Canons and Sparrows II*: The Enhanced Bernoulli Exact Method for Determining Statistical Significance and Effect Size in the Meta-Analysis of k 2 x 2 Tables

Lawrence Marc Paul (lmpaul@sciencesupportconsulting.com)
Bell Laboratories

Methodology

Keywords: Meta-analysis, Categorical Analysis, Dichotomous Analysis, Sparse Tables, Mantel-Haenszel, DerSimonian, Exact Solution, Inverse Variance, Convolution, Heterogeneity, Rare Events

Posted Date: January 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-139437/v1

Version of Record: A version of this preprint was published at Emerging Themes in Epidemiology on August 3rd, 2021. See the published version at https://doi.org/10.1186/s12982-021-00101-8.

Canons and Sparrows II*: The Enhanced Bernoulli Exact Method for Determining Statistical Significance and Effect Size in the Meta-Analysis of *k* 2 x 2 Tables

Lawrence M. Paul

Bell Laboratories, Retired

E-mail: lmpaul@sciencesupportconsulting.com

Running head: Non-parametric exact tests for meta-analysis of categorical data

* *"Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow. The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data."* *(R. A. Fisher, 1925)*

ABSTRACT

Background: The use of meta-analysis to aggregate the results of multiple studies has increased dramatically over the last 40 years. For homogeneous meta-analysis, the Mantel-Haenszel technique has typically been utilized. In such meta-analyses, the effect size across the contributing studies of the meta-analysis differ only by statistical error. If homogeneity cannot be assumed or established, the most popular technique developed to date is the inverse-variance DerSimonian & Laird (DL) technique **[1]**. However, both of these techniques are based on large sample, asymptotic assumptions. At best, they are approximations especially when the number of cases observed in any cell of the corresponding contingency tables is small.

Results: This research develops an exact, non-parametric test for evaluating statistical significance and a related method for estimating effect size in the meta-analysis of *k* 2 x 2 tables for any level of heterogeneity as an alternative to the asymptotic techniques. Monte Carlo simulations show that even for large values of heterogeneity, the Enhanced

Bernoulli Technique (EBT) is far superior at maintaining the pre-specified level of Type I Error than the DL technique. A fully tested implementation in the R statistical language is freely available from the author. In addition, a second related exact test for estimating the Effect Size was developed and is also freely available.

Conclusions: This research has developed two exact tests for the meta-analysis of dichotomous, categorical data. The EBT technique was strongly superior to the DL technique in maintaining a pre-specified level of Type I Error even at extremely high levels of heterogeneity. As shown, the DL technique demonstrated many large violations of this level. Given the various biases towards finding statistical significance prevalent in epidemiology today, a strong focus on maintaining a pre-specified level of Type I Error would seem critical.

Keywords: Meta-analysis; Categorical Analysis; Dichotomous Analysis; Sparse Tables; Mantel-Haenszel; DerSimonian; Exact Solution; Inverse Variance, Convolution, Heterogeneity, Rare Events

BACKGROUND

The use of meta-analysis in epidemiological research has been increasing at a very rapid rate. A review of the National Library of Medicine's online database ("Pub Med") shows that in 1977 there was only a single research article with the term "meta-analysis" in its title. This number had increased to 138 in 1991, 966 in 2005 and to 17,205 in 2019 (see Figure 1).

Insert About Here

Fig. 1 Number of articles containing "meta-analysis" in the title by year of publication.

Part of this growth may be due to the widespread availability of powerful personal computer software making meta-analysis techniques more feasible to implement. More importantly, the need to draw meaningful conclusions from an aggregation of small studies may help explain this exponential growth.

The use of meta-analytic techniques is controversial when the contributing studies are not randomized control trials (RCT). Many researchers feel that it is highly misleading to attempt to combine a series of disparate studies **[2]** while others maintain that, with proper safeguards, meta-analysis allows an extremely useful pooling of smaller studies **[3]**, **[4]**. A discussion of the appropriateness of meta-analysis is beyond the scope of this paper. Rather, the focus here will be on minimizing unnecessary error in testing the overall statistical significance of a meta-analysis and in estimating the Effect Size.

Overview of 2 x 2 x k Categorical Meta-Analysis

The "2 x 2 x k" categorical meta-analysis paradigm is probably the most frequently encountered situation in meta-analysis. It consists of a series of *k* contributing studies each described by a 2 x 2 contingency table. Every cell of each 2 x 2 table contains the number of occurrences of an event (e.g., disease cases) for the particular combination of row and column variables. For the sake of illustration, we can associate the two columns

of each table with Disease Manifestation vs. No Disease Manifestation and the two rows with Exposure vs. No Exposure. Table 1 represents the results of one of these *k* studies

Table 1 Typical contributing study (one of *k*) in a dichotomous meta-analysis

	Disease Status		
Exposure	**Disease Manifestation**	**No Disease Manifestation**	**Total**
Exposure	4	96	100
No Exposure	2	98	100
Total	6	194	200

In most meta-analyses, there are typically two distinct components: 1) A statistical test of the overall difference between the Exposure and No Exposure groups across the *k* contributing studies and 2) A method to pool the observed differences between groups across the *k* studies in order to estimate the true difference (the Effect Size).

Surprisingly, in recent years, many epidemiologists employing meta-analytic techniques have greatly deemphasized the first component. Borenstein *et al.* **[3]** conclude:

> ". . . However, meta-analysis also allows us to move beyond the question of statistical significance, and address questions that are more interesting and also more relevant." (pp. 11-12).

Similarly, Higgins *et al.* **[4]** rather dismissively state:

> ". . . If review authors decide to present a p value with the results of a meta-analysis, they should report a precise p value, together with the 95% confidence interval" (pp. 371-372).

A method is developed that maintains the Type I error ("false alarm rate") at the desired level, but which has good power to detect true differences across a large range of event probability, number of contributing studies, sample size and level of heterogeneity. An argument can be made that maintaining the Type I error at a pre-specified level is more important than the power (1 – Type II error rate) to detect true differences between conditions. The framers of modern statistical testing called such errors "Errors of the First Kind" and placed a special emphasis on them. Neyman & Pearson in 1933 stated:

> "A new basis has been introduced for choosing among criteria available for testing any given statistical hypothesis, H_0, with regard to an alternative H_t. If Θ_1 and Θ_2 are two such possible criteria and if in using them there is the same chance, ε, of rejecting H_0 when it is in fact true, we should choose that one of the two which assures the minimum chance of accepting H_0 when the true hypothesis is H_t." **[5]**[p. 336]

Thus, while Neyman & Pearson supported the effort to choose criteria that yield the greatest power to detect true differences, this effort is secondary to maintaining a pre-specified level of Type I error. A second exact method is developed to estimate the effect size of any statistically significant finding.

"Rare" Events and Meta-Analysis

The probability of occurrence of a disease is often categorized as "rare" although no specific definition exists. As an example, Higgins *et al.* state that "There is no single risk at which events are classified as 'rare' " but gives as examples 1 in a 100 or 1 in a 1,000 (see **[6]**, p. 520). An obvious related issue is observing zero cases in one or more cells of

a contingency table. Table 2 shows the expected cell sizes from various realistic combinations of disease probability and contributing study sample size.

Table 2 Expected number of disease cases as a function of disease probability and individual study sample size (each arm)

		Individual Study Sample Size (each arm)		
Disease/Condition	**Approximate Disease Prob.**	**100**	**500**	**1000**
Myocardial Infarction Incidence Rate for Age ≥ 60 years	.011 **[7]**	1.1	5.5	11
Parkinson's Disease Incidence Rate (60 -69 Age Group)	.00058 **[8]**	.06	.29	.58
Alzheimer's Disease Incidence Rate (60 -74 Age Group)	.002 **[9]**	.2	1	2
Lung Cancer Incidence Rate for White Males	.00051 **[10]**	.05	.26	.51

Table 2 supports the notion that "rare" events are a focus of many epidemiological studies.

For homogeneous meta-analysis (i.e. where the effect across studies may be assumed to be the same within statistical variation), the two techniques typically used for categorical data are the Mantel-Haenszel and Peto techniques. Both of these techniques rely on the Mantel-Haenszel Chi Square to test for the overall statistical significance. For heterogeneous meta-analyses, the asymptotic DerSimonian-Laird (DL) inverse variance technique is typically used **[1]**.

The problem in applying large sample asymptotic techniques to meta-analyses involving small numbers of cases will be illustrated in the older and much more developed domain of homogeneous meta-analyses. Mantel developed what is probably the most widely used technique for homogeneous meta-analyses **[11]**. In applying his technique, he

showed that a minimum of approximately five cases was required in each of the 4 cells of each of the 2 x 2 tables for each of the *k* studies comprising the meta-analysis **[12]**. This is the same heuristic requirement typically used without any particular justification for the simple chi-square test. Mantel & Fleiss reviewed the options when a reasonable number of cases was not present in all cells:

> "The investigators could have obtained data from very many more tables to make things more asymptotic for use of M-H [note: this is the Mantel-Haenszel technique], or they could readily have applied a more exact procedure for the data at hand" [p. 134].

R. A. Fisher made essentially the same plea in 1925 in the preface to the first edition of his well-known *Statistical Methods for Research Workers* **[13]**:

> "Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow. The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data."

Both criticisms suggest the use of exact methods to handle the sparseness of the underlying contingency tables at least for the disease examples contained in Table 2. All but two of the combinations of individual study sample size and disease probability shown in Table 2 would yield fewer than five cases per cell leading to violations of the minimum cell size in the Mantel-Haenszel (MH) Chi Square test, and thus the test would be potentially flawed. In addition, these two cases were for sample size equal to 500 and 1,000 which may not represent many realistic studies. While this limitation of the MH Chi Square test was known to Mantel and others (e.g. **[12]**), it seems to generally have been forgotten for meta-analysis of 2 x 2 x k categorical data. The continued use of an asymptotic test in situations not suited for its use is unacceptable given the computer power that is now available to all researchers.

Heterogeneity vs. Homogeneity in Meta-Analyses

The term "heterogeneity" refers to the fact that studies done at different times and by different researchers might be expected to yield different results. The expectation is that a variable of interest may be dependent, at least in part, to one or more other variables. The meta-analysis researcher, J. P. T. Higgins stated "As Heterogeneity is to be expected in a meta-analysis: it would be surprising if multiple studies, performed by different teams in different places with different methods, all ended up estimating the same underlying parameter." [**[14]**, p. 158]. While researchers may agree that heterogeneity is to be expected, there is very little agreement on how to quantify this variability. The obvious candidate is τ^2, the estimated variability between studies. However, τ^2 is not invariant across study designs and its interpretation may not be intuitive. Alternatives

include $\boldsymbol{I^2}$, the ratio of the inter-study variability to the total variability and the *Q* statistic, which is mathematically related to $\boldsymbol{I^2}$ (see, e.g., **[15]**).

In the technique described in this paper, heterogeneity will be mathematically manipulated through τ^2 using the logit distribution as developed by Bhaumik *et al.* **[16]**. Namely:

$$\boldsymbol{x_{Ci} \sim B(p_{Ci}, n_{Ci}),\ \ x_{Ei} \sim B(p_{Ei}, n_{Ei}),} \quad \textbf{[1]}$$

$$\boldsymbol{logit(p_{Ci}) = \mu + \varepsilon_{1i},\ \ logit(p_{Ei}) = \mu + \theta + \varepsilon_{1i} + \varepsilon_{2i}} \quad \textbf{[2]}$$

$$\boldsymbol{\varepsilon_{1i} \sim N(0, \gamma^2)} \quad \textbf{[3]}$$

$$\boldsymbol{\varepsilon_{2i} \sim N(0, \tau^2)} \quad \textbf{[4]}$$

Where:

B is the Binomial Distribution

N is the Normal Distribution

X_{Ci}, X_{Ei} are the observed number of cases in the control and exposure groups respectively of the *ith* study

p_{Ci}, p_{Ei} are the event probabilities in the control and exposure groups respectively of the *ith* study

n_{Ci}, n_{Ei} are the sample sizes in the two groups of the *ith* study

μ corresponds to the background event (disease) probability in the exposure and control groups

$\boldsymbol{\theta}$ corresponds to the overall ratio of the event probabilities for the exposure group relative to the control group

$\boldsymbol{\gamma^2}$ is a variance corresponding to the uncertainty of the observed disease probability in both the exposure and control groups of the *k* contributing studies.

$\boldsymbol{\tau^2}$ is a variance corresponding to the heterogeneity or the "heterogeneity parameter" only in the exposure group.

$\boldsymbol{\varepsilon_{1i}}$ is a Normal distribution deviation in background event (disease) probability for both the exposure and control groups of each of the contributing studies.

$\boldsymbol{\varepsilon_{2i}}$ is a Normal distribution deviation in background disease probability due to heterogeneity in the exposure group of each of the contributing studies

The Basic Principles of the Dersimonian-Laird (DL) Method

As stated above, this research specifically contrasts an exact method for conducting meta-analyses in *k* 2 x 2 tables with heterogeneity with the most popular approach which was developed by DerSimonian and Laird (DL) **[1]**.

For each contributing study, the DL technique calculates the logarithm of the sample odds ratio and a corresponding estimate of the variance of this measure based on the asymptotic distribution of these logarithms. Adjustments are made for entries in the individual 2 x 2 tables that contain a zero-cell count. Equations 5-8 below capture the core DL approach. In Equation 5, an estimate of the interstudy variability, $\boldsymbol{\tau^2}$, is first derived from Cochran's Q statistic and the weights assigned to each of the *k* contributing studies, $\boldsymbol{\omega_i}$. These weights are equal to the inverse of the square of the standard error of the estimate of the odds ratio, $\boldsymbol{\widehat{\theta}_i}$, in each of the *k* contributing studies.

$$\hat{\tau}^2 = \frac{Q-(k-1)}{\sum \omega_i - \left(\frac{\sum \omega_i^2}{\sum \omega_i}\right)} \quad (5)$$

As shown in Equation 6, a new set of weights, ω_i', are then calculated based on the estimated value of $\hat{\tau}^2$ from Equation 5 and the standard errors of the contributing studies.

$$\omega_i' = \frac{1}{SE(\hat{\theta}_i)^2+\tau^2} \quad [6]$$

These new weights are then used to calculate estimates of both the overall log odds ratio, θ_{DL} and its standard error as shown in Equation 7 and 8.

$$\hat{\theta}_{DL} = \frac{\sum \omega_i' \hat{\theta}_i}{\sum \omega_i'} \quad [7]$$

$$SE(\hat{\theta}_{DL}) = \frac{1}{\sqrt{\sum \omega_i'}} \quad [8]$$

A test of statistical significance is then based on a large sample normal distribution. The DL technique requires asymptotic assumptions regarding both the Q statistic used to

estimate the interstudy variability, $\boldsymbol{\tau^2}$, and the normal distribution required to test for statistical significance. A more subtle issue is the possibility of distorting correlations between the individual estimates of the effect size for each contributing study, $\boldsymbol{\theta_i}$, and the individual weights used for each of these contributing effect sizes.

RESULTS

A Non-Parametric Exact Test of Overall Statistical Significance for Dichotomous Categorical Meta-Analysis

Jakob Bernoulli's notion of what is now called a Bernoulli Trial offers the basis for a non-parametric approach to aggregating multiple epidemiological studies based on dichotomous categorical data. The enhancements to the Bernoulli method developed in this paper offer a practical exact method for assessing the overall statistical significance and effect size of a dichotomous meta-analysis.

One of the many important contributions of this outstanding 17th century mathematician was the idea of the fixed probability of an event over a sequence of independent trials which led to what is now called Bernoulli Trials and to the related Binomial Distribution. In brief, Bernoulli viewed a set of statistical events as a series of independent coin flips with each flip having a probability p of obtaining a head and q = 1 - p of obtaining a tail. This hypothetical coin is often treated as a fair coin where both p and q equal .5. The simplest Bernoulli Trials approach encompasses a series of *n* flips and answers questions of the type: what is the probability of observing x heads in n such flips? (See for example Rosner *[17]*). In epidemiology, one could consider each of the *k* contributing studies of a meta-analysis as a single Bernoulli Trial with p = .5. Then the combination of the *k*

studies could be analyzed as a binomial distribution. This is the standard Sign Test (see, for example, *[18]*).

For example, for a meta-analysis of 20 studies, if 15 out of 20 studies had more cases in the exposure group than in the control group, we could ask: What is the probability that 15 or more of the 20 studies could have shown a larger effect in the exposure group strictly by chance alone? If this cumulative probability is less than a pre-specified level of Type I error (e.g. .05), one would reject the null hypothesis and conclude there probably exists a statistically reliable relationship between exposure and the end point used.

The principal reason that this approach has seen little use in practical epidemiology is that it suffers from two critical deficits. First, the dichotomous Bernoulli heads vs. tails approach doesn't deal with the third possibility of a tie. The author of this study believes that no truly useful method to date has been offered to deal with those situations when there are an identical number of events in each of the exposure and the control arms of a study other than to discard the study. Second, a truly exact EBT method requires a complete convolution of the frequency distributions of the contributing studies in order to derive the combined frequency distribution. Even for equal sample size, each of the k contributing studies could have a different Bernoulli probability, p, requiring a full convolution to determine the null distribution of the total number of times there were more cases in the exposure group relative to the control group across the k contributing studies. Before dealing with the ties problem, the determination of the combined distribution will be outlined.

Combining the Individual Studies Contributing to the Meta-Analysis. A critical problem is finding a method for combining the individual study binomial distributions of the *k* contributing studies each with a possibly different p value into an overall frequency distribution.

Prior to the widespread availability of computing power, the convolution of a large number of individual binomial distributions was typically handled by approximate methods given the unwieldy nature of the calculations. Even with the advent of available computer power, convolution is still often impractical. As an example, for a meta-analysis involving 24 studies each with a unique binomial distribution p , there are over 2 million unique combinations of the studies that need to be considered just to calculate the single discrete probability that exactly 12 of the 24 studies have more cases in the exposure group than in the control group.[1] However, an exact algorithm was laid out in a readily implementable fashion by Butler and Stephens in a 1993 technical report *[19]* which can easily be implemented even on a personal computer. The algorithm yields the exact probability distribution of the convolution of individual binomial distributions corresponding to the specific studies contributing to the meta-analysis. The method makes use of a recurrence relationship inherent in the binomial distribution which allows the semi-automatic calculation of its probabilities without resort to the simple but overwhelmingly inefficient enumeration of all of the possible combinations of studies. This easily established relationship can be stated as:

[1] This number of combinations is simply C (24,12) = 2,704,156.

311 $$P(X = 0) = (1 - p)^n \qquad if\ j = 0$$

312

313 $$P(X = j) = \left\{ \frac{(n - j + 1)}{j} \right\} * \left\{ \frac{p}{(1 - p)} \right\} * P(X = j - 1) \qquad if\ j \geq 1$$

314

315 Figure 2 compares the estimated number of computer executable steps required in the
316 Butler & Stephens method compared to a traditional convolution.

317

318 **Insert About Here**

319 **Fig. 2 Estimated computer executable steps per Butler & Stephens vs. traditional**
320 **convolution.**

321

322 As can be seen, a traditional convolution is only tractable when the number of
323 contributing studies is less than or equal to approximately 20.

324 **The Ties Problem.** The next problem in adapting the standard Bernoulli Trials
325 technique to practical meta-analysis is a procedure to deal with the situation where there
326 are an identical number of cases in both the exposure and control arms of a study
327 contributing to the meta-analysis. In studies with small sample sizes and/or low disease
328 probabilities, the highest probability tie is typically the "0/0" tie in which no cases are
329 observed in either the exposure or the control arms.

330 A first step in dealing with ties is to more clearly define the criteria for a "success". The
331 present EBT approach defines a success as there being a **strictly** greater number of cases
332 in the exposure group relative to the control group. Under this definition, the same

number of cases in both arms of the study or more cases in the control arm of the study is considered a "failure". In essence, this is a trinomial situation. There are successes, failures and ties. We are simply combining the failures where there are more cases in the control group relative to the exposure group and tie situations and calling the combination "failures."

Equation 9 below forms the basis of the EBT method. The Greek capital letter "Π has been chosen to specify the probabilities of there being more cases in one arm of the study relative to the other to differentiate these parameters from the underlying disease probabilities:

$$\boldsymbol{\Pi_{E_i} + \Pi_{C_i} + prob\ (tie)_i = 1} \quad [9]$$

where:

$\boldsymbol{\Pi_{E_i}}$ = probability of there being strictly more cases in the exposure group relative to the control group in Study *i*

$\boldsymbol{\Pi_{C_i}}$ = probability of there being strictly more cases in the control group relative to the exposure group in Study *i*

$prob(tie)_i$ = probability of finding exactly same number of cases in both groups of Study *i.*

Assuming that $\boldsymbol{\Pi_{E_i}}$ and $\boldsymbol{\Pi_{C_i}}$ would be equal under the null hypothesis of no difference between exposure and control groups and rearranging terms, we have:

$$\boldsymbol{2\Pi_{E_i} + prob(tie)_i = 1} \quad [10]$$

Solving for $\boldsymbol{\Pi_{E_i}}$ we have:

$$\boldsymbol{\Pi_{E_i}} = \frac{\mathbf{1}-\boldsymbol{prob}(\boldsymbol{tie})_i}{\mathbf{2}} \quad [11]$$

Thus, the only requirement for calculating the $\boldsymbol{\Pi_{E_i}}$ parameter for each contributing study is to first determine the probability of all tie situations for the study.

This is a very straightforward procedure. To determine $\boldsymbol{prob}(\boldsymbol{tie})_i$ for each of the contributing studies, all of the tie situations need to be enumerated and then summed together.

As a simple example, assume that Study *i* has 100 participants in each of its exposure and control arms and that the underlying event (disease) probability *p* is 01.

The probability that there are no cases among these 100 participants in the exposure arm would then be:

$$\boldsymbol{Prob}(\mathbf{0}\ \boldsymbol{cases}) = .\mathbf{01}^{\mathbf{0}} * (\mathbf{1}-.\mathbf{01})^{\mathbf{100}} = .\mathbf{99}^{\mathbf{100}} = .\mathbf{37}$$

Similarly, the probability of there being no cases in the control arm would also be .37.

Thus, the probability of a "0,0" tie would be $.\mathbf{37}^{\mathbf{2}} = .\mathbf{13}$ which is surprisingly large.

Table 3 lists the probabilities for the first five tie situations and sums these probabilities to determine $\boldsymbol{prob}(\boldsymbol{tie})_i$.[2]

Table 3 Probability of observing exactly the same number of cases in both the exposure and control groups for background event probability equal to .01 and sample size equal to 100 as a function of the number of observed cases

Number of Cases in Each Group	Probability

[2] The probabilities for six or more ties decrease to extremely small values. However, in actuality, the EBT method calculates all possible ties in calculating $\boldsymbol{prob}(\boldsymbol{tie})_i$

0	.13
1	.137
2	.034
3	.004
4	.0002
5	.00001
Total	.309

As shown in Table 3, there is over a 30% probability of obtaining a tie for 0 cases through 5 cases in both the exposure and control groups. Applying Equation [11] to this hypothetical study, we see that, under the null hypothesis of equal probabilities, $\boldsymbol{\Pi_{E_i}}$ and $\boldsymbol{\Pi_{C_i}}$ are both equal to .35. Thus, due to ties, the nominal .50 value for $\boldsymbol{\Pi_{E_i}}$ **and** $\boldsymbol{\Pi_{C_i}}$ has been greatly reduced.

The EBT technique is indeed a "vote counting" method and such methods have been greatly disparaged by Rothman **[20]** among others. However, unlike a simple Sign Test, the EBT method is based on a reasonable approach to the ties problem and combines the individual $\boldsymbol{P_{E_i}}$ values by doing the equivalent of a formal convolution of the frequency distributions of the individual contributing studies.

A Non-Parametric Exact Method for the Estimation of Effect Size for Dichotomous Categorical Meta-Analysis

Basic Estimation Technique

A second exact technique was developed to estimate the effect size for dichotomous categorical meta-analysis. As a starting point, one might simply from the ratio of the average observed event probabilities, $\boldsymbol{p_{E_i}}$ and $\boldsymbol{P_{C_i}}$, in the exposure and control groups respectively of each study and average these ratios across the *k* contributing studies. This simple approach, however, is highly biased. As shown in the underlying model that is described in Equations 1 – 4, the number of observed "successes" in the exposure and control arms of the *k* contributing studies each depend on an identical source of variation captured by $\boldsymbol{\varepsilon_{1i}}$ in the model. The exposure group, however, contains an additional source of variation, captured by $\boldsymbol{\varepsilon_{2i}}$ in the model. Figure 3 illustrates the problem of estimating the effect size by simply forming the ratio of $\boldsymbol{p_E}$ to $\boldsymbol{p_C}$.

Insert About Here

Fig. 3 Demonstration of inappropriateness of comparing the $\boldsymbol{p_E}$ and $\boldsymbol{p_C}$ distributions to estimate Effect Size.

Even for the relative risk of 1.0 depicted in the figure, the exposure distribution will have positive excursions that are not compensated for by equally robust negative excursions at least for small (rare) values of event probability.

The differential skew of the $\boldsymbol{p_{E_i}}$ distribution relative to the $\boldsymbol{p_{C_i}}$ distribution was used to address this issue. The additional skew in the exposed group due to the source of $\boldsymbol{\varepsilon_{2i}}$ was estimated by taking the difference between the total exposure group skew and the expected skew from a pure binomial with the same observed event probability. The

observed average $\boldsymbol{p_E}$ across the *k* contributing studies was then reduced by a factor proportional to this difference in skew levels.

Monte Carlo Simulation of The EBT And DL Techniques for Statistical Significance and Effect Size Estimation

A series of Monte Carlo simulations was conducted to evaluate the EBT statistical significance test and the effect size estimation techniques and to compare them to the typically used DerSimonian-Laird Inverse Variance technique. The simulation was written and executed in *R: A Programming Environment for Data Analysis and Graphics.* **[21]**. The DerSimonian results were calculated using the "meta" package in R.

Five levels of relative risk (ratio of exposure group to control group event probability) of 1.0, 1.25, 1.5, 1.75, and 2.0 were crossed with three levels of disease background event probability (.005, .01, and .05), and three levels of sample size (50,100 and 200). Finally, the number of studies entering into each meta-analysis was chosen to be 5, 10, 20, or 40 studies.

In addition, the heterogeneity between the contributing studies, τ^2 [Equation 4], was evaluated at 0 (homogeneity), .4, and .8. This last value of .8 represents a very large variance among the studies and was partially chosen to be able to compare the results with previous work (**[22]** and **[16]**). As an example, at τ^2 =.8 a nominal exposure group event probability $\boldsymbol{p_E}$ of .05 would vary from of .009 to .24 which is almost a 30:1 ratio. Finally, the common variability in both the exposure and control groups represented by $\boldsymbol{\gamma^2}$ in Equation 1 was chosen to be .5 to allow direct comparison with earlier work (**[22]** and **[16]**).

427 The statistical significance and effect size were evaluated using both the EBT and
428 DerSimonian techniques for each replication. All simulation runs were conducted with
429 10,000 replications. A value of .05 was used as the pre-specified level of Type I Error.
430 The "Mid-P" technique advocated by Agresti **[23]** and others was used to determine the *p*
431 values in a less conservative manner leading to more realistic power levels.
432

433 **Results from the Monte Carlo Simulations: Testing Statistical Significance**

434 Figures 4 and 5 show the results using the EBT and DL methods respectively.
435 To simplify presentation, only scenarios in which the expected number of cases was
436 greater than or equal to two were utilized. Table 4 shows the included scenarios.

Table 4 Scenarios included in the analysis of statistical significance

Background Event Probability	Sample Size	Expected Number of Observed Cases
.01	200	2
.05	50	2.5
.05	100	5
.05	200	10

437
438
439 **Insert About Here**

440 **Fig. 4 Power for number of studies, relative risk, and heterogeneity for EBT. A-D**
441 **correspond to studies equal to 5, 10, 20 and 40.**

442
443 **Insert About Here**

Fig. 5 Power for number of studies, relative risk, and heterogeneity for DL. A-D correspond to studies equal to 5, 10, 20 and 40.

The basic finding was that the EBT method maintained the prespecified level of Type I error for both the homogeneous and heterogeneous scenarios while the DL method had many violations of this level for heterogeneous scenarios. This limitation of the DL method is highlighted in the inserts in Figure 5 which show the power when the Relative Risk is one (i.e. the false alarm rate). For the homogeneous scenario where $\tau^2 = 0$, both the EBT and the DL methods respect the prespecified Type I error level. However, for $\tau^2 = .4$ and for $\tau^2 = .8$, the DL method exhibits large violations of this level. Not unexpectedly, as the number of contributing studies increases, the power for Relative Risk greater than one increases for both the EBT and DL methods for values of Relative Risk of greater than 1.0. A separate analysis showed that the standard deviation of the power estimates in Figures 4 and 5 was less than or equal to .42% (i.e. .0042).

In actuality, comparing the power between the EBT & DL techniques for event ratios greater than 1.0 is not possible due to the large number of violations of the pre-specified Type 1 Error violations for the DL technique.

Figure 6 is a comparison of Type I Error (false alarm rate) for the current EBT technique and the DL technique as a function of heterogeneity (τ^2).

Insert About Here

Fig. 6 Type I error for EBT and DL methods as a function of heterogeneity.

As can be clearly seen, the current EBT technique is relatively resistant to the effects of increasing heterogeneity over a very large range. The DL technique, however exhibits a

monotonically increasing sensitivity to heterogeneity. A related aspect of any meta-analysis technique's ability to perform well in the face of heterogeneity is its resistance to "contamination" from one or a small number of "rogue studies." Since the EBT method does not directly allow such rogue studies to directly affect the test statistic, it should be much more resistant to these distortions.

The large costs of discreteness have been studied by Agresti **[24]** and others.

A first cost of discreteness results when the number of contributing studies is small. The general issue of overcoverage is highlighted in Figure 7.

Insert About Here

Fig. 7 Interval overcoverage as a function of the number of contributing studies.

The overcoverage is greatest for the smallest number of *k* contributing studies, and generally decreases as the number of contributing studies increases. As Figure 7 demonstrates, even an unrealistic level of 500 contributing studies is still associated with a relatively large level of overcoverage. While such discreteness clearly reduces power, it could be argued that a statistically significant finding based on extremely sparse tables and a handful of studies requires stronger evidence. Unfortunately, the majority of meta-analyses consist of fewer than two or three studies as Kontopantelis *et al.* have shown in their extensive analysis of all meta-analyses in the Cochrane Library **[25]**.

Additional Monte Carlo testing was done for unbalanced designs (unequal sample sizes in the exposure and control arms of the contributing studies) and meta-analyses with unequal sample sizes across contributing studies. Table 5 shows the sample sizes for the two groups for a typical unbalanced design in which the control group sample size is

490 twice the exposure group sample size. The sum of the two sample sizes across both arms
491 of the study was chosen to be 200 yielding an average sample size of 100 to allow
492 comparison with the balanced designs of Figures 4 and 5.
493

Table 5 Sample sizes for simulation of unbalanced designs

Number of studies = 10

	Study #									
Group	1	2	3	4	5	6	7	8	9	10
Exposure	66	66	66	66	66	66	66	66	66	66
Control	134	134	134	134	134	134	134	134	134	134

494

495 Table 6 below shows the results of the simulation for heterogeneity values $\tau^2 = 0$ and $\tau^2 =$
496 .8, Event ("disease") Probability of .05, Number of Studies = 10, and Sample Size (avg.)
497 = 100 at the same five levels of Relative Risk used above. The simulation run consisted
498 of 10,000 replications as in Figures 4 and 5.
499

Table 6 Power (%) for the unbalanced design of table 5

τ^2 (heterogeneity) equal to 0 and .8; event probability = .05;

Number of studies equal to 10; Sample size (per study arm) equal to 100

	Heterogeneity									
	0					.8				
	Risk Ratio									
Technique	1.0	1.25	1.5	1.75	2.0	1.0	1.25	1.5	1.75	2.0
EBT	2.1	14.4	43.3	71.0	88.4	4.2	11.0	21.6	35.0	46.9
DerSimonian & Laird	2.2	16.5	57.2	87.5	97.7	11.5	24.0	41.6	59.2	72.8

As the results in Table 6 indicate, the Type I Error (Relative Risk = 1.0) remained below the specified value of five percent for the EBT technique and was far above this point for the DerSimonian technique when the heterogeneity was equal to .8.

Table 7 below shows the sample sizes for the exposure and control groups for each of the contributing studies for a design with unequal sample size across the contributing studies. This particular design was chosen as a relatively extreme case. As can be seen, the average sample size for both the groups was maintained at 100 to allow comparison of the simulation results with the equal sample size scenarios of Figures 4 and 5.

Table 7 Sample sizes for simulation of unequal sample size designs

Number of studies equal to 10

	Study #									
Group	1	2	3	4	5	6	7	8	9	10
Exposure	175	25	175	25	175	25	175	25	175	25
Control	175	25	175	25	175	25	175	25	175	25

Table 8 below shows the results of the simulation for a heterogeneity values of $\tau^2 = 0$ and $\tau^2 = .8$, Event ("disease") Probability of .05, and Sample Size (individual study arm average) = 100, at the same five levels of Relative Risk as used above. The simulation run consisted of 10,000 replications as in Figures 4 and 5.

Table 8 Power (%) for the unbalanced design of table 7

τ^2 (heterogeneity) equal to 0 and to .8; Event probability = .05;

Number of studies equal to 10; Sample size (avg. per individual study arm) equal to 100

	Heterogeneity									
	0					**.8**				
	Risk Ratio									
Technique	1.0	1.25	1.5	1.75	2.0	1.0	1.25	1.5	1.75	2.0
EBT	2.0	14.9	43.0	70.0	87.7	4.3	11.4	22.4	34.5	47.9
DerSimonian & Laird	2.4	16.6	56.6	87.2	97.8	11.4	25.1	41.6	58.2	73.5

Most importantly, the EBT Technique was superior at protecting the pre-specified level of Type I Error relative to the DL technique at a heterogeneity level of .8.

A clear finding of the Monte Carlo simulations common to both meta-analysis techniques studies is the apparent fruitlessness of searching for small effect sizes. Both the EBT and DL techniques are very poor at reliably finding statistically significant results until the relative risk approaches 2.0. While this finding does not directly bear on the issues studied in this report, it does serve as a cautionary tale to those who continue to try to tease out very small effects especially from sparse data.

Results from the Monte Carlo Simulations: Effect Size Estimation

Figures 8 and 9 capture the basic findings for estimating the Effect Size.

Insert About Here

Fig. 8 Effect Size as function of Relative Risk and heterogeneity. A & B correspond to the EBT and DL methods.

Again, only simulation scenarios in which the expected number of observed cases was greater than or equal to two were utilized. Since the effect of the number of studies contributing to the meta-analysis was small for this effect size estimation, results were averaged across this variable. Figure 8 shows the results for the EBT method and Figure 9 for the DL method. As shown in the figures, both methods were reasonably successful at estimating the levels of relative risk. However, the EBT method generally underestimated the relative risk for $\tau^2 = 0$ and overestimated it for $\tau^2 = .8$ while the DL method tended to underestimate the relative risk. Finally, the interquartile range for the DL method was considerably smaller than for the EBT method as shown in Figures 10 and 11.

Insert About Here

Fig. 9 Semi-Interquartile Range as function of Relative Risk and heterogeneity. A & B correspond to the EBT and DL methods.

CONCLUSIONS AND SUGGESTIONS FOR THE FUTURE

This research has developed an exact test for the meta-analysis of dichotomous, categorical data and a related method to estimate the size of the effect.

The Enhanced Binomial Technique (EBT) to Assess Statistical Significance

The EBT technique was greatly superior to the DerSimonian technique in maintaining a pre-specified level of Type I Error. As shown, the DerSimonian technique demonstrated many large violations of this level when heterogeneity was present. Given the various biases towards finding statistical significance prevalent in epidemiology today, a strong focus on maintaining a pre-specified level of Type I Error would seem critical (see, e.g., **[26]**). The EBT approach is greatly superior at maintaining this pre-specified value of Type I Error in the face of even extreme heterogeneity.

The Enhanced Binomial Technique (EBT) to Estimate Effect Size

A related but separate method was developed to estimate the effect size. This new technique was comparable to the often-used DL method although both methods demonstrated some accuracy issues. The DL method exhibited a somewhat smaller Semi-IQR variability. The fact that the EBT method was clearly superior in assessing statistical significance while the DL method demonstrated a smaller variability in estimating effect Size supports the possible utility of separating these two procedures as outlined at the beginning of this article. One possibility is to use the EBT and DL methods for statistical significance assessment and effect size estimation respectively.

While statistical programs providing exact solutions already exist such as Cytel's StatXact, they are beyond the means of most practicing statisticians and epidemiologists.

For example, Cytel Inc. currently lists a price of over $900 USD for their current version, StatXact 11 **[27]**

The techniques developed here are written in the almost universal statistical language of R and are freely available from the author. As such, it is hoped that other researchers would be able to extend and improve these initial versions.

As outlined in this report, the use of meta-analysis in epidemiology is increasing very rapidly and appears to be meeting an important need. Fortunately, inexpensive and readily available computer power has also vastly increased in the past forty years. For example, task speed as measured in Million Instructions per Second ("MIPS") has increased from .64 for the IBM370 mainframe computer in 1972 to 238,000 for an Intel Pentium processor personal computer in 2014. **[28]**. By using the techniques developed here and the computer power available to all researchers today, the determination of statistical significance and the estimation of effect size can be readily accomplished without unnecessary error.

Declarations

Ethics approval and consent to participate

No experimental participants

Consent for publication

No consent required

Availability of data and material

Both software programs are freely available from the author

Competing interests

The author has no competing interests

Funding

No funding was obtained for this work

Authors' contributions

Work was fully done by L. Paul

Acknowledgements

No acknowledgements

REFERENCES

1. **DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7.3: p. 177-188.**
2. **Shapiro S. Meta-analysis/Shmeta-analysis. Am J Epidemiol. 1994 Nov 1; 140(9): p. 771-8.**
3. **Borenstein M, et al.. Introduction to meta-analysis.: John Wiley & Sons; 2009.**
4. **Higgins JP, editor. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell; 2008.**
5. **Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character. 1933: p. 289-337.**

6. **Higgins J, Deeks JJ, Altman DG. Special topics in statistics. In Higgins J, editor. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. Chichester: John Wiley; 2008. p. 481-529.**
7. **Mons U, Müezzinler A, Gellert C, Schöttker B, Abnet C, Bobak M, et al. Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium. bmj. 2015 Apr 20; 350(h1551).**
8. **Hirsch L, Jette N, Frolkis A, Steeves T, Pringsheim T. The incidence of Parkinson's disease: a systematic review and meta-analysis. Neuroepidemiology. 2016; 46(4): p. 292-300.**
9. **Alzheimer's Association 2015 Alzheimer's disease facts and figures. Alzheimer's & dementia. Journal of the Alzheimer's Association. 2015 Mar; 11(3): p. 332.**
10. **Torre L, Siegel R, Ward E, Jemal A. Global cancer incidence and mortality rates and trends—an update. Cancer Epidemiology and Prevention Biomarkers. 2016 Jan; 25(1): p. 16-27.**
11. **Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959 Apr; 22(4): p. 719-48.**
12. **Mantel N, Fleiss J. Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure. Am J Epidemiol. 1980; 112(1): p. 129-34.**
13. **Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.**
14. **Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. International Journal of Epidemiology. 2008; 37: p. 1158-1160.**
15. **Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. British Medical Journal. 2003; 327(7414): p. 557.**
16. **Bhaumik DK, Amatya A, Normand SLT, Greenhouse J, Kaizar E, Neelon B, et al. Meta-analysis of rare binary adverse event data. Journal of the American Statistical Association. 2012; 107(498): p. 555-567.**
17. **Rosner B. Fundamentals of Biostatistics. 5th ed.: Duxbury Press; 1999.**

18. **Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. 2nd ed.; 1988.**

19. **Butler K, Stephens M. Distribution of a Sum of Binomial Random Variables. Technical Report No. 467 prepared under contract N00014-92-5-1264 (NR-042-267) for The Office of Naval Research. Palo Alto: Stanford University; 1993.**

20. **Rothman K, Greenland S. Meta-analysis. Some methods to avoid: Qualitative tally (vote counting) and Quality scoring. New York: Lippincott Williams Wilkins; 1998.**

21. **The R Project for Statistical Computing.. Available from: https://www.r-project.org/.**

22. **Paul LM. Cannons and sparrows: an exact maximum likelihood non-parametric test for meta-analysis of k 2× 2 tables. Emerging themes in epidemiology. 2018 December; 15(1).**

23. **Agresti A. A survey of exact inference for contingency tables. Statistical science. 1992; 7(1).**

24. **Agresti A. Dealing with discreteness: makingexact'confidence intervals for proportions, differences of proportions, and odds ratios more exact. Statistical Methods in Medical Research. 2003 Feb; 12(1).**

25. **Kontopantelis E, Springate D, Reeves D. A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses. PloS one. 2013 Jul 26; 8(7).**

26. **Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005: p. e124.**

27. **Cytel Inc. Cytel. [Online].; 2020 [cited 2020 November 17]. Available from: https://store.cytel.com/products/statxact?hsCtaTracking=be2ed66d-9346-4239-ad8b-c19193bfcda0%7C40b5c432-854f-4116-a2d2-079223b15428.**

28. **Wikipedia. [Document: "Instructions per Second"].; 2016 [cited 2016 June 3. Available from: https://en.wikipedia.org/wiki/Instructions_per_second.**

614

Figures

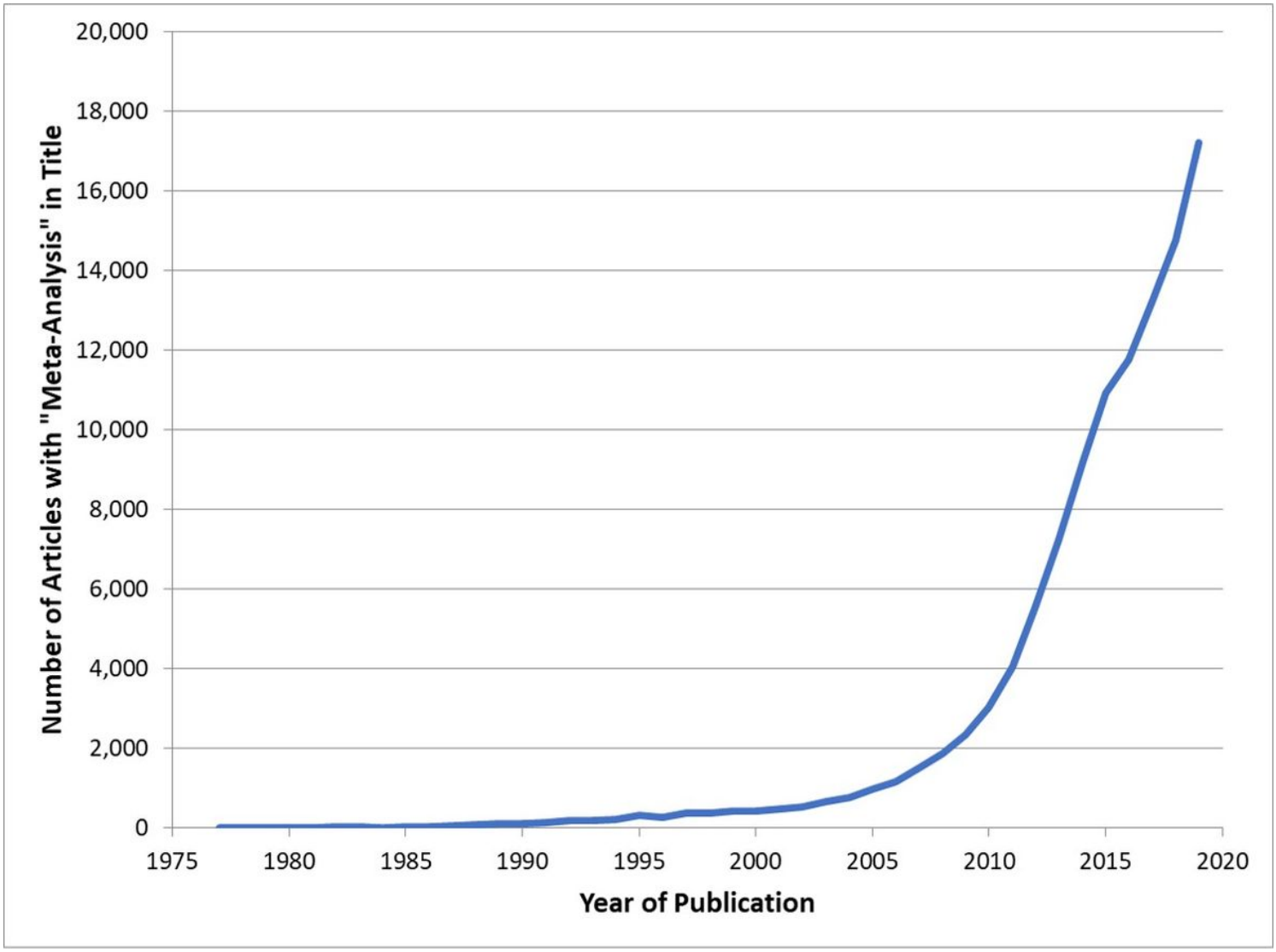

Figure 1

Number of articles containing "meta-analysis" in the title by year of publication.

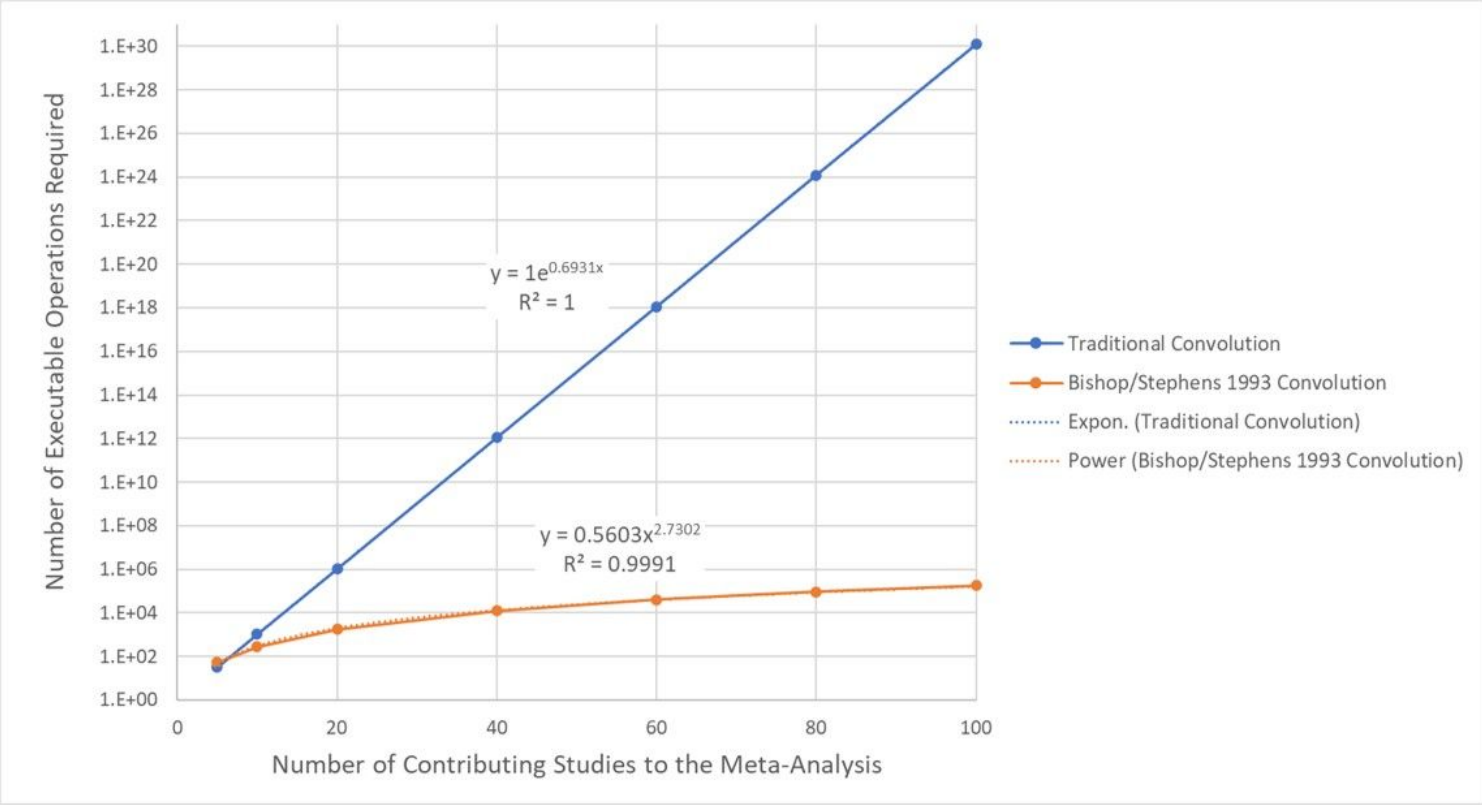

Figure 2

Estimated computer executable steps per Butler & Stephens vs. traditional convolution.

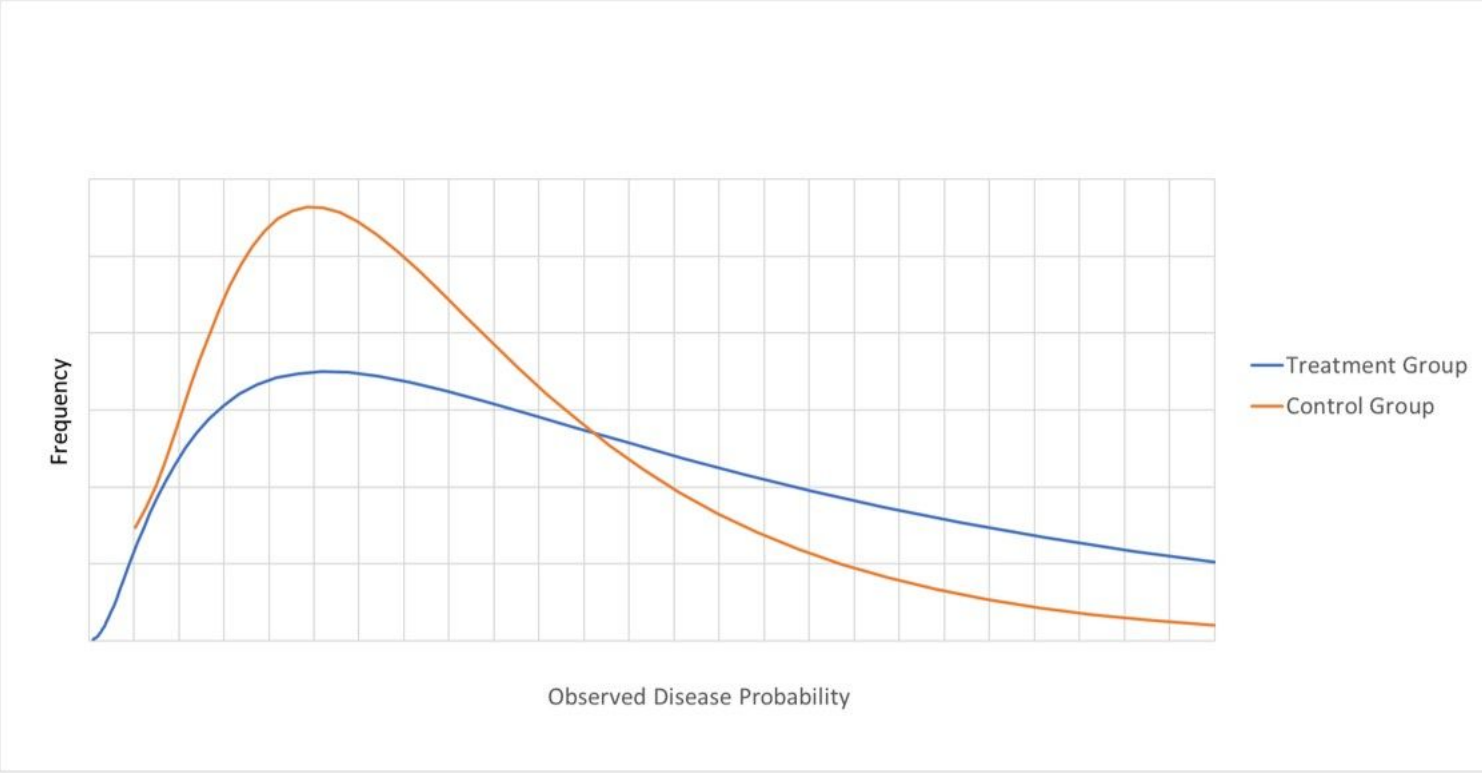

Figure 3

Demonstration of inappropriateness of comparing the PE and PC distributions to estimate Effect Size.

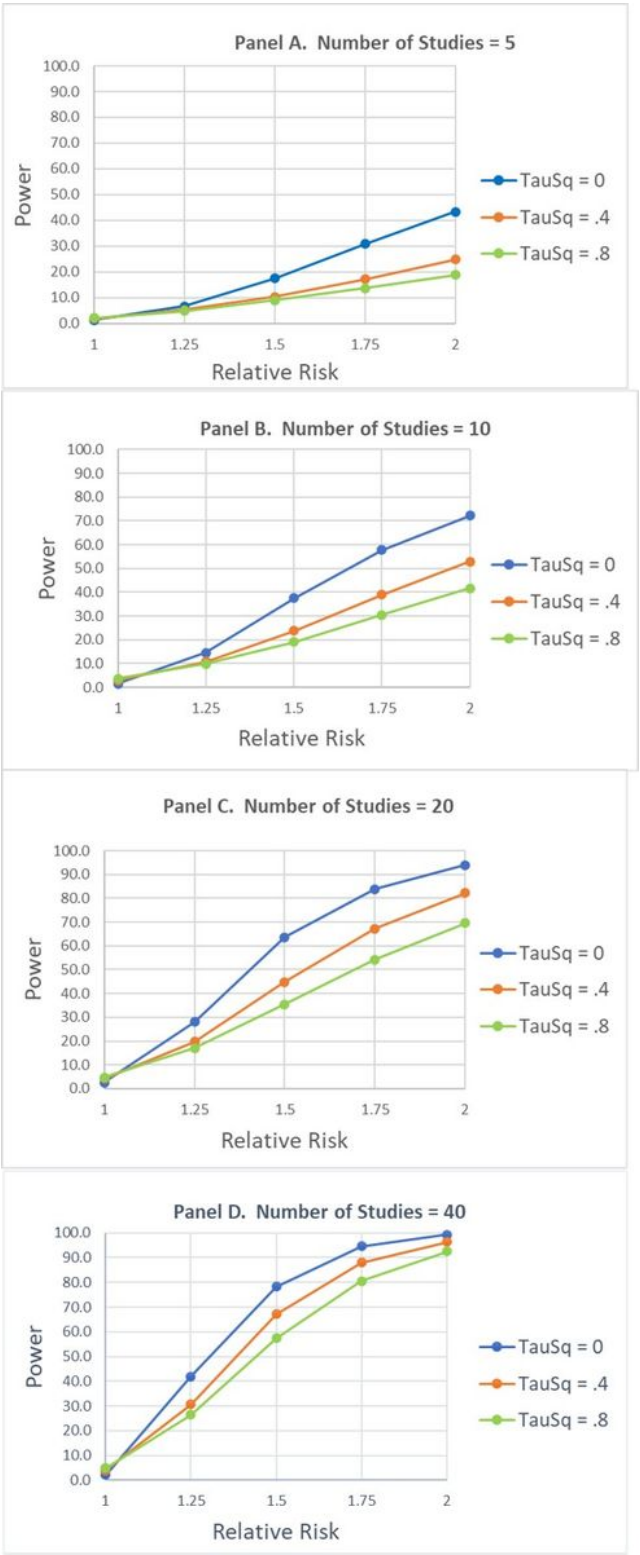

Figure 4

Power for number of studies, relative risk, and heterogeneity for EBT. A-D correspond to studies equal to 5, 10, 20 and 40.

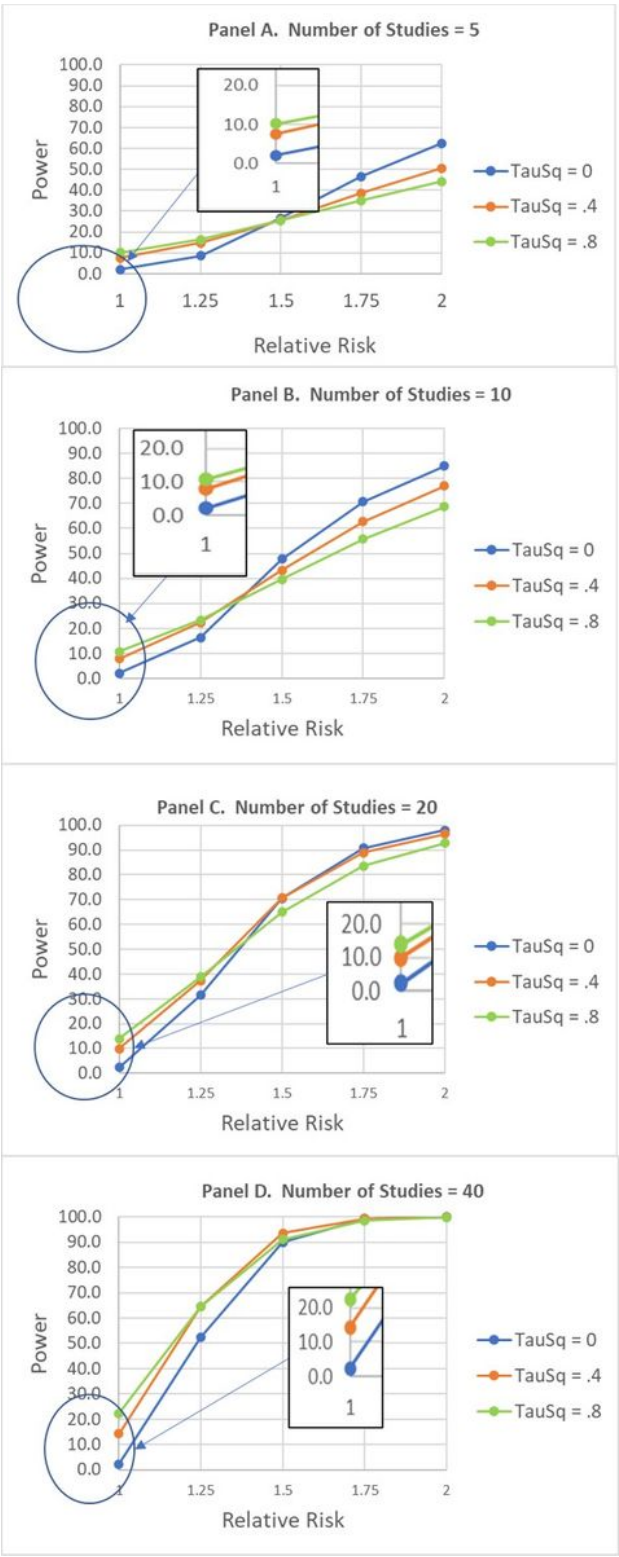

Figure 5

Power for number of studies, relative risk, and heterogeneity for DL. A-D correspond to studies equal to 5, 10, 20 and 40.

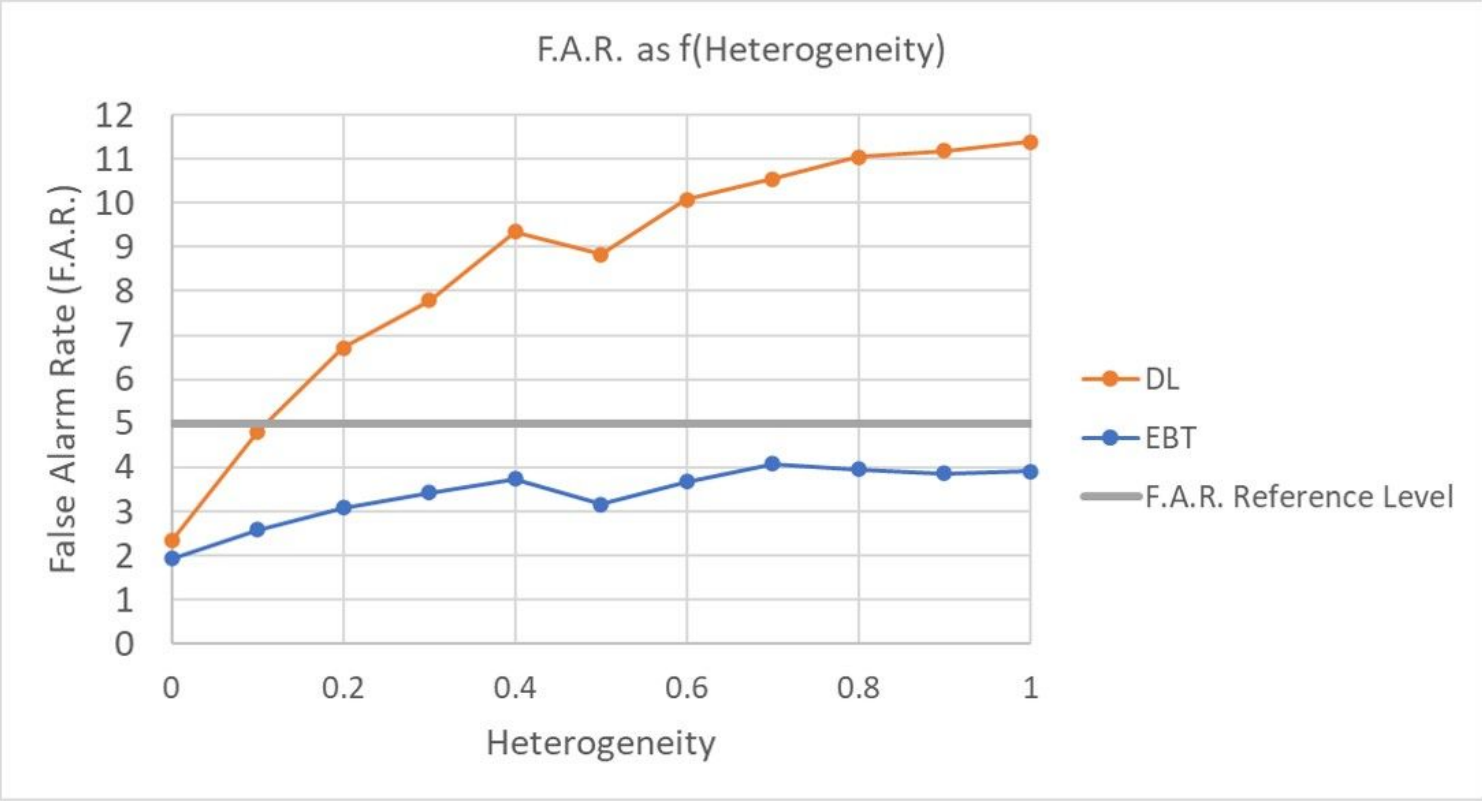

Figure 6

Type I error for EBT and DL methods as a function of heterogeneity.

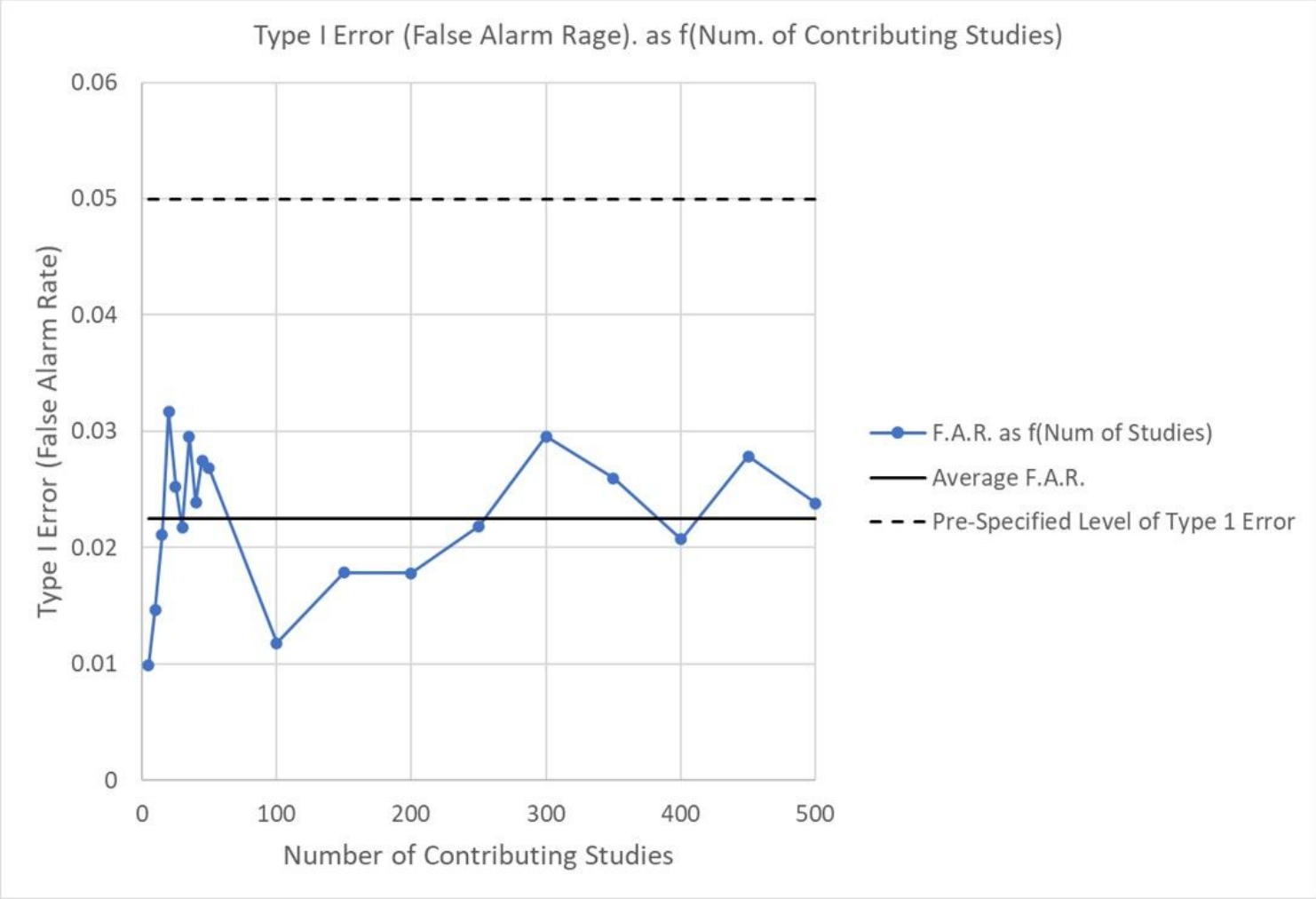

Figure 7

Interval overcoverage as a function of the number of contributing studies.

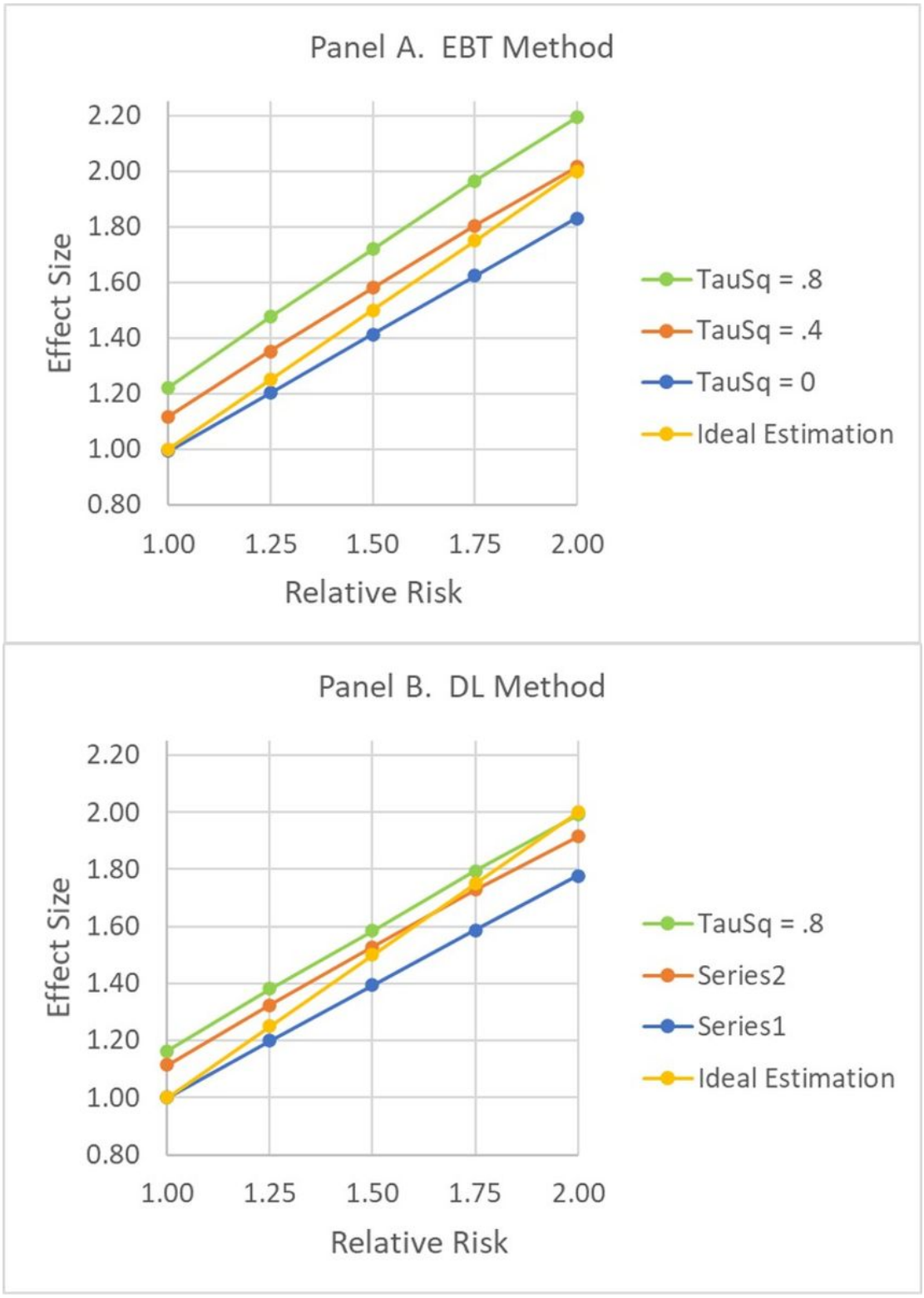

Figure 8

Effect Size as function of Relative Risk and heterogeneity. A & B correspond to the EBT and DL methods.

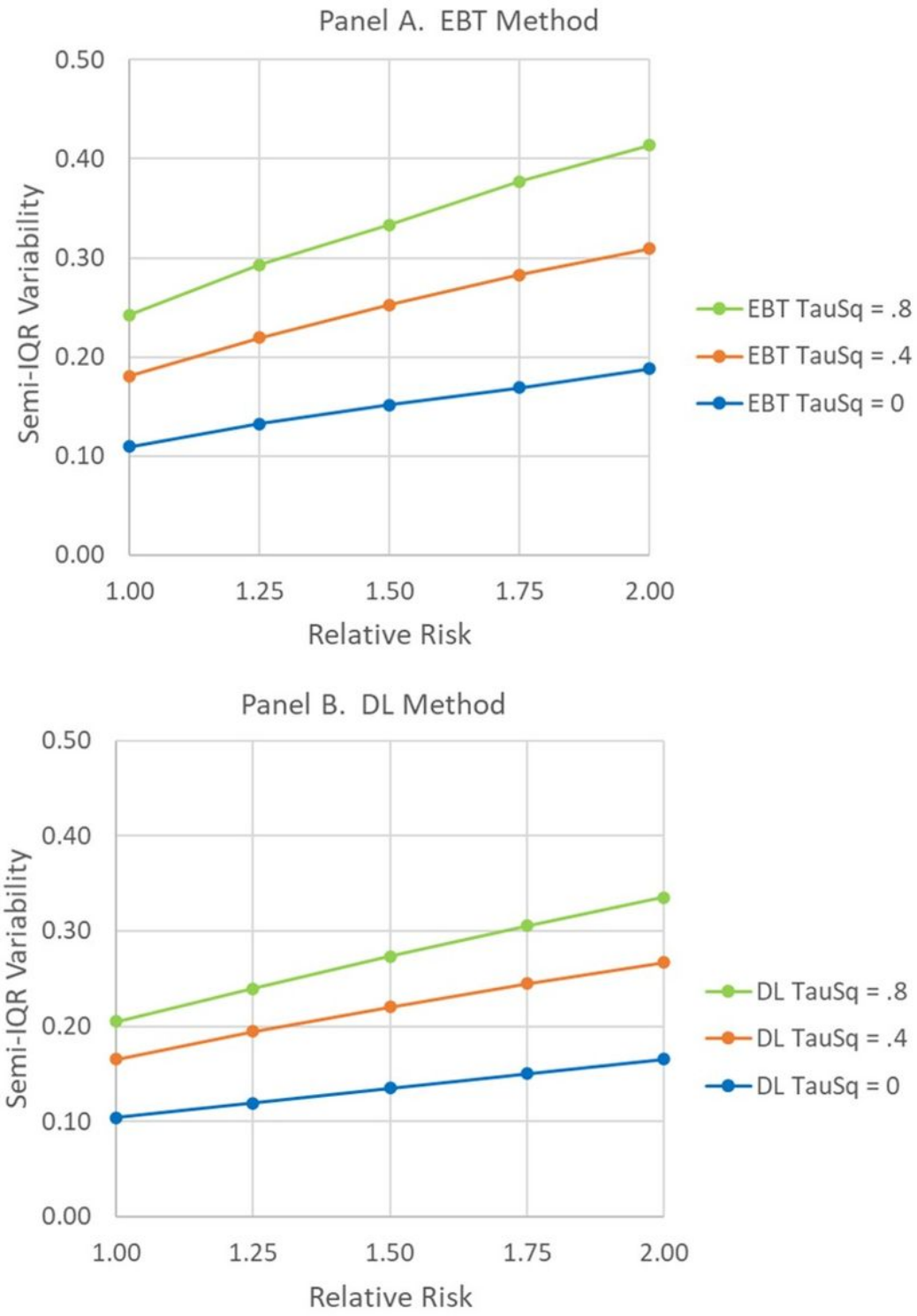

Figure 9

Semi-Interquartile Range as function of Relative Risk and heterogeneity. A & B correspond to the EBT and DL methods.