Tutorial in Biostatistics

Statistics in Medicine

Received: 22 December 2014, Accepted: 19 November 2015 Published online 5 January 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6844

# Low-event-rate meta-analyses of clinical trials: implementing good practices

**Jonathan J. Shuster[a]*† and Michael A. Walker[b]**

**Meta-analysis of clinical trials is a methodology to summarize information from a collection of trials about an intervention, in order to make informed inferences about that intervention. Random effects allow the target population outcomes to vary among trials. Since meta-analysis is often an important element in helping shape public health policy, society depends on biostatisticians to help ensure that the methodology is sound. Yet when meta-analysis involves randomized binomial trials with low event rates, the overwhelming majority of publications use methods currently not intended for such data. This statistical practice issue must be addressed. Proper methods exist, but they are rarely applied. This tutorial is devoted to estimating a well-defined overall relative risk, via a patient-weighted random-effects method. We show what goes wrong with methods based on 'inverse-variance' weights, which are almost universally used. To illustrate similarities and differences, we contrast our methods, inverse-variance methods, and the published results (usually inverse-variance) for 18 meta-analyses from 13 *Journal of the American Medical Association* articles. We also consider the 2007 case of rosiglitazone (Avandia), where important public health issues were at stake, involving patient cardiovascular risk. The most widely used method would have reached a different conclusion. © 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.**

**Keywords:** clinical trial; low event rates; meta-analysis; random effects; relative risk

## 1. Introduction

Meta-analysis is often used to assist policymakers assemble information on important health policy issues. We recognize that because of selection bias, reporting bias, and the likelihood of errors in the data from contributing studies, it is imperfect as a scientific method. But when meta-analysis is conducted, its methods must be statistically rigorous. The primary purposes of this tutorial are to (i) make potential analysts and journal reviewers aware that the overwhelming majority of reports of random-effects meta-analysis of low-event-rate clinical trials are using inverse-variance methods that are not appropriate for this situation; (ii) present parameterizations of relative risk, the most popular metric for meta-analysis of binomial data, and argue for a survey-sampling approach; (iii) present in detail the method of Shuster, Guo and Skylar (SGS) [1] as one possible remedy; and (iv) present a comparative analysis of 18 meta-analyses from 13 *Journal of the American Medical Association* (JAMA) articles as published, using the method of DerSimonian and Laird (DL) [2] and using SGS [1]. The scope of this article is on good practices in the estimation of the overall relative risk for low-event-rate random-effects meta-analysis of randomized binomial trials. Issues related to how to properly conduct other aspects of a meta-analysis are beyond the scope of this tutorial.

In random-effects meta-analysis, the method most commonly used for summarizing relative risk for independent two-sample binomial trials, DL [2], has serious theoretical deficiencies when the event rates are low. As of 08/04/2015, according to the Web-of-Science, this is the most-cited paper on meta-analysis (nearly 13,000). Yet some of the most important clinical trials related applications of meta-analysis are precisely in this arena, as when event rates are low, it takes large numbers of patients and large numbers of trials to accurately assess the safety and efficacy of interventions. In their final paragraph, DL [2] mildly cautioned users about problems in estimating variances when sample sizes are small. Section 16.9.5 of

[a] *Department of Health Outcomes and Policy, College of Medicine, University of Florida, Gainesville, FL, 32610-0177, U.S.A.*
[b] *Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, IA, 52242, U.S.A.*
**Correspondence to: Jonathan J Shuster, Department of Health Outcomes and Policy, College of Medicine, University of Florida, PO Box 100177, Gainesville FL, 32610-0177 U.S.A.*
† *E-mail: shusterj@ufl.edu*

the Cochrane Handbook [3] expressly states that 'Methods that should be avoided with rare events are the inverse-variance methods (including the DerSimonian and Laird random-effects method)'. Further, the Cochrane Handbook also states that as of 2011, 'The DerSimonian and Laird method is the only random-effects method commonly available in meta-analytic software'. [These statements also appear in Section 16.9.5 of the 2008 version.] This leaves applied researchers with a serious gap between computational capability and sound biostatistical theory. In their Section 5, SGS [1] present mechanistic reasons that there is potential for major differences in accuracy within studies between the large-sample estimates and the actual parameters they are trying to estimate. The issues center on rare events, even when no arms have zero events. Hence, the theoretical problems are not resolved by continuity corrections (perhaps more appropriately termed bias adjustments) in zero-event arms of trials.

The major issue with inverse-variance methods in low-event-rate situations is that the variance estimate for an individual-study-level log of the relative risk is associated with the direction of the sampling error, inducing bias. The estimate of within-study asymptotic variance when, for both groups, the observed number of events is not zero is

$$\hat{v}_j = \left[N_{1j}\hat{P}_{1j}/\left(1-\hat{P}_{1j}\right)\right]^{-1} + \left[N_{2j}\hat{P}_{2j}/\left(1-\hat{P}_{2j}\right)\right]^{-1} \quad (1)$$

where the $N_{ij}$ and $\hat{P}_{ij}$ are the sample sizes and event proportions for study j, treatment i.

When the sampling error for an event proportion is in the positive (negative) direction, the impact is to increase (decrease) the weights, respectively. For large samples without rare events, this is a minor consideration. But it is a major problem for low-event-rate situations, even when no zero-event arms occur.

The common practice of assessing heterogeneity using Cochran's Q statistic, and using the result to decide between fixed and random effects, is generally not acceptable. Borenstein *et al.* [4], page 84 entitles a section: 'Model should not be based on the test for heterogeneity'. In other words, the choice should be made according to the nature of the trials being combined, and not on empirical evidence supporting or rejecting homogeneity. Given the exceedingly low sensitivity of the Cochran's Q statistic when event rates are low, the only plausible conclusions are that (i) homogeneity is implausible and (ii) homogeneity is inconclusive. In either case, we do not have much confidence in homogeneity. Since random effects are valid whether or not fixed effects are valid, it is prudent to use random effects, unless the trials being combined are truly conducted under universal conditions, something that will occur only rarely.

To fill the methodological gap, Section 4 of SGS [1] presents a patient-weighted alternative random-effects method that they vetted in nearly 40,000 rare-event meta-analysis scenarios where the number of studies being combined is small, 5–20. The large-sample theory applies to large numbers of studies being combined, so when the number of studies is small, the authors had concern about the accuracy of their normal distribution and *t*-distribution approximations. The normal distribution approximations fared poorly, but their *t*-distribution approximations were much more accurate. For these, the real coverage of the 95% confidence intervals (CIs) averaged nearly 95%, with only modest departures from 95% in the individual scenarios. To help users conduct the analyses using these methods, they offer a SAS (Statistical Analysis System) macro at http://actstat.org/associated-links.html.

We chose to concentrate our review of published meta-analyses on the JAMA because it publishes a large number of highly cited meta-analyses of low-event-rate clinical trials. Our purpose is not to second guess individual articles but rather to see how the published papers' results line up with the methods of SGS. The review aims to answer two questions. (i) Do the published results differ from SGS? (ii) Do DL and SGS produce substantially different results for these studies? Specifically, do the analyses reach the same conclusions? Do the methods differ systematically on effect size estimates and lengths of CIs? An excellent motivating example for the clinical importance of this investigation is the Nissen and Wolski [5] meta-analysis of myocardial infarction in randomized trials of rosiglitazone (Avandia) in type 2 diabetes. This will be presented in the Discussion section.

It seems that despite the warnings from [3], analytic practice has not changed. Using the Web-of-Science, we looked at the three most-cited 2014 low-event-rate papers with keywords 'clinical trial' and 'meta-analysis' as of August 14, 2014: Kishimoto *et al.* [6], Williams *et al.* [7] and Monami *et al.* [8]. All used DL [2].

## 2. Parameterization of relative risk

In this section, we look at two approaches to creating a target population parameter: (i) effects at random and (ii) studies at random. We also review inverse-variance approaches and briefly summarize the major application issues raised in Section 5 of SGS [1]. Finally, we present the large-sample distribution theory



 

for the summary estimate for the patient-weighted approach of SGS [1]. In effects at random, we presume conceptually that each study design in the universe is a fixed entity and the effect size is drawn randomly from a single urn of effect sizes, independent of the study design. For example, in effects at random, there is no correlation between study size and the study's true effect. In studies at random, we presume conceptually that the studies form a random sample from a universe of studies, allowing the study-specific effect sizes to be associated with the design. But if we make the additional assumption that the effect size is independent of the design, studies at random and effects at random will coincide. Hence, effects at random is a special case of studies at random.

### 2.1. *Effects at random*

In random-effects meta-analysis, the usual model is

$$\widehat{\theta}_j \;=\; \Theta_j \;+\varepsilon_j \tag{2}$$

where j is the index for the j-th study, j = 1, 2, 3, …, M, with studies considered as independent, and the (possibly vector-valued) estimate $\widehat{\theta}_j$ has conditional mean $\ominus_j$, given the study, making the random error term $\varepsilon_j$ satisfy $E(\varepsilon_j)=0$. That is, $\widehat{\theta}_j$, the study estimate of the study-specific parameter, is unbiased for its population counterpart $\ominus_j$, given the selected study. Physically, we think of the study-specific set $\{\ominus_j\}$ that comprise the meta-analysis as a random sample from a population whose mean is $\ominus = E(\ominus_j)$.

The statistical task is to estimate $\ominus$, or functions of components of $\ominus$, if vector-valued.

SGS called this sampling model 'Effects at random'. It has a very attractive feature in that all weighted combinations of the $\widehat{\theta}_j$, where the weights are fixed (non-random) and sum to 1, are unbiased for $\ominus$, and thus, it makes sense to optimize the weights.

There is an inherent assumption that because the $\{\ominus_j\}$ are presumed to come from a single population, there can be no association between the study design parameters, including sample size, and the particular $\ominus_j$ for the study.

Consider a weighted estimate of $\ominus$, with non-random weights $W_j$:

$$\widehat{\theta}_W \;=\; \Sigma W_j, \hat{\theta}_j \tag{3}$$

where the sum of the weights $= \Sigma W_j = 1$.

With effects at random, model (2) ensures that

$$E\left(\widehat{\theta}_W\right) \;=\; \Theta \text{ and } Var\left(\widehat{\theta}_W\right) \;=\; \Sigma {W_j}^2 {\sigma_j}^2 \tag{4}$$

with $\sigma_j^2$ the unconditional variance of the study estimator $\widehat{\theta}_j$ in equation (2).

For estimating the log of the relative risk (or the log of the odds ratio), DL used weights inversely proportional to the variance of $\widehat{\theta}_j$, namely

$$W_j \;=\; {\sigma_j}^{-2} / \,\Sigma\, {\sigma_k}^{-2} \tag{5}$$

This choice would minimize $Var(\widehat{\theta}_W)$ if these variances were known constants (at least to a very high degree of certainty), but in practice, they are unknown. Since these variances involve both between-study and within-study variance components, they must be estimated. Accuracy and bias are major problems for combining low-event-rate studies as the weights become random variables, subject to bias and sampling error. The DL approach in general ignores the systematic and sampling errors in deriving the weights, leading to validity issues. For further information on this issue, see Böhning *et al.* [9] and Hamza *et al.* [10].

When event rates are low, this approach has three obvious issues, as well as a critical but subtle issue that should make us look for alternative approaches to analyze these collections of studies. These issues are illustrated through an example in Section 5 of SGS [1].

Issue 1: Whether or not there are zero-event cells, the individual logs of relative risk estimates, $\widehat{\theta}_j$, have substantial bias in estimating $\ominus_j$.

Issue 2: Whether or not there are zero-event cells, the variance estimates for within-study variance are inaccurate. {See equation (1)}

The two issues above also compromise estimation of between-study variance as well as true heterogeneity.

Issue 3: When event rates are low, for inverse estimated variance-related weights, the contribution of a single arm of a single study to the weight is approximately proportional to the event probability for that arm [see equation (1)], leading to a strong association between the weights $W_j$ and the estimates $\widehat{\theta}_j$.

The randomness of the weights due to the within-study properties is not considered in the inverse-variance weighting formulation of the effects at random meta-analysis. However, another connected issue should make a user reluctant to apply these methods.

Issue 4: A challenge to the effects at random concept. In actuality, the weights should be seen as random unless they are fixed as $W_j = 1/M$. Without loss of generality, we can randomly permute the indices $j = 1, 2, \ldots, M$ in equation (3), as after this permutation the estimate in equation (3) is unchanged. For ease of notation, we continue to label the studies 1, 2, …, M rather than (1), (2), …, (M) after the permutation. This permutation tool is an enabling concept that allows us to employ powerful techniques borrowed from clustering methods in survey sampling. After this random permutation, each study has a 1/M chance of occupying each index 1, 2, …, M. Now there is no controversy as to whether the weights are random variables. Further, this permutation makes the vectors $(W_j, \widehat{\theta}_j)$ exchangeable over j and therefore identically distributed. From this exchangeability, equations (2) and (3) and the fact that the weights $W_j$ sum to 1, it follows (for all j) that

$$E\left(\widehat{\theta}_W\right) = ME\left(W_j\widehat{\theta}_j\right),\ E\left(\widehat{\theta}_j\right) = \Theta \text{ and } E\left(W_j\right) = 1/M \tag{6}$$

Note that equation (6) is valid as long as the studies can be viewed as a random sample from a universe of studies, a more general situation than effects at random, which as noted previously is a special case. We shall work under this more general set-up in the following.

Using equation (6), the bias in $\widehat{\theta}_W$ can be expressed as

$$B = E\left(\widehat{\theta}_W\right) - \Theta = M\left\{E\left(W_j\widehat{\theta}_j\right) - \Theta(1/M)\right\} = MCov\left(W_j, \widehat{\theta}_j\right)$$

It follows that to avoid bias, the weights must be uncorrelated with the point estimates.

This problem goes well beyond issue 3 stated previously, as violations of the unverifiable 'no correlation' assumption would render effects at random a biased method. This 'no correlation' assumption is also a problem for other meta-analysis settings, including Bayesian approaches.

Since no other weighting system can guarantee unbiasedness in all circumstances, Shuster, Jones and Salmon [11] suggested the use of unweighted methods. Their focus was on literally estimating $\ominus$ rather than seeking out an alternative target parameter. The unweighted method is legitimate and may be the only bias-free method involving equation (3) to estimate $\ominus$, but it is intuitively unappealing to most end-users. Section 4 of SGS [1] used a survey-sampling approach and thereby chose a different target parameter.

We can envision important situations where effects at random may not be a reasonable presumption. For example, early studies of a drug may be smaller and have shorter follow-up than later studies. Further, as side-effect profiles become clearer, eligibility criteria and concomitant medication can differ from earlier (smaller) to later (larger) trials.

### 2.2. *Studies at random: a cluster sampling approach*

Conceptually, we think of studies as being a random sample of potential studies, taken from a large urn of studies. Our inference will be aimed at the totality of studies in the urn. The inference we will make will be to the totality of conceptual patients in the studies in the urn, treating the actual sampled studies as completed. The robustness of this concept lies in the fact that after a random permutation of the study indices 1, 2, …, M, the vectors of parameters (including design information and outcomes) are identically distributed across studies. As we shall see, total-sample-size weighting is a very simple approach, with readily evaluable statistical properties. One can also view the study selection as casting a net into the large urn of potential studies and drawing a sample of M studies from the urn without labeling them.

A key difference between other methods and SGS [1] is that their recommended methods estimate individual proportions and do not rely on individual-study relative-risk estimation, which as noted previously, is biased and has difficulty estimating variances when event rates are low. Even for small samples, proportions can be estimated without bias. We estimate a global event proportion for each treatment and



 

estimate the relative risk by the ratio of these proportions. We use weights proportional to the total sample size for the study. Using arm-specific weighting could create bias if, for example, there was an unbalanced randomization (say, 3 : 1) in a study where the overall event rate was high on both treatments. Studies with one or both arms having zero events are included without continuity corrections.

One easily understood physical definition of relative risk follows naturally from the studies at random concept. Step 1: Draw an unassigned patient at random from the universe of trials, with each hypothetical patient having the same chance of being drawn. What is the ratio of the probability of an event given that patient is assigned to Arm 2, to the probability of an event given that patient is assigned to Arm 1? In this hypothetical experiment, the probability that a patient is drawn from a given trial is proportional to the total sample size for that trial, irrespective of the arm-specific sample-size ratio. Specifically, if we denote the true event rate for Arm = i and Study = j as $P_{ij}$, and the total sample size for study j as $N_j$, then the true overall probability of an event for the randomly selected patient, given assignment to Arm i = 1 or Arm i = 2 is

$$\Pi_i = \Sigma N_j P_{ij} / \Sigma N_j \tag{7}$$

where summation is over the universe of studies.

The true relative risk for this experiment is therefore

$$RR = \Pi_2 / \Pi_1 = \Sigma N_j P_{2j} / \Sigma N_j P_{1j} \tag{8}$$

Equation (8) gives us another intuitive interpretation of this relative risk. The numerator (denominator) is the hypothetical expected number of events in the universe of trials if all patients received Arm 2 (Arm 1). $RR = 2$ would imply that we would expect twice as many events on Arm 2 had all patients in the universe been uniformly treated on Arm 2, rather than if all patients in the universe received Arm 1.

Next, for our actual experiment, we are drawing a random sample of studies from the target universe of studies.

For treatment i = 1, 2 and study j = 1, 2, …, M, let

$$A_{ij} = N_j \hat{P}_{ij} \tag{9}$$

be the predicted number of events on study $j$ if all patients received treatment $i$, where $N_j$ is the total sample size for study $j$, and $\widehat{P}_{ij}$ represents the sample proportion of events for treatment i, study j. Since the proportions are conditionally unbiased, based on the studies at random concept:

$$E(A_{ij}) = E[N_j \hat{P}_{ij}] = E[E\{N_j \hat{P}_{ij} \mid \text{Study} = j\}] = E[N_j P_{ij}] \tag{10}$$

with the unconditional expectation taken over the universe of studies, from which the actual studies are a conceptual random sample. The sample proportions given the study ID are unbiased for the true underlying proportion for that study.

We define the sample means of the exchangeable $A_{ij}$ as follows for the actual studies in the analysis:

$$\overline{A}_i = \sum_j A_{ij} / M$$

Since $\bar{A}_i$ is the sample mean of the exchangeable $A_{ij}$, j = 1, 2, …, M, it follows from equation (10) that

$$E(\overline{A}_i) = E[N_j P_{ij}] \tag{11}$$

If we divide the numerator and denominator in equation (8) by $N_S$, the number of studies in the universe, making both the transformed numerator and denominator population means for the projected number of events when all subjects in the study would receive treatment 2 (numerator) or treatment 1 (denominator), it follows that

$$RR = E[N_j P_{2j}] / E[N_j P_{1j}] \tag{12}$$

and hence $RR$ can be estimated simply by

$$\widehat{RR} = \overline{A}_2 / \overline{A}_1 \tag{13}$$

The $\bar{A}_i$ are unbiased for the numerator (i = 2) and denominator (i = 1) for the true relative risk, defined in equation (12). Moreover, from the method of moments, see Shuster [12], they are nonparametrically minimum variance for the numerator and denominator among all unbiased competitors.

*2.3. Summary notes on effects at random versus studies at random*

(A) If effects at random holds, then studies at random also holds, but not the converse.
(B) When event rates are low, the estimation of the logarithm of a summary relative risk from the individual studies' logarithms of relative risks for effects at random involves biased estimates and poor large-sample approximation of weights and variances.
(C) For effects at random, the target transformation is a log of the relative risk, not a directly estimated relative risk. The mean of a function can differ from the function of the mean, especially when event rates are low. The studies at random approach directly estimates a well-defined relative risk.
(D) Using studies at random, both the random-effects concept and the target relative risk are easier for lay individuals to grasp than they are for effects at random. No model equation is needed in studies at random.

*2.4. Obtaining p*-values, point and interval estimates using studies at random

In this subsection, we provide the asymptotic sampling properties of $\log(\widehat{RR})$, defined in equation (13), obtained by the delta method in SGS [1], Section 4 for M, a 'large number' of studies in the analysis.

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

$$\mathrm{SE}^2 = \left[\left\{\mathrm{S}(A_{1j})/\overline{A}_1\right\}^2 + \left\{\mathrm{S}(A_{2j})/\overline{A}_2\right\}^2 - 2\left\{\mathrm{C}(A_{1j}, A_{2j})/\left(\overline{A}_1\,\overline{A}_2\right)\right\}\right]/\mathrm{M} \tag{14}$$

where $S(\,)$ represents the sample standard deviation and $C(\,,\,)$ represents the sample covariance, denominators M − 1. The standard error of $\log(\widehat{RR})$ is SE = SQRT(SE$^2$).

By asymptotic t, we mean that $[\log(\widehat{RR}) - \log(\mathrm{RR})]/[\mathrm{SE}]$ is approximately central t-distributed with M − 2 degrees of freedom for large M. This is asymptotically equivalent to asymptotic normality, but empirically it gives much more accurate approximations than those based on normality. For small M (5–20), SGS [1], Section 6, have vetted the methods in nearly 40,000 scenarios, with 100,000 simulations for each, with good accuracy. This forms the basis for obtaining p-values and, after taking antilogs, CIs for RR. Specifically, the endpoints of the $100(1-\alpha)$ CI for RR are

$$\exp\left\{\log\left(\widehat{RR}\right) \pm \mathrm{TINV}(\mathrm{M}-2,\ \alpha/2)\mathrm{SE}\right\} \tag{15}$$

with TINV($n$, $\gamma$) defined as the upper $100\gamma$ percentile of the central t-distribution with n degrees of freedom.

$$\text{P-value} = 2^{*}\mathrm{PROBT}\left(-\left|\log\left(\widehat{RR}\right)\right|/\mathrm{SE},\ \mathrm{M}-2\right) \tag{16}$$

with PROBT(t, n) defined as the probability that an observation from a central t-distribution with n degrees of freedom falls below *t*.

## 3. Review of 13 highly cited JAMA articles

In this section, we assess the potential impact of the use of inverse-variance methods for low-event-rate meta-analysis of clinical trials published in the JAMA. This journal was selected because at the time of our selection process, it had the second highest impact factor, behind only the *New England Journal of Medicine* (NEJM), and unlike the NEJM, it published a large number of meta-analyses. We found that all of the eligible articles basically ignored the warnings in [3] and [4] about (i) the use of inverse-variance random-effects methods or (ii) testing for heterogeneity and using a fixed-effects method when the test for heterogeneity was not significant. Our primary purpose is to see how the published results, DL [2] and SGS [1] agree or disagree.

*3.1. Eligibility criteria for inclusion of JAMA articles*

Criteria for inclusion: (1) highly cited article published from 2007 to 2013, as searched in the Web-of-Science as of December 2013 [we prioritized selection by times cited in two strata: (i) 2007–2011 and (ii) 2012–2013]; (2) reported on a review of a collection of randomized independent binomial trials; (3) had at least one low-event-rate study with expected events at most 5; (4) used relative risk (*RR*) as its metric; and (5) had fully retrievable numerator and denominator data on events. [One potential article had to be excluded for this reason.]



 

We identified 13 eligible articles [13–25] and conducted analyses on all low-event-rate binomial endpoints in the article, except that no subset analyses were conducted. The total number of meta-analyses we reviewed from JAMA was 18. Table I lists the meta-analyses that qualified for inclusion in our reanalyses, along with the definition of the endpoints studied.

### *3.2. Results of JAMA review*

For each study, the analysis is provided as published and, more importantly, by the DL [2] method and by the SGS [1] method. Comprehensive Meta-Analysis 2.0 was used for DL, with standard continuity corrections (adding 0.5 to all cells for trials with one zero-event arm and excluding trials where both arms had zero events). Some authors did not fully report the method of meta-analysis used in their papers. Most of these meta-analyses used similar analytical methods. The authors who reported the RR methodology for their results used random effects and fixed effects (some using DL and some using Mantel–Haenszel analysis). Thirteen of these 18 meta-analyses apparently used DL, where our DL results agree with the published results to sufficient accuracy. Those JAMA authors who did not report their methods failed (and evidently were not required) to comply with the recommendation of the International Committee of Medical Journal Editors.

Table II displays point estimates and 95% CIs for each eligible analysis, as published, by DL [2], and by SGS [1]. DL and SGS give similar results for most of the point estimates and CIs. We did find five analyses with substantially different results from SGS. The last column provides the ratio of lengths for the CIs. Analyses with major differences between DL and SGS are highlighted.

## 4. Discussion

An example of the strong motivation for the public health importance of using appropriate methods is a 2007 meta-analysis for myocardial infarction in 48 trials of rosiglitazone (Avandia) in type 2 diabetes. The sentinel danger signal was published by Nissen and Wolski [5] (May 2007), and the FDA held a hearing in July 2007, leading to a Black Box Warning and a major reduction in written prescriptions for rosiglitazone. Although the meta-analysis was not the sole basis for this action, it probably would not have occurred so rapidly without it. Yet, on the basis of the software available to these authors (and still widely used today), the ultimate inferences for both Nissen and Wolski [5] (fixed-effects Peto

**Table I.** Low-event-rate meta-analyses published in JAMA between 2007 and 2013.

| Reference #, multiple analyses .1 and .2 | *M* | Endpoint | Lead author |
|---|---|---|---|
| **13** | 27 | Suicide ideation/attempt | Bridge (2007) |
| 14 | 63 | Antibiotic-associated diarrhea | Hempel (2012) |
| 15 | 15 | Risk of low birth weight | Kayentao (2013) |
| 16 | 15 | Venous thromboembolism | Nalluri (2008) |
| 17.1 | 8 | Lung injury | Neto (2012) |
| **17.2** | 9 | Mortality | Neto (2012) |
| 18.1 | 8 | Cardiovascular deaths | Nguyen (2011) |
| 18.2 | 11 | Prostate cancer-specific mortality | Nguyen (2011) |
| 19.1 | 21 | Incident pancreatitis in 21 large statin trials | Preiss (2012) |
| 19.2 | 7 | Incident pancreatitis in 7 large fibrate trials | Preiss (2012) |
| 20 | 16 | Fatal adverse events | Ranpura (2011) |
| 21 | 17 | All-cause mortality | Rizos (2012) |
| 22.1 | 17 | Major cardiovascular events—inhaled anticholinergics | Singh (2008) |
| 22.2 | 5 | Major cardiovascular events—long term Inhaled anticholinergics | Singh (2008) |
| **23.1** | 5 | Major cardiovascular events | Udell (2013) |
| **23.2** | 5 | Cardiovascular mortality | Udell (2013) |
| 24 | 25 | In hospital mortality | Wiener (2008) |
| **25** | 35 | Mortality | Zarychanski (2013) |

Studies where DerSimonian and Laird and Shuster, Guo and Skylar differ substantially are highlighted.
*M* = number of studies in analysis.

**Table II.** Results as published versus DL (1986) versus SGS (2012).

| Ref. from Table I | Method | As published | DL | SGS | Ratio lengths DL:SGS |
|---|---|---|---|---|---|
| **13** | DL (cc) | **1.7 (1.1–2.7){0.017}** | **1.73 (1.11–2.70)** | **2.13 (1.32–3.44){0.003}** | **0.75** |
| 14 | DL (cc) | 0.58 (0.50–0.68){<0.001} | 0.58 (0.49–0.68) | 0.58 (0.49–0.68){<0.001} | 1.00 |
| 15 | DL | 0.80 (0.69–0.94){0.006} | 0.81 (0.69–0.94) | 0.79 (0.68–0.92){0.005} | 1.04 |
| 16 | Fixed | 1.33 (1.13–1.56){<0.001} | 1.35 (1.14–1.58) | 1.36 (1.15–1.61){0.002} | 0.96 |
| 17.1 | Fixed (cc) | 0.33 (0.23–0.47){<0.001} | 0.41 (0.30–0.56) | 0.39 (0.28–0.55){0.001} | 0.96 |
| **17.2** | Fixed (cc) | **0.64 (0.46–0.86){0.007}** | **0.71 (0.55–0.93)** | **0.70 (0.44–1.11){0.11}** | **0.57** |
| 18.1 | Fixed | 0.93 (0.79–1.10){0.41} | 0.94 (0.79–1.10) | 0.94 (0.80–1.10){0.36} | 1.03 |
| 18.2 | DL (cc) | 0.69 (0.56–0.84){<0.001} | 0.69 (0.56–0.84) | 0.72 (0.59–0.88){0.004} | 0.97 |
| 19.1 | DL (cc) | 0.79 (0.65–0.95){0.01} | 0.79 (0.65–0.95) | 0.78 (0.68–0.90){0.001} | 1.36 |
| 19.2 | DL (cc) | 1.39 (1.00–1.95){0.053} | 1.40 (1.00–1.95) | 1.40 (1.00–1.98){0.052} | 0.97 |
| 20 | DL (cc) | 1.33 (0.95–1.86){0.094} | 1.33 (0.95–1.86) | 1.42 (0.99–2.06){0.058} | 0.85 |
| 21 | DL (cc) | 0.96 (0.91–1.02){0.17} | 0.96 (0.91–1.02) | 0.96 (0.91–1.01){0.097} | 1.10 |
| 22.1 | DL (cc) | 1.58 (1.21–2.06){0.001} | 1.57 (1.19–2.06) | 1.60 (1.28–2.01){0.001} | 1.16 |
| 22.2 | Fixed | 1.73 (1.27–2.36){<0.001} | 1.71 (1.26–2.33) | 1.74 (1.31–2.31){0.008} | 1.07 |
| **23.1** | DL | **0.57 (0.39–0.82){0.003}** | **0.57 (0.39–0.82)** | **0.54 (0.32–0.91){0.032}** | **0.73** |
| **23.2** | DL | **0.81 (0.36–1.83){0.61}** | **0.81 (0.36–1.83)** | **0.77 (0.19–3.03){0.58}** | **0.52** |
| 24 | DL | 0.93 (0.85–1.03){0.15} | 0.93 (0.85–1.03) | 0.93 (0.84–1.03){0.15} | 0.95 |
| **25** | DL | **1.07 (1.00–1.14){0.05}** | **1.07 (1.00–1.14)** | **1.07 (1.02–1.12){0.009}** | **1.40** |

Entries in columns 3–5 are point estimate of relative risk (95% CI){two-sided *p*-value}. Studies where DL and SGS differ substantially are highlighted. DL is calculated from Comprehensive Meta-Analysis version 2.0 and also employs standard continuity corrections for zero-event cells. DL, DerSimonian and Laird; SGS, Shuster, Guo and Skylar; cc, continuity corrections for zero-event cells.

method after a preliminary test for heterogeneity) and Diamond and Kaul [26] (both Bayes and DL [2] with standard continuity corrections), reaching conflicting conclusions, were flawed methodologically.

The Nissen and Wolski meta-analysis was published using a summary odds ratio, so we reconstructed the results using relative risk (*RR*) as a metric. However, for low event rates, the distinction is slight, and relative risk (the ratio of event probabilities) is easier to interpret than the odds ratio (ratio of event odds). The FDA decision had considerable impact on averting potential harm to patients, on large ongoing rosiglitazone trials, and on financial losses to the manufacturer (sales and lawsuits). The meta-analytic basis of the decision, which turned out to be correct, can only be attributed to good fortune, in that NW used the Peto fixed-effect method rather than the DL method (default in Comprehensive Meta-Analysis, the program they used). That program forces the user to see the results of DL before the user can select alternative methods. The results are contrasted in Table III, with clear-cut added risk in the SGS [1] analyses, but equivocal CIs in both the DL and Peto analyses. Although we would not exclude studies in a *de novo* analysis, we also present SGS results after eliminating studies with no events on both arms as a parallel to what was published.

In 2010, Nissen and Wolski [27] added eight studies and further follow-up to their original meta-analysis. For all three methods, the 2010 results agree well with the respective 2007 results, and so details are not shown.

One might argue that the Peto and Mantel–Haenszel methods are valid for low-event-rate collections in assessing the signal, that is, testing that the true relative risk is 1.00 for all studies in the universe. This reduces the testing problem to fixed effects under this null hypothesis. However, this simplification has

**Table III.** Nissen–Wolski (2007) analysis and re-analyses.

| Method | Outcome | Point est. | LCL | UCL | Two-sided *p*-value |
|---|---|---|---|---|---|
| Peto | OR | 1.43 | 1.03 | 1.98 | 0.032 |
| DL | OR | 1.29 | 0.94 | 1.76 | 0.12 |
| DL | RR | 1.28 | 0.94 | 1.75 | 0.12 |
| SGS (1) | RR | 1.41 | 1.14 | 1.75 | 0.0026 |
| SGS (2) | RR | 1.41 | 1.13 | 1.76 | 0.0031 |

(1) includes all 48 studies; (2) excludes 10 studies with no event on both arms.
DL, DerSimonian and Laird; SGS, Shuster, Guo and Skylar; RR, relative risk; OR, odds ratio; LCL(UCL)=lower (upper) 95% confidence limit.



 

two issues. First, when random effects are present, these methods do not produce valid point estimates and confidence limits, both of which are exceedingly important. Second, with random effects, there can be a true overall relative risk of 1.00, with some studies having true relative risks above the neutral value of 1.00, counterbalanced by other studies with true relative risks below 1.00. Now both the Peto and Mantel–Haenszel methods' theoretical presumptions are not applicable under this less restrictive null hypothesis and are likely to misstate the precision of their estimates.

The overwhelming majority of clinical investigators are very reluctant to use Bayesian methods in meta-analysis. Biostatisticians and other methodologists should encourage their clinical colleagues as to their merits in appropriate situations. As of 21 November 2014, there were 15.8 million Google, 2.6 million Google Scholar and 13,000 PUBMED hits for the terms Clinical Trial and Meta-Analysis. When we added the term Bayes, the numbers dropped to 370,000 (2.4%), 23,000 (0.9%) and 118 (0.9%), respectively. Therefore, methodologists need to be much more proactive in this arena.

We draw one distinction between this article and SGS [1], in that SGS did not choose a recommended strategy among three metrics and two weighting methods. We recommend (i) the use of patient-weighted over unweighted analysis and (ii) relative risk as the metric of choice. While SGS showed slightly more accurate coverage for unweighted methods, the patient-weighted methods had considerably narrower confidence intervals, and we consider this added precision to be more important. Further, in a non-binomial article by Shuster [28], with discussion from Laird, Fitzmaurice and Ding [29], Waksman [30] and Thompson and Higgins [31], with response by Shuster *et al.* [32], the net message for unweighted methods is that, although valid, they are highly inefficient. There was no criticism of a patient-weighted method of Shuster [28], also presented in that article. As for choosing among the three metrics in SGS [1], their relative risk analysis needs to estimate far fewer parameters (five sample moments) than their odds-ratio analysis (14 sample moments). Absolute differences in proportions are not in common use in meta-analysis of low-event-rate binomial trials, and they overly weight studies with very low event rates.

Of related interest, Hamza and colleagues [10] have proposed an alternative and superior method of estimating variances via likelihood methodology, rather than the traditional methods.

The majority of these recently published JAMA meta-analyses give similar results when analyzed with DL [2] or SGS [1]. This might give us some comfort in that the majority of published low-event-rate meta-analyses using the DL or Peto method are likely to reach similar conclusions to SGS, including fairly similar confidence limits. But we expect a substantial minority will have major issues with the conclusions and CIs for relative risk. A wider review of the most-cited low-event-rate meta-analyses of clinical trials in other publications is therefore essential. Such a review could assess what study properties exist when DL is accurate versus inaccurate. Such an assessment was well beyond the scope of our small JAMA review. For the future, it is critically important to heed the warnings issued by the Cochrane Handbook and avoid the use of DL when event rates are low. New warnings in major software packages would also help. SGS and a Hypergeometric/Normal Bayesian method per Stijnen *et al.* [33] are attractive alternatives.

Other authors who have approached the low-event-rate problem include Tian *et al.* [34] and Lane [35], but in practice DL continues to be predominantly used. Advantages of SGS over other methods for low-event-rate meta-analysis include the following: (a) it targets a more easily understood population parameter; (b) its estimates do not rely on asymptotic properties within studies; (c) it accommodates a more conservative *t*-approximation rather than a normal approximation when the number of studies is small; (d) it is valid in the more general studies at random setting, whereas its competitors all use the more restrictive effects at random model, assuming the effect drawn for a given study is independent of the study design; (e) it has been vetted for combining small numbers of studies in nearly 40,000 low-event-rate scenarios, with 100,000 simulations each; (f) zero events on one or both arms of a study are handled no differently than any other study (in fact, if no events occur on both arms, the same point estimate is obtained with the study included or excluded [not recommended], but these studies have impact on standard errors; for more on zero event arms, see Kuss[36]); and (g) it is more robust when some trials have group sequential designs. The framework of inference is that the actual trials are complete, that they represent a random sample from a large conceptual universe of trials and that the inference is to the actual potential participants in this universe of trials. This immunizes the inference from biases of raw proportions within group sequential trials. Perceived disadvantages include the following: (i) the method does not directly estimate heterogeneity of relative risks (users can still run a test of heterogeneity of odds ratios, preferably an exact one, but SGS works with or without heterogeneity; one can also readily look for heterogeneity in the proportions, but that has very limited utility); (ii)



 

**Table IV.** Neto [17] study data.

| Study | Arm 1 | Arm 2 |
|---|---|---|
| 1 | 2/26 | 1/26 |
| 2 | 3/23 | 2/13 |
| 3 | 27/163 | 69/212 |
| 4 | 13/558 | 15/533 |
| 5 | 24/76 | 23/74 |
| 6 | 3/154 | 1/75 |
| 7 | 1/75 | 2/74 |
| 8 | 0/50 | 1/50 |
| 9 | 1/20 | 1/20 |

Entries are events/sample size.

the inferential framework is to a conceptual population of studies with the actual completed studies considered to be a random sample (but most alternate methods emanate from equation (2) without a true physical population that allows associations between weights and estimates; moreover, the exchangeability after a random permutation allows us to legitimately use the exchangeability in our inference, even if the targeted population is not fully defined for convenience sampling as opposed to random sampling of studies); and (iii) when the number of studies is small, and the sample sizes and/or event rates are highly diverse, the *t*-approximation may not be accurate.

A question posed and answered by a reviewer is as follows: Why does the DL method receive nearly universal use for these low-event-rate binomial meta-analyses, despite the warning in the Cochrane Handbook? Neither of the two main software packages (RevMan 5 or Comprehensive Meta-Analysis 3.0) issue user warnings when studies have low event rates. Further, the Deeks and Higgins 2010 publication on the statistical algorithms in RevMan 5, http://www.researchgate.net/profile/Jonathan_Deeks2/publication/241313811_Standard_statistical_algorithms_in_Cochrane_reviews_Ve_r_s_i_o_n_5/links/54d159b70cf28370d0e07f9f.pdf, does not issue a warning. A recent article by Cornell *et al.* [37] suggests sunsetting the method in these low-event scenarios.

A completely counterintuitive application can be seen in the second Neto [17] analysis in Table II (individual study data shown in Table IV). DL gives a point estimate for relative risk at 0.71, 95% CI 0.55–0.93, $p=0.004$. If we double the data (every numerator and every denominator), one would think the significance would be amplified. Yet the DL point estimate changes to 0.78 and the CI widens by nearly 40% to 0.56–1.09, $p=0.15$. In the actual data, thanks to the Q statistic being less than the degrees of freedom, the fixed and random-effects analysis coincided. When all entries are doubled, a random-effects analysis was mandated as Q more than doubled, thereby making the standard error increase. This anomaly cannot occur with SGS.

The following are good topics for future advances: (i) Since we claim validity, not optimality of SGS [1], it is of interest to see how its precision compares with Bayesian methods (a tutorial on various Bayesian methods would be a good addition to the literature on low-event-rate meta-analysis); (ii) since SGS' validity does not require low event rates, its properties for small numbers of studies should be investigated when event rates are not low; and (iii) it would be of further interest to see the gain in precision for methods that rely on patient-level data over SGS. With mandatory raw data deposits recently implemented for European clinical trials and with ClinicalTrials.gov considering similar requirements, patient-level data should become available in the not too distant future, without worries of selection bias. Further, we recommend that a doctoral level biostatistician or quantitative epidemiologist be part of the research team for conducting any meta-analysis. Finally, when called upon to review a manuscript that presents results of a meta-analysis involving clinical trials with low event rates, make sure that the analytic methods used are appropriate and adequately documented before recommending acceptance. These meta-analyses can play major contributing roles in setting health policy and in multi-million dollar litigation. Using inappropriate statistical methods can cause substantial damage.

## Acknowledgements

This work was partially supported by NIH grant 1UL1TR000064 from the National Center for Advancing Translational Sciences. Special thanks go to Dr Ingram Olkin, Stanford University, for his helpful comments on an



 

early draft of this paper. We are grateful to the reviewers, especially one whose constructive suggestions had a major impact on the content and quality of the manuscript.

## References


1. Shuster JJ, Guo JD, Skyler JS. Meta-analysis of safety for low event-rate binomial trials. *Research Synthesis Methods* 2012; **3**:30–50.
2. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
3. Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews for Interventions, Version 5.1.0*. 2011; Wiley Publication, Chichester, UK.
4. Borenstein M, Hedges L, Higgins J, Rothstein H. *Introduction to Meta-Analysis*. Wiley: Chichester, UK, 2009.
5. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine* 2007; **356**(24):2457–2471.
6. Kishimoto T, Robenzadeh A, Leucht C, Leucht S, Watanabe K, Mimura M, Borenstein M, Kane JM, Correll CU. Long-acting injectable vs oral antipsychotics for relapse prevention in schizophrenia: a meta-analysis of randomized trials. *Schizophrenia Bulletin* 2014; **40**(1):192–213.
7. Williams CJM, Peyrin-Biroulet L, Ford AC. Systematic review with meta-analysis: malignancies with anti-tumour necrosis factor-α therapy in inflammatory bowel disease. *Alimentary Pharmacology and Therapeutics* 2014; **39**:447–458.
8. Monami M, Dicembrini I, Mannucci E. Dipeptidyl peptidase-4 inhibitors and pancreatitis risk: a meta-analysis of randomized clinical trials. *Diabetes, Obesity, and Metabolism* 2014; **16**:48–56.
9. Böhning D, Malzahn U, Dietz E, Schlattmann P. Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostatistics* 2002; **3**:445–457.
10. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology* 2008; **61**:41–51.
11. 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**(24):4375–4385.
12. Shuster JJ. Nonparametric optimality of the sample mean and sample variance. *American Statistician* 1982; **36**:176–178.
13. Bridge JA, Iyengar S, Salary CB, Barbe RP, Birmaher B, Pincus HA, Ren L, Brent DA. Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment: a meta-analysis of randomized controlled trials. *JAMA* 2007; **297**(15):1683–1696. PubMed PMID: 17440145.
14. Hempel S, Newberry SJ, Maher AR, Wang Z, Miles JN, Shanman R, Johnsen B, Shekelle PG. Probiotics for the prevention and treatment of antibiotic-associated diarrhea: a systematic review and meta-analysis. *JAMA* 2012; **307**(18):1959–1969.
15. Kayentao K, Garner P, van Eijk AM, Naidoo I, Roper C, Mulokozi A, MacArthur JR, Luntamo M, Ashorn P, Doumbo OK, ter Kuile FO. Intermittent preventive therapy for malaria during pregnancy using 2 vs 3 or more doses of sulfadoxine-pyrimethamine and risk of low birth weight in Africa: systematic review and meta-analysis. *JAMA* 2013; **309**(6):594–604.
16. Nalluri SR, Chu D, Keresztes R, Zhu X, Wu S. Risk of venous thromboembolism with the angiogenesis inhibitor bevacizumab in cancer patients: a meta-analysis. *JAMA* 2008; **300**(19):2277–2285.
17. Neto AS, Cardosa SO, Manetta JA, Pereira VG, Espósito DC, Pasqualucci Mde O, Damasceno MC, Schultz MJ. Association between use of lung-protective ventilation with lower tidal volumes and clinical outcomes among patients without acute respiratory distress syndrome: a meta-analysis. *JAMA* 2012; **308**(16):1651–1659.
18. Nguyen PL, Je Y, Schutz FA, Hoffman KE, Hu JC, Parekh A, Beckman JA, Choueiri TK. Association of androgen deprivation therapy with cardiovascular death in patients with prostate cancer: a meta-analysis of randomized trials. *JAMA* 2011; **306**(21):2359–2366.
19. Preiss D, Tikkanen MJ, Welsh P, Ford I, Lovato LC, Elam MB, LaRosa JC, DeMicco DA, Colhoun HM, Goldenberg I, Murphy MJ, MacDonald TM, Pedersen TR, Keech AC, Ridker PM, Kjekshus J, Sattar N, McMurray JJ. Lipid-modifying therapies and risk of pancreatitis: a meta-analysis. *JAMA* 2012; **308**(8):804–811.
20. Ranpura V, Hapani S, Wu S. Treatment-related mortality with bevacizumab in cancer patients: a meta-analysis. *JAMA* 2011; **305**(5):487–494 (Corrected *JAMA*.2011; 305 (8): 229-229, not relevant to exercise in this paper).
21. Rizos EC, Ntzani EE, Bika E, Kostapanos MS, Elisaf MS. Association between omega-3 fatty acid supplementation and risk of major cardiovascular disease events: a systematic review and meta-analysis. *JAMA* 2012; **308**(10):1024–1033.
22. Singh S, Loke YK, Furberg CD. Inhaled anticholinergic and risk of major adverse cardiovascular events in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *JAMA* 2008; **300**(12):1439–1450.
23. Udell JA, Zawi R, Bhatt DL, Keshtkar-Jahromi M, Gaughran F, Phrommintikul A, Ciszewski A, Vakili H, Hoffman EB, Farkouh ME, Cannon CP. Association between influenza vaccination and cardiovascular outcomes in high-risk patients: a meta-analysis. *JAMA* 2013; **310**(16):1711–1720.
24. Wiener RS, Wiener DC, Larson RJ. Benefits and risks of tight glucose control in critically ill adults: a meta-analysis. *JAMA* 2008; **300**(8):933–944. (Corrected *JAMA*.2009; 301(9): 936-936, not relevant to exercise in this paper).
25. Zarychanski R, Abou-Setta AM, Turgeon AF, Houston BL, McIntyre L, Marshall JC, Fergusson DA. Association of hydroxyethyl starch administration with mortality and acute kidney injury in critically ill patients requiring volume resuscitation: a systematic review and meta-analysis. *JAMA* 2013; **309**(7):678–688.
26. Diamond GA, Kaul S. Rosiglitazone and cardiovascular risk. *New England Journal of Medicine* 2007; **357**(9):938–939.
27. 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.
28. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine* 2010; **29**(12):1259–1265.
29. Laird N, Fitzmaurice G, Ding X. Comments on 'Empirical vs natural weighting in random effects meta-analysis'. *Statistics in Medicine* 2010; **29**(12):1266–1267.




 

30. Waksman JA. Comments on 'Empirical vs natural weighting in random effects meta-analysis'. *Statistics in Medicine* 2010; **29**(12):1268–1269.
31. Thompson SG, Higgins JPT. Comments on 'Empirical vs natural weighting in random effects meta-analysis'. *Statistics in Medicine* 2010; **29**(12):1270–1271.
32. 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**(12):1272–1281.
33. 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**:3046–3067.
34. Tian L, Cai T, Pfeffer MA, Piankov N, Cremieux PY, Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 x 2 tables with all available data but without artificial continuity correction. *Biostatistics* 2009; **10**(2):275–281.
35. Lane PW. Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research* 2013; **22**(2):117–132.
36. Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Statistics in Medicine* 2015; **34**(7):1097–1116.
37. Cornell JE, Mulrow CD, Localio R, Stack CB, Melbohm AR, Guallar E, Goodman SN. Random-effects meta-analysis of inconsistent effects: a time for change. *Annals of Internal Medicine* 2014; **160**(4):267–270.



 