromes: a meta-analysis of contemporary randomized clinical trials. *J Am Coll Cardiol*. 2006; 48:1319–25. [PubMed: 17010789]
- Bellamy L, Casas JP, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *Lancet*. 2009; 373:1773–9. [PubMed: 19465232]
- Berger JS, Krantz MJ, Kittelson JM, Hiatt WR. Aspirin for the prevention of cardiovascular events in patients with peripheral artery disease: a meta-analysis of randomized trials. *JAMA*. 2009; 301:1909–19. [PubMed: 19436018]
- Bohlius J, Schmidlin K, Brillant C, Schwarzer G, Trelle S, Seidenfeld J, et al. Recombinant human erythropoiesis-stimulating agents and mortality in patients with cancer: a meta-analysis of randomised trials. *Lancet*. 2009; 373(9674):1532–42. [PubMed: 19410717]
- Bonett DG. Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*. 2009; 14:225–238. [PubMed: 19719359]

- Borenstein, M.; Hedges, L.; Higgins, J.; Rothstein, H. *Introduction to Meta-Analysis*. Wiley Publication; Chichester, UK: 2009.
- Borenstein M, Hedges L, Higgins J, Rothstein H. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*. 2010; 1:97–111.
- Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007; 26:53–77. [PubMed: 16596572]
- Brar SS, Leon MB, Stone GW, Mehran R, Moses JW, Brar SK, Dangas G. Use of drug-eluting stents in acute myocardial infarction: a systematic review and meta-analysis. *J Am Coll Cardiol*. 2009 May 5;53:1677–89. [PubMed: 19406344]
- Burr D, Doss H. A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*. 2005; 100:242–251.
- Cai T, Parast L, Ryan L. Meta-analysis for rare events. *Stat Med*. 2010; 29(20):2078–89. [PubMed: 20623822]
- Carpenter J, Rucker G, Schwarzer G. Comments on ‘Fixed vs. random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death’ by J. J. Shuster, L. S. Jones and D. A. Salmon. *Statistics in Medicine*. 2008; 27:3910–3912. [PubMed: 18069734]
- Collet JP, Montalescot G, Le May M, Borentain M, Gershlick A. Percutaneous coronary intervention after fibrinolysis: a multiple meta-analyses approach according to the type of strategy. *J Am Coll Cardiol*. 2006; 48:1326–35. [PubMed: 17010790]
- de Denus S, Sanoski CA, Carlsson J, Opolski G, Spinler SA. Rate vs. rhythm control in patients with atrial fibrillation: a meta-analysis. *Arch Intern Med*. 2005; 165:258–62. [PubMed: 15710787]
- De Ferrari GM, Sanzo A. T-wave alternans in risk stratification of patients with nonischemic dilated cardiomyopathy: can it help to better select candidates for ICD implantation? *Heart Rhythm*. 2009; 6(3 Suppl):S29–35. [PubMed: 19168397]
- De Luca G, Ucci G, Cassetti E, Marino P. Benefits from small molecule administration as compared with abciximab among patients with ST-segment elevation myocardial infarction treated with primary angioplasty: a meta-analysis. *J Am Coll Cardiol*. 2009; 53:1668–73. [PubMed: 19406342]
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986; 7(3):177–88. [PubMed: 3802833]
- Diamond GA, Bax L, Kaul S. Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Ann Intern Med*. 2007; 147:578–581. [PubMed: 17679700]
- Emerson JD, Hoaglin DC, Mosteller F. A comparison of procedures for combining risk differences in sets of 2×2 tables from clinical trials. *Journal of the Italian Statistical Society*. 1993; 2:269–290.
- Emerson JD, Hoaglin DC, Mosteller F. Simple robust procedures for combining risk differences in sets of 2×2 tables. *Statistics in Medicine*. 1996; 15:1465–1488. [PubMed: 8855475]
- Higgins, JPT.; Green, S. *Cochrane handbook for systematic reviews for interventions*. Wiley Publication; Chichester, UK: 2011. Version 5.1.0
- Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J Roy Statistical Society, Series A*. 2009; 172:137–159.
- Home PD, Pocock SJ, Beck-Nielsen H, et al. RECORD Study Team. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. *Lancet*. 2009; 373:2125–2135. [PubMed: 19501900]
- Laird N, Fitzmaurice G, Xiao D. Comments on ‘Empirical vs. natural weighting in random effects meta-analysis. *Statistics in Medicine*. 2010; 29:1266–1267. [PubMed: 20499329]
- Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med*. 2009; 18;151(4):W65–W94.
- Mottillo S, Filion KB, Bélisle P, Joseph L, Gervais A, O’Loughlin J, Paradis G, Pihl R, Pilote L, Rinfret S, Tremblay M, Eisenberg MJ. Behavioural interventions for smoking cessation: a meta-analysis of randomized controlled trials. *Eur Heart J*. 2009; 30:718–30. [PubMed: 19109354]
- Myung SK, McDonnell DD, Kazinets G, Seo HG, Moskowitz JM. Effects of Web- and computer-based smoking cessation programs: meta-analysis of randomized controlled trials. *Arch Intern*

- Med. 2009; 169:929–37. Erratum in: *Arch Intern Med* 2009; 169:1194–1194. [PubMed: 19468084]
- Nissen SE, Wolski K. Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med*. 2007 Jun 14; 356(24):2457–71. [PubMed: 17517853]
- Nissen SE, Wolski K. Rosiglitazone revisited: An updated Meta-analysis for myocardial infarction and cardiovascular mortality. *Archives of Internal Medicine*. 2010; 170 (14):1191–1201. [PubMed: 20656674]
- Piccini JP, Berger JS, O'Connor CM. Amiodarone for the prevention of sudden cardiac death: a meta-analysis of randomized controlled trials. *Eur Heart J*. 2009; 30:1245–53. [PubMed: 19336434]
- Rucker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*. 2009; 28:721–738. [PubMed: 19072749]
- Rucker G, Schwarzer G, Carpenter J, Schumacher M. Natural weighting-Is it Natural? *Statistics in Medicine*. 2010; 29:2963–2965. [PubMed: 21104939]
- Senn S. The many modes of meta. *Drug Information Journal*. 2000; 34:535–549.
- Senn S. Hans van Houwelingen and the art of summing up. *Biometrical Journal*. 2010; 52:85–94. [PubMed: 20140900]
- Serfling, RJ. Approximation theorems in mathematical statistics. John Wiley Publication; New York: 1980. p. 118-125.
- Shuster JJ. Nonparametric Optimality of the Sample Mean and Sample Variance. *American Statistician*. 1982; 36:176–178.
- Shuster JJ. Empirical vs. natural weighting in random effects meta-analysis. *Statistics in Medicine*. 2010a; 29:1259–1265. [PubMed: 19475538]
- Shuster JJ. Reply to Rucker G, Schwarzer G, Carpenter J, Schumacher M. 'Natural weighting-Is it Natural?'. *Statistics in Medicine*. 2010b; 29:2965–2966.
- Shuster, JJ. Review: Cochrane handbook for systematic reviews for interventions, Version 5.1.0, published 3/2011. In: Higgins, Julian PT.; Green, Sally, editors. *Research Synthesis Methods*. Vol. 2. 2011. p. 126-130.
- Shuster JJ, Hatton RC, Hendeles L, Winterstein AG. Reply to Discussion of "Empirical vs natural weighting in random effects meta-analysis". *Statistics in Medicine*. 2010; 29:1272–1281.
- Shuster JJ, Jones LS, Salmon DA. Fixed vs. random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death. *Statistics in Medicine*. 2007; 26:4375–4385. [PubMed: 17768699]
- Shuster JJ, Jones LS, Salmon DA. Rebuttal to Carpenter *et al.* Comments on 'Fixed vs. random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death'. *Statistics in Medicine*. 2008; 27:3912–3914.
- Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*. 2010; 29:3046–67. [PubMed: 20827667]
- Strippoli GF, Navaneethan SD, Johnson DW, Perkovic V, Pellegrini F, Nicolucci A, Craig JC. Effects of statins in patients with chronic kidney disease: meta-analysis and meta-regression of randomised controlled trials. *BMJ*. 2008; 336:645–51. Erratum in: *BMJ* 2009; 339:b29510–b2591. [PubMed: 18299289]
- Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*. 2004; 23:1351–1375. [PubMed: 15116347]
- Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Stat Med*. 1993; 12:2273–2284. [PubMed: 7907813]
- Verdecchia P, Angeli F, Cavallini C, Gattobigio R, Gentile G, Staessen JA, Reboldi G. Blood pressure reduction and renin-angiotensin system inhibition for prevention of congestive heart failure: a meta-analysis. *Eur Heart J*. 2009; 30:679–88. [PubMed: 19168534]

APPENDIX 1: Non-Technical Supplement for Section 2–5 and 8

Section 2: There are three levels of assumptions one can make for low event-rate binomial meta-analysis. (1) Fixed effects assume that the true effect size is exactly the same in every study, and each study provides a separate estimate of this single number. (2) Random effects with Effects at Random (ER) presume each study is independent but allows each study to have its own randomly drawn effect size. However, ER makes the strong assumption that there is absolutely no relationship between a study's effect size and the size of the study. Under ER, provided its model is correct, any unweighted or weighted estimate (provided weights are totally based upon sample size and add up to one) will provide a valid unbiased estimate of the true effect size. By using weights inversely proportional to the variance, we can optimize the estimate. The argument leveled against unweighted methods in this context is strictly about efficiency not validity. (3) Studies at Random (SR): Here, the only assumption is that the studies are independent. We allow any relationship between study effect size and study design (including its size). Here, different weighting systems will estimate systematically different effect size parameters. In other words, comparing different weighting schemes is no longer a question of optimization as they estimate fundamentally different effect sizes. Since there is no diagnostic test that can prove or disprove that there is no relationship between study size and effect size, it is prudent statistical practice to presume SR, in order to avoid making an inference which may be an artifact of our assumptions. Under SR, we must therefore define the parameter we are targeting in advance of doing any analysis. It needs to be pointed out that it is a mathematical fact that under SR, the only weighting system that guarantees unbiased estimation of the model parameter θ in (1) is the unweighted estimator.

In **Section 3**, we target a **study level metric**, which is envisioned as picking a study at random from the population (each study regardless of size having equal chance of being drawn). What is the metric (e.g. relative risk) if a yet to be assigned patient is randomly assigned to Treatment #2 to Treatment #1? In **Section 4**, although we write our model (1), we note that under SR, the target parameter in the model is the unweighted mean (study level mean). We therefore define an alternate target parameter, θ_w , in Section 4, namely a **patient level metric**, which is envisioned as picking a patient at random from the population of past, present, and future patients with each patient having equal chance of being drawn. What is the metric (e.g. relative risk) if a yet to be assigned patient is randomly assigned to Treatment #2 to Treatment #1? Here, the weight of a study is proportional to its total sample size.

In **Section 5**, we take a close look at a specific example which demonstrates amongst other things, the utter breakdown of what is needed for the DerSimonian-Laird (1986) method to work as envisioned for low event-rate binomial studies. We took a specific example to look strictly at what it needs to be true within a study vs. what it actually delivers. When we have equal sample sizes of 100 per group, there are $101 \times 101 = 10,201$ possible combinations of failure numbers on the two treatments. We can easily compute all of the properties presumed by DL against what they really are. No simulation is needed. To be unbiased, DL treats the weights as fixed (non-random), and presumes no correlation between the weights and the estimates. Section 5 demonstrates that the weights are volatile random variables and that there is strong correlation between these weights and the effect size estimates, making DL biased. Formula (4) is valid for the combination of a weighted sum of estimates only when the weights are fixed (non-random, at least to a good approximation). This formula ignores random variation in the weights, and this example clearly shows this is not a correct presumption. Finally, we looked at the quality of the variance estimation for within study errors. We see from this example, it is highly inaccurate, even when no zero event cells occur. This unreliable estimation makes estimation of the between study variance (except for

risk differences), completely unreliable, and such reporting should be avoided for relative risks and odds ratios in the low event scenario. We therefore recommend against reporting the purported between study variance, τ^2 , in low event-rate relative risk and odds ratio meta-analyses.

In **Section 8**, we looked at a large number of random effects scenarios under Studies at Random to see how the methods perform for a small number of studies being combined. There are 800–880 scenarios for each of 16 values of M , the number of studies being combined from 5–20, and our overall interest is to see how the so called approximate 95% confidence intervals really perform when the number of studies being combined is small. We are not especially interested in the individual scenarios, but cover a wide range of diversity (large vs. small random effects), sample sizes, and failure rates. The ranges tell us much about the consistency of coverage, while the means are also very informative even though we did not choose the parameter settings totally at random. For the first two experiments (Tables 5 and 6), we forced the weighted and unweighted methods to estimate the same target parameter. Note that as expected, the weighted estimate performed better in terms of width of the confidence interval. In Table 7, they estimate different parameters, and this gives a user a very good idea of the apples vs. oranges nature of weighted vs. unweighted estimation when indeed they target different parameters.

In summary, this article provides users with valid assumption-free methods (both unweighted and weighted) for low event-rate meta-analysis. It also provides powerful evidence for avoiding the use of the DL method for low event-rate binomial applications. User friendly software is available from the link in the discussion section.

APPENDIX 2: Other Metrics and their Error Properties

Odds Ratio- Unweighted (See Section 3 for Definitions of Proportions)

$$OR = \{\pi_2(1-\pi_1)\} / \{\pi_1(1-\pi_2)\} \quad (A1)$$

Point Estimate:

$$\widehat{OR} = \{\hat{\pi}_2(1-\hat{\pi}_1)\} / \{\hat{\pi}_1(1-\hat{\pi}_2)\} \quad (A2)$$

Based on the Delta method per Serfling (1980), and using natural logs,

$$Z_{\text{LOR}} = \{\text{Log}(\widehat{OR}) - \text{Log}(OR)\} / \text{SE}$$

has an asymptotic t-distribution with $M-1$ degrees of freedom where

$$\text{SE} = \text{SQRT}\{(Q+R-2S)/M\} \quad (A3)$$

where

$$\begin{aligned} Q &= C_{11} / [\hat{\pi}_1 (1 - \hat{\pi}_1)]^2 \\ R &= C_{22} / [\hat{\pi}_2 (1 - \hat{\pi}_2)]^2 \\ S &= C_{12} / [\hat{\pi}_1 (1 - \hat{\pi}_1) \hat{\pi}_2 (1 - \hat{\pi}_2)] \end{aligned}$$

See (10) for the definition of the C_{kl} $k=1,2$, $l=1,2$.

Risk Difference- Unweighted (See Section 3 for Definitions of Proportions)

$$RD = \pi_2 - \pi_1$$

Point Estimate:

$$\begin{aligned} \widehat{RD} &= \hat{\pi}_2 - \hat{\pi}_1 \\ Z_{RD} &= (\widehat{RD} - RD) / SE \end{aligned}$$

has an asymptotic t-distribution with $M-1$ degrees of freedom where SE^2 can be calculated as the sample variance of the study-specific estimated differences in proportions divided by M , or equivalently, as

$$SE = \text{SQRT}\{(Q_2 + R_2 - 2S_2)/M\} \quad (A4)$$

where

$$\begin{aligned} Q_2 &= C_{11} \\ R_2 &= C_{22} \\ S_2 &= C_{12} \end{aligned}$$

See (10) for the definition of the C_{kl} $k=1,2$, $l=1,2$.

Odds Ratio- Weighted

For $i=1,2$ and $j=1,2,\dots,M$ let $A_{ij} = U_j \hat{P}_{ij}$ and $B_{ij} = U_j (1 - \hat{P}_{ij})$ be the adjusted number of events and non-events respectively.

Let us define sample means as follows:

$$\begin{aligned} \bar{A}_i &= \sum_j A_{ij} / M, & \bar{B}_i &= \sum_j B_{ij} / M & \text{and} \\ \bar{U} &= \sum_j U_j / M \end{aligned}$$

Since the estimated proportions, $\hat{P}_{ij}$ and B_{ij} have a common denominator U_j , the odds ratio is estimated as the following ratio:

$$\widehat{OR} = \bar{A}_2 \bar{B}_1 / \{\bar{A}_1 \bar{B}_2\} \quad (A5)$$

This represents a consistent estimator of the odds ratio (the patient level odds ratio analogous to the patient level relative risk defined in Section 4.) Furthermore, using the delta method and 4-dimensional central limit theorem for independent, identically distributed vectors, $\log(\widehat{OR})$ is asymptotically t-distributed (M-2 df) with mean $\log(OR)$ and asymptotic variance

$$\begin{aligned} SE_{LOR}^2 = & \{ \{S(A_{1j})/\bar{A}_1\}^2 + \{ \{S(A_{2j})/\bar{A}_2\}^2 + \{S(B_{1j})/\bar{B}_1\}^2 + \{S(B_{2j})/\bar{B}_2\}^2 \\ & + 2\{C(A_{1j}, B_{2j})/(\bar{A}_1 \bar{B}_2)\} + 2\{C(A_{2j}, B_{1j})/(\bar{A}_2 \bar{B}_1)\} \\ & - 2\{C(A_{1j}, A_{2j})/(\bar{A}_1 \bar{A}_2)\} - 2\{C(B_{1j}, B_{2j})/(\bar{B}_1 \bar{B}_2)\} \\ & - 2\{C(A_{1j}, B_{1j})/(\bar{A}_1 \bar{B}_1)\} - 2\{C(A_{2j}, B_{2j})/(\bar{A}_2 \bar{B}_2)\} \} / M \end{aligned} \quad (A6)$$

where $S()$ represents the sample standard deviation and $C(,)$ represents the sample covariance. The Standard error of $\log(\widehat{OR})$ is $SQRT(SE_{LOR}^2)$.

Risk Difference-Weighted

This will be a ratio estimate, but unfortunately, logs cannot be applied as risk differences can be negative.

Let us define, with the A's as above for the Odds Ratio,

$$\begin{aligned} D_j &= (A_{2j} - A_{1j}) \\ U_j &= (N_{1j} + N_{2j})/2. \end{aligned} \quad (A7)$$

Then if $\bar{D}$ and $\bar{U}$ represent the sample means of the D_j and U_j respectively,

$$\widehat{RD} = \bar{D}/\bar{U} \quad (A8)$$

is consistent for the risk difference. Further, from the central limit theorem and the delta method,

$\widehat{RD}$ is asymptotically t (M-2 df) with mean equal to RD per the weighted analog of the patient level metric defined in Section 4, and asymptotic variance

$$SE_{RD}^2 = \{ \{S^2(D_j)/\bar{U}^2\} + \{ \bar{D}^2 S^2(U_j)/\bar{U}^4 \} - 2\{ \bar{D} C(D_j, U_j)/\bar{U}^3 \} \} / M \quad (A9)$$

where $S()$ is the sample standard deviation and $C(,)$ is the sample covariance.

Table 1

Example of Actual vs. Ideal Properties for Estimation of Study-Specific Log(RR) for a two independent binomial sample trial with Sample Size: N=100 per group. (RR=2)

Mode	True Failure Rates	Bias (A)	Coef_var for Variance est (B)	Ratio (C)	Correlation of Relative Weight (1/est_var) with Estimate (D)
1	2% vs. 1%	-0.10	41%	1.79	-0.13
2		-0.07	41%	1.64	-0.17
3		-0.35	31%	2.17	0.03
1	4% vs. 2%	0.06	56%	1.24	-0.36
2		0.06	55%	1.24	-0.36
3		-0.11	42%	1.50	-0.22
1	10% vs. 5%	0.06	61.8%	0.93	-0.56
2		0.06	61.8%	0.93	-0.56
3		0.05	50.7%	0.98	-0.54
Ideal		0.00	0.00	1.00	0.00

$$\text{Ratio} = E(\text{Est_var}) / \text{MSE}[\log(\widehat{RR})]$$

Ideal=Presumptions for DerSimonian-Laird method [8] on Logits (Comprehensive Meta-Analysis Default) to work. Letters (A,B,C,D) refer to the letters defined in the paragraph above, "Parameters of interest...."

Mode 1: No exclusions; Mode 2: Excludes results with no failures on both treatments; Mode 3: Excludes results with no failures either on one or both treatments.

Table 2

Random Survey of Cardiovascular Literature 2005–2010

Reference	Endpoint	DL Result Published	Unweighted Result Section 3	Weighted Result Section 4
Myung (2009)	RR(M=22)	1.44 (1.27–1.69)	1.53(1.30–1.82)	1.32(1.13–1.54)
deDenus (2005)	OR(M=5)	0.87 (0.74–1.02)	1.01(0.61–1.66)	0.87(0.79–0.97)
Bath (2005)	OR(M=28)	1.29(1.13–1.47)	1.47(0.80–2.69)	1.31(1.18–1.44)
Strippoli (2009)	RR(M=5)	0.81(0.74–0.89)	0.68(0.31–1.53)	0.81(0.79–0.82)
Mottillo (2009)	OR	Bayes	N/A	N/A
Piccini (2009)	OR(M=15)	0.72(0.61–0.84)	0.62(0.38–1.00)	0.71(0.62–0.82)
Verdecchia (2009)	OR(M=8)	0.76(0.55–1.04)	1.00(0.61–1.62)	0.83(0.56–1.24)
DeFerrari (2009)	RR	?? Fixed Effects	N/A	N/A
Abbate (2008)	OR(M=10)	0.49(0.26–0.94)	0.42(0.19–0.91)	0.74 (0.41–1.39)
Bavry (2006)	RR(M=7)	0.75(0.63–0.90)	0.52(0.28–0.96)	0.75(0.63–0.89)
Brar (2009)	RR(M=13)	0.89(0.70–1.14)	0.79(0.52–1.19)	0.90(0.70–1.14)
Collet (2006)	OR(M=4)	0.63(0.39–0.99) ^a	0.44(0.13–1.56)	0.58(0.34–1.00)
Berger (2009)	RR(M=18)	0.88(0.76–1.04)	0.98(0.71–1.35)	0.87(0.75–1.02)
De Luca (2009)	OR(M=6)	1.14(0.64–2.04)	0.87(0.33–2.31)	1.14(0.57–2.31)
Bellamy (2009)	RR(M=20)	7.43(4.79–11.51)	11.2(5.8–21.5)	12.2(11.3–13.3)

^aPublished upper limit was 0.99, but the correct upper limit is 1.01

M=Number of Studies; OR=Odds Ratio; RR=Relative Risk

Ratio of lengths of Confidence intervals in the Log Scale for DerSimonian-Laird estimate (based on 13 results) to:

Unweighted estimate per Section 3. Median=0.49. Quartiles 0.32, 0.65.

Weighted estimate per Section 4. Median=1.03. Quartiles 0.92, 1.32.

Table 3

Rosiglitazone data from Nissen-Wolski {NW} (2010).

Contrast of Results by Method (56 trials vs. 55 trials excluding RECORD)

Method	Est OR	95% CL	P-Value 2-sided	Ratio of CI Lengths Log Scale W/O to With RECORD
Myocardial Infarction				
NW-Peto Method	1.28	1.02–1.63	0.04	131%
Without RECORD	1.39	1.02–1.89	0.04	
Unweighted (Per Sec 3)	1.43	0.94–2.16	0.092	111%
Without RECORD	1.46	0.92–2.30	0.10	
CMA (Logits)	1.22	0.97–1.54	0.092	129%
Without RECORD	1.27	0.94–1.71	0.12	
Weighted (Per Sec 4)	1.28	1.11–1.48	0.0014	122%
Without RECORD	1.38	1.14–1.68	0.0017	
Cardiac Deaths				
NW-Peto Method	1.03	0.78–1.36	0.86	167%
Without RECORD	1.46	0.92–2.33	0.11	
Unweighted (Per Sec 3)	1.49	1.01–2.20	0.043	110%
Without RECORD	1.72	1.12–2.64	0.015	
CMA (Logits)	0.97	0.74–1.28	0.85	161%
Without RECORD	1.23	0.79–1.91	0.37	
Weighted (Per Sec 4)	1.04	0.78–1.38	0.80	131%
Without RECORD	1.51	1.10–2.08	0.012	

CMA=Comprehensive Meta-Analysis for Odds Ratio (Default version)

Table 4

On study Mortality for Erythropoiesis Stimulating Agents in Cancer

Method	Point Est	95% CL	P= Value 2-sided
HR as Published by Bohlius (2009)	1.17	1.06–1.30	0.001
RR: CMA- Binomial	1.15	1.05–1.26	0.003
RR: Unweighted (Per Section 3)	1.05	0.95–1.17	0.33
RR: Weighted (Per Section 4)	1.14	1.04–1.25	0.0054

HR=Hazard Ratio RR=Relative Risk

Table 5

Results of 100,000 Simulations, each with 800 Scenarios for each M, in Simulation #1.

M	Unweighted Coverage (Normal)	Unweighted Coverage (T: M-1)	Weighted Coverage (Normal)	Weighted Coverage (T: M-2)	Ratio Length Wt/Unw(%)
5	88.6(88.0–89.5)	95.6(95.1–96.2)	85.8(84.1–87.4)	95.7(94.6–96.6)	100.9(91.8–108.9)
6	89.9(89.5–90.6)	95.5(95.1–96.1)	87.5(86.1–88.8)	95.0(94.1–95.9)	95.6(86.6–104.0)
7	90.8(90.3–91.4)	95.5(95.1–96.0)	88.7(87.4–89.8)	94.7(93.9–95.6)	93.4(84.2–102.3)
8	91.4(90.9–92.0)	95.4(95.1–96.0)	89.5(88.3–90.6)	94.6(93.8–95.5)	92.3(83.0–101.5)
9	91.8(91.5–92.3)	95.4(95.0–95.8)	90.2(89.2–91.2)	94.5(93.7–95.3)	91.7(82.2–101.3)
10	92.2(91.8–92.6)	95.4(95.0–95.8)	90.7(89.7–91.7)	94.5(93.7–95.3)	91.3(81.7–101.2)
11	92.5(92.2–92.9)	95.3(95.0–95.7)	91.1(90.2–91.9)	94.5(93.7–95.1)	91.0(81.4–101.1)
12	92.7(92.5–93.1)	95.3(95.0–95.7)	91.5(90.6–92.2)	94.5(93.8–95.1)	90.9(81.1–101.2)
13	92.9(92.6–93.3)	95.3(95.0–95.7)	91.8(91.0–92.5)	94.5(93.9–95.1)	90.8(80.9–101.2)
14	93.1(92.8–93.4)	95.3(95.0–95.6)	92.0(91.4–92.7)	94.5(93.9–95.1)	90.7(80.7–101.3)
15	93.2(92.9–93.5)	95.2(95.9–95.5)	92.2(91.6–93.0)	94.5(93.9–95.2)	90.6(80.5–101.4)
16	93.3(93.0–93.6)	95.2(95.0–95.5)	92.4(91.9–93.1)	94.5(94.0–95.1)	90.5(80.4–101.4)
17	93.4(93.2–93.7)	95.2(95.0–95.5)	92.6(92.0–93.1)	94.5(94.0–95.1)	90.5(80.3–101.4)
18	93.5(93.3–93.8)	95.2(95.0–95.5)	92.7(92.2–93.3)	94.5(94.0–95.1)	90.5(80.2–101.6)
19	93.6(93.3–93.9)	95.2(94.9–95.5)	92.8(92.3–93.3)	94.5(94.1–95.0)	90.4(80.1–101.6)
20	93.7(93.4–93.9)	95.2(94.9–95.4)	92.9(92.4–93.4)	94.6(94.1–95.0)	90.4(80.1–101.6)

Entries are Mean (Range) in Percent (%): M=Number of Studies in Meta-Analysis

Table 6

Results of 100,000 Simulations, each with 800 Scenarios for each M, in Simulation # 2.

M	Unweighted Coverage (Normal)	Unweighted Coverage (T: M-1)	Weighted Coverage (Normal)	Weighted Coverage (T: M-2)	Ratio Length Wt/Unwt(%)
5	88.5(87.7–89.4)	95.6(95.0–96.2)	85.7(83.6–87.4)	95.7(94.6–96.5)	100.6(91.7–108.4)
6	89.8(89.0–90.5)	95.5(95.0–96.1)	87.4(85.6–88.8)	95.0(93.9–95.9)	95.4(86.6–103.2)
7	90.7(90.0–91.4)	95.5(94.9–96.0)	88.6(87.0–89.9)	94.7(93.6–95.6)	93.2(84.2–101.4)
8	91.3(90.7–91.9)	95.4(94.9–95.9)	89.5(88.1–90.6)	94.6(93.6–95.4)	92.1(83.0–100.8)
9	91.8(91.3–92.3)	95.4(94.9–95.8)	90.1(88.9–91.1)	94.5(93.6–95.2)	91.5(82.2–100.5)
10	92.1(91.6–92.7)	95.3(94.8–95.8)	90.6(89.5–91.5)	94.5(93.5–95.1)	91.1(81.6–100.4)
11	92.4(91.9–92.9)	95.3(94.9–95.7)	91.1(90.0–91.8)	94.4(93.6–95.1)	90.8(81.3–100.4)
12	92.6(92.2–93.1)	95.3(94.9–95.7)	91.4(90.4–92.2)	94.4(93.7–95.1)	90.7(81.0–100.5)
13	92.8(92.5–93.2)	95.2(94.8–95.7)	91.7(90.7–92.4)	94.4(93.6–95.1)	90.6(80.8–100.5)
14	93.0(92.5–93.4)	95.2(94.8–95.6)	91.9(91.1–92.6)	94.4(93.7–95.0)	90.5(80.6–100.7)
15	93.2(92.8–93.5)	95.2(94.9–95.6)	92.2(91.3–92.8)	94.5(93.8–95.0)	90.4(80.5–100.8)
16	93.3(92.9–93.6)	95.2(94.8–95.6)	92.4(91.6–93.0)	94.5(93.8–95.0)	90.4(80.4–100.9)
17	93.4(93.0–93.7)	95.2(94.8–95.5)	92.5(91.8–93.0)	94.5(93.9–94.9)	90.3(80.3–100.9)
18	93.5(93.1–93.8)	95.2(94.8–95.5)	92.7(92.0–93.2)	94.5(93.9–95.0)	90.3(80.2–101.0)
19	93.6(93.2–93.9)	95.2(94.8–95.5)	92.8(92.2–93.3)	94.5(93.9–95.0)	90.3(80.1–101.1)
20	93.6(93.3–94.0)	95.1(94.8–95.5)	92.9(92.3–93.4)	94.5(94.0–95.0)	90.3(80.0–101.2)

These have proportions distributed as Beta-Distributions with same Mean(P) and Var(P) as Table 5

Entries are Mean (Range) in Percent (%): M=Number of Studies in Meta-Analysis

Table 7

Results of 100,000 Simulations, each with 880 Scenarios for each M, in Simulation #3.

M	Unweighted Coverage (Normal)	Unweighted Coverage (T: M-1)	Unweighted Coverage of Weighted Param	Weighted Coverage (Normal)	Weighted Coverage (T: M-2)	Weighted Coverage of Unweighted Param
5	87.5(86.9–88.4)	94.7(94.1–95.3)	93.1 (89.7–95.4)	85.0(82.9–87.1)	95.0(93.6–96.3)	93.1(90.0–95.6)
6	88.9(88.4–89.7)	94.7(94.2–95.2)	92.1(87.6–95.2)	86.9(85.2–88.6)	94.4(93.2–95.7)	90.7(85.5–94.9)
7	89.9(89.4–90.6)	94.7(94.1–95.2)	91.0(84.9–94.9)	88.2(86.7–89.6)	94.2(93.0–95.4)	88.6(81.1–94.4)
8	90.6(90.1–91.3)	94.7(94.2–95.2)	89.8(81.9–94.7)	89.2(87.8–90.4)	94.1(92.9–95.2)	86.6(77.0–94.1)
9	91.2(90.6–91.9)	94.7(94.3–95.2)	88.5(78.9–94.5)	89.9(88.6–91.0)	94.2(93.0–95.1)	84.7(73.3–93.8)
10	91.6(91.1–92.2)	94.7(94.2–95.2)	87.1(75.7–94.3)	90.5(89.3–91.5)	94.2(93.2–95.0)	82.9(69.3–93.5)
11	91.9(91.4–92.5)	94.8(94.3–95.2)	85.8(72.8–94.2)	91.0(89.8–91.8)	94.2(93.2–95.0)	81.2(65.6–93.3)
12	92.2(91.6–92.7)	94.7(94.3–95.2)	84.3(69.4–94.0)	91.3(90.4–92.2)	94.3(93.4–95.0)	79.5(62.1–93.0)
13	92.4(91.9–92.9)	94.8(94.3–95.2)	82.9(66.4–93.8)	91.6(90.7–92.4)	94.3(93.4–95.0)	77.8(58.9–92.9)
14	92.6(92.0–93.1)	94.8(94.2–95.2)	81.5(63.4–93.7)	91.9(91.1–92.6)	94.3(93.7–95.0)	76.2(55.7–92.7)
15	92.8(92.3–93.3)	94.8(94.3–95.2)	80.1(60.3–93.5)	92.1(91.4–92.9)	94.4(93.7–95.0)	74.6(52.5–92.7)
16	92.9(92.4–93.4)	94.8(94.3–95.2)	78.6(57.5–93.3)	92.3(91.6–92.9)	94.4(93.7–94.9)	73.0(49.9–92.4)
17	93.1(92.6–93.5)	94.8(94.4–95.2)	77.2(54.7–93.1)	92.5(91.8–93.2)	94.5(93.8–95.1)	71.4(46.7–92.2)
18	93.2(92.7–93.6)	94.8(94.4–95.2)	75.8(51.8–92.8)	92.7(92.1–93.3)	94.5(93.9–95.0)	69.9(44.1–92.1)
19	93.3(92.8–93.7)	94.8(94.4–95.3)	74.4(49.4–92.6)	92.8(92.1–93.4)	94.5(93.9–95.0)	68.4(41.3–91.9)
20	93.4(92.9–93.7)	94.8(94.4–95.2)	73.0(46.8–92.4)	92.9(92.3–93.5)	94.5(94.0–95.1)	66.9(38.9–91.7)

Entries are Mean (Range) in Percent (%): M=Number of Studies in Meta-Analysis