Published in final edited form as:

Stat Methods Med Res. 2019 January ; 28(1): 263–274. doi:10.1177/0962280217721246.

Meta-Analysis of Rare Binary Events in Treatment Groups with Unequal Variability

Lie Li [recent Ph.D. graduate] and Xinlei Wang* [Professor]

Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, P O Box 750332, Dallas, Texas 75275-0332

Abstract

Meta-analysis has been widely used to synthesize information from related studies to achieve reliable findings. However, in studies of rare events, the event counts are often low or even zero, and so standard meta-analysis methods such as fixed-effect models with continuity correction may cause substantial bias in estimation. Recently, Bhaumik et al.¹ developed a simple average (SA) estimator for the overall treatment effect based on a random-effects (RE) model. They proved that the SA method with the continuity correction factor 0.5 (SA_0.5) is the least biased for large samples, and showed via simulation that it has superior performance when compared with other commonly used estimators. However, the RE models used in previous work are restrictive because they all assume that the variability in the treatment group is equal to or always greater than that in the control group. Under a general framework that explicitly allows treatment groups with unequal variability but assumes no direction, we prove that SA_0.5 is still the least biased for large samples. Meanwhile, to account for a trade-off between the bias and variance in estimation, we consider the mean squared error (MSE) to assess estimation efficiency and show that SA_0.5 fails to minimize the MSE. Under a new RE model that accommodates groups with unequal variability, we thoroughly compare the performance of various methods for both large and small samples via simulation, and draw conclusions about when to use which method in terms of bias, MSE, type I error and confidence interval coverage. A data example of rosiglitazone meta-analysis is used to provide further comparison.

Keywords

bias; fixed effect; log odds ratio; mean squared error; random effects

1 Introduction

Meta-analysis, a useful statistical procedure to synthesize information across a collection of relevant studies, has been widely used to obtain reliable findings in medical research^{2,3}. In this paper, we focus on meta-analysis of rare binary events to evaluate treatment/intervention effects between two experimental conditions by combining data from multiple clinical trials.

* All correspondence should be addressed to Xinlei Wang (swang@smu.edu).

In many applications, the number of events of interest can be very low or even zero, mainly because the background incidence rate of the event is low and/or the sample size is not sufficiently large. For example, when the hormone replacement therapy is used to alleviate menopausal symptoms, it could cause serious diseases such as breast cancer and heart disease. These adverse events are rare but important to study⁴. It is evident that reliably detecting such adverse effects from an individual study is difficult. Thus, meta-analysis of rare binary events has been found to be useful and so received a lot of attention in the past.

Suppose there are K independent studies, in each of which there are treatment and control groups, and the outcome variable is binary. The whole data include K independent 2×2 tables. Under this simple setup, various meta-analysis approaches have been developed to estimate treatment effects on binary outcomes, based on either fixed-effect (FE) or random-effects (RE) models. FE models assume treatment effects to be the same across all studies. Popular FE meta-analysis methods of binary events include the Mantel-Haenszel (MH) method⁵, the empirical logit (EL) method⁶ and Peto's method⁷. On the contrary, RE models allow the treatment effects to vary from study to study. Among many, DerSimonian and Laird⁶ proposed the classical DerSimonian and Laird method (DSL) method with weights depending on the within and between-study variance estimates.

When the focus is narrowed down to *rare* binary events, early attention was largely paid to the FE meta-analysis approaches^{8;9}. For rare binary events, ordinary estimators of the (log) odds ratio and (log) relative risk are often undefined or unstable. A commonly used approach to deal with this problem is adding a continuity correction factor. Sweeting et al.⁸ found that, the MH method performed better than the inverse-variance weighted method (i.e., the EL method) for imbalanced group sizes under all kinds of continuity correction. Bradburn et al.⁹ compared the performance of twelve methods such as the MH method without zero-cell correction and logistic regression for pooling rare events. Both recommended the MH method with an appropriate continuity correction when analyzing sparse data with homogeneous treatment effects.

In practical situations, the FE assumption can be restrictive, where RE models are more appropriate. Through a specific rosiglitazone example involving rare binary events, Shuster et al.¹⁰ showed meta-analysis based on RE models was superior to that based on FE models. Cai et al.¹¹ proposed likelihood methods based on Poisson RE models. More recently, based on a RE model, Bhaumik et al.¹ developed a simple average (SA) estimator, which, as will be seen in (3) and (4), simply takes the average of the estimated treatment effects (i.e., log odd ratios with a positive continuity correction factor a) from individual studies. Bhaumik et al.¹ set $a = \frac{1}{2}$ so that the estimator is proved to be asymptotically unbiased. They further compared the SA estimator with $a = \frac{1}{2}$ (SA_0.5) to commonly used estimators including MH, EL and DSL via extensive simulation and a data example, and concluded SA_0.5 was the least biased, and performed better than the competitors. Due to its competitive performance as well as simplicity, SA_0.5 appears to be the most compelling method for meta-analysis of rare binary events, among many.

However, Bhaumik et al.¹ only used the bias measure to develop SA_0.5 and evaluate the performance of the different estimators, and so did not account for a trade-off between the bias and variance in estimation. In addition, they did not consider small sample sizes in their simulation where the asymptotic unbiasedness of SA_0.5 might not hold. More importantly, the RE models used in Bhaumik et al.¹ and other existing work¹² are restrictive because they all assume that the variability in the treatment group is equal to or always greater than that in the control group.

In this paper, we first propose a framework for RE models, which explicitly allows treatment groups with unequal variability but assumes no direction. Under this general framework, we prove that SA_0.5 is still the least biased for large samples. Meanwhile, to account for a trade-off between the bias and variance in estimation, we consider the mean squared error (MSE) to assess the estimation efficiency, and show that SA_0.5 cannot achieve the optimal MSE of its kind. Under a new RE model that accommodates groups with unequal variability, we thoroughly compare the bias and MSE of various methods for both large and small samples via simulation; and their performance on hypothesis testing and interval estimation is further compared using the type I error and coverage of 95% confidence intervals. A data example of rosiglitazone meta-analysis is used for comparison as well. Based on results of simulation and real data analysis, we provide practical guidelines for meta-analysis of rare binary events under the general framework.

2 A General Framework for RE models

Consider a meta-analysis including K independent studies. Throughout this paper, we assume K is a fixed constant. In the k th study, the treatment group includes n_{kt} subjects, and the control group includes n_{kc} subjects. Let x_{kt} and x_{kc} be the number of events of interest, and p_{kt} and p_{kc} be the probability of the event being observed in the treatment and control groups of the k th study, respectively, where $x_{kt} \sim \text{Binomial}(n_{kt}, p_{kt})$, and $x_{kc} \sim \text{Binomial}(n_{kc}, p_{kc})$. We denote the treatment effect of the k th study by θ_k , which is defined by the log odds ratio

$$\theta_k \equiv \ln \left(\frac{p_{kt}/q_{kt}}{p_{kc}/q_{kc}} \right),$$

where $q_{kt} = 1 - p_{kt}$ and $q_{kc} = 1 - p_{kc}$.

Our general framework for RE models requires a minimal set of assumptions:

1. The treatment effects θ_k 's are random and follow an unknown distribution $D(\theta, \tau_\theta^2)$ with $E(\theta_k) = \theta$ and $\text{Var}(\theta_k) = \tau_\theta^2$, where θ is the mean treatment effect over different studies, and τ_θ^2 reflects between-study heterogeneity.
2. For any two studies k and k' , (p_{kc}, p_{kt}) and $(p_{k'c}, p_{k't})$ are jointly independent.
3. The counts x_{kt} 's and x_{kc} 's are mutually independent given p_{kt} 's and p_{kc} 's.

Note that Assumption 2, combined with Assumption 1, implies that θ_k 's are identically and independently distributed (i.i.d).

We mention that the RE model considered in Bhaumik et al.¹ fits in the above general framework, which is given by

$$\text{logit}(p_{kc}) = \mu_k, \quad \text{logit}(p_{kt}) = \mu_k + \theta_k, \quad \theta_k \stackrel{iid}{\sim} N(\theta, \tau_\theta^2), \quad (1)$$

where any two components of $(\mu_1, \dots, \mu_K; \theta_1, \dots, \theta_K)$ are assumed to be independent. This model implicitly assumes that the variance of $\text{logit}(p_{kc})$ is not greater than the variance of $\text{logit}(p_{kt})$. Bhaumik et al.¹ further assumed $\mu_k \equiv u$ in their simulation evaluation, where μ is a constant. Also, the model used in Smith et al.¹² assumes the equality of the variances, which is given by

$$\text{logit}(p_{kc}) = \mu_k - \frac{\theta_k}{2}, \quad \text{logit}(p_{kt}) = \mu_k + \frac{\theta_k}{2}, \quad \theta_k \stackrel{iid}{\sim} N(\theta, \tau_\theta^2). \quad (2)$$

By contrast, our general framework does not imply the equality or any specific direction between the variances of the treatment and control groups.

3 Properties of SA_0.5 under the General Framework

As mentioned in the introduction, Bhaumik et al.¹ considered the SA estimator $\hat{\theta}_a$, namely

$$\hat{\theta}_a = \frac{1}{K} \sum_{k=1}^K \hat{\theta}_{ka}, \quad (3)$$

where $\hat{\theta}_{ka}$ is an estimator of the treatment effect θ_k given by

$$\hat{\theta}_{ka} = \ln \frac{x_{kt} + a}{n_{kt} - x_{kt} + a} - \ln \frac{x_{kc} + a}{n_{kc} - x_{kc} + a}, \quad (4)$$

with a constant $a \in (0, 1]$ for continuity correction. They proved that $\hat{\theta}_a$ is asymptotically unbiased when $a = \frac{1}{2}$ under the specific RE model (1).

Below we examine properties of the SA estimator under the RE framework proposed in Section 2. We show that the asymptotic unbiasedness of SA_0.5 still holds; but SA_0.5 fails to achieve the optimality in terms of minimizing the MSE of $\hat{\theta}_a$.

3.1 Estimation bias

Gart et al.¹³ showed that for any binomial random variable x , based on a sample of size m with parameter p , the following equation holds

$$E\left(\ln \frac{x+a}{m-x+a}\right) = \ln \frac{p}{q} + \frac{(\frac{1}{2}-a)(p-q)}{mpq} + \frac{(p-q)}{12(mpq)^2} \{6a^2 - 12(1-pq)a + (5-6pq)\} \quad (5)$$

$$+ O(m^{-3}),$$

where $q = 1-p$. Based on (5), we have the following conditional means for $k = 1, \dots, K$:

$$E\left(\ln \frac{x_{kt}+a}{n_{kt}-x_{kt}+a} | p_{kt}\right) = \ln \frac{p_{kt}}{q_{kt}} + \frac{(\frac{1}{2}-a)(p_{kt}-q_{kt})}{n_{kt}p_{kt}q_{kt}} + O(n_{kt}^{-2});$$

$$E\left(\ln \frac{x_{kc}+a}{n_{kc}-x_{kc}+a} | p_{kc}\right) = \ln \frac{p_{kc}}{q_{kc}} + \frac{(\frac{1}{2}-a)(p_{kc}-q_{kc})}{n_{kc}p_{kc}q_{kc}} + O(n_{kc}^{-2}).$$

Let $n \equiv \min[(n_{kt}, n_{kc})_{k=1}^K]$ be the overall minimum number of subjects, and let

$\mathbf{p} = (p_{kt}, p_{kc})_{k=1}^K$. Under our general framework, we can have

$$E\left(\hat{\theta}_a | \mathbf{p}\right) = E\left(\frac{1}{K} \sum_{k=1}^K \hat{\theta}_{ka} | \mathbf{p}\right) = \frac{1}{K} \sum_{k=1}^K \theta_k + \frac{\frac{1}{2}-a}{K} \sum_{k=1}^K \frac{p_{kt}-q_{kt}}{n_{kt}p_{kt}q_{kt}} - \frac{\frac{1}{2}-a}{K} \sum_{k=1}^K \frac{p_{kc}-q_{kc}}{n_{kc}p_{kc}q_{kc}} + O(n^{-2}).$$

So

$$\text{Bias}\left(\hat{\theta}_a\right) = E\left(\hat{\theta}_a\right) - \theta = E\left[E\left(\hat{\theta}_a | \mathbf{p}\right)\right] - \theta = \frac{\frac{1}{2}-a}{K} \left[\sum_{k=1}^K E p_{kt} \left(\frac{p_{kt}-q_{kt}}{n_{kt}p_{kt}q_{kt}} \right) - \sum_{k=1}^K E p_{kc} \left(\frac{p_{kc}-q_{kc}}{n_{kc}p_{kc}q_{kc}} \right) \right] + O(n^{-2}).$$

That means, if n is large enough so that the higher order term $O(n^{-2})$ can be ignored, $\text{SA}_{0.5}$ is an asymptotically unbiased estimator, i.e., $\lim_{n \rightarrow +\infty} \text{Bias}\left(\hat{\theta}_a\right) = 0$ under our general

framework.

3.2 Mean squared error

To derive the MSE of the SA estimator $\hat{\theta}_a$ under our general framework, we first consider the variance of $\hat{\theta}_a$. For any binomial random variable x , based on a sample of size m with parameter p , we can show

$$V\left(\ln\frac{x+a}{m-x+a}\right) = \frac{1}{mpq} + \frac{(\frac{3}{2}-2a) + (4a-4)pq}{(mpq)^2} + O(m^{-3}),$$

where $V(f(x))$ denotes the variance of $f(x)$. Based on the above result, we have the following conditional variance:

$$V(\hat{\theta}_a | \mathbf{p}) = \frac{1}{K^2} \left[\sum_{k=1}^K \frac{1}{n_{kt} p_{kt} q_{kt}} + \sum_{k=1}^K \frac{1}{n_{kc} p_{kc} q_{kc}} + O(n^{-2}) \right].$$

So

$$\begin{aligned} V(\hat{\theta}_a) &= E_{\mathbf{p}} \left[V(\hat{\theta}_a | \mathbf{p}) \right] + V_{\mathbf{p}} \left[E(\hat{\theta}_a | \mathbf{p}) \right] = E_{\mathbf{p}} \left[\frac{1}{K^2} \left[\sum_{k=1}^K \frac{1}{n_{kt} p_{kt} q_{kt}} + \sum_{k=1}^K \frac{1}{n_{kc} p_{kc} q_{kc}} \right] \right] \\ &+ V_{\mathbf{p}} \left[\frac{1}{K} \sum_{k=1}^K \theta_k + \frac{\frac{1}{2}-a}{K} \sum_{k=1}^K \frac{p_{kt}-q_{kt}}{n_{kt} p_{kt} q_{kt}} - \frac{\frac{1}{2}-a}{K} \sum_{k=1}^K \frac{p_{kc}-q_{kc}}{n_{kc} p_{kc} q_{kc}} \right] + O(n^{-2}). \\ &= \frac{1}{K^2} \left[\sum_{k=1}^K E_{p_{kt}} \left(\frac{1}{n_{kt} p_{kt} q_{kt}} \right) + \sum_{k=1}^K E_{p_{kc}} \left(\frac{1}{n_{kc} p_{kc} q_{kc}} \right) \right] + \frac{1}{K^2} \sum_{k=1}^K [V_{\mathbf{p}}(\theta_k) + \\ &+ (1-2a) \text{Cov} \left(\theta_k, \frac{p_{kt}-q_{kt}}{n_{kt} p_{kt} q_{kt}} \right) - (1-2a) \text{Cov} \left(\theta_k, \frac{p_{kc}-q_{kc}}{n_{kc} p_{kc} q_{kc}} \right)] + O(n^{-2}). \end{aligned}$$

Combined with the bias of the SA estimator, we can get:

$$\begin{aligned} \text{MSE}(\hat{\theta}_a, \theta) &= \frac{1}{K^2} \left[\sum_{k=1}^K E_{p_{kt}} \left(\frac{1}{n_{kt} p_{kt} q_{kt}} \right) + \sum_{k=1}^K E_{p_{kc}} \left(\frac{1}{n_{kc} p_{kc} q_{kc}} \right) \right] \\ &+ \frac{1}{K^2} \left[K \cdot \tau_{\theta}^2 + (1-2a) \sum_{k=1}^K \left[\text{Cov} \left(\theta_k, \frac{p_{kt}-q_{kt}}{n_{kt} p_{kt} q_{kt}} \right) - \text{Cov} \left(\theta_k, \frac{p_{kc}-q_{kc}}{n_{kc} p_{kc} q_{kc}} \right) \right] \right] + O(n^{-2}). \end{aligned}$$

As $n \rightarrow +\infty$, we can ignore the higher order term $O(n^{-2})$. Thus, to minimize the MSE of the SA estimator for large n , the best choice of a in $(0, 1]$ is given by

$$a = \begin{cases} \varepsilon & \text{if } \sum_{k=1}^K \left[\text{Cov} \left(\theta_k, \frac{p_{kt}-q_{kt}}{n_{kt} p_{kt} q_{kt}} \right) - \text{Cov} \left(\theta_k, \frac{p_{kc}-q_{kc}}{n_{kc} p_{kc} q_{kc}} \right) \right] < 0, \\ 1 & \text{if } \sum_{k=1}^K \left[\text{Cov} \left(\theta_k, \frac{p_{kt}-q_{kt}}{n_{kt} p_{kt} q_{kt}} \right) - \text{Cov} \left(\theta_k, \frac{p_{kc}-q_{kc}}{n_{kc} p_{kc} q_{kc}} \right) \right] > 0, \\ \text{any number in } (0, 1] & \text{if } \sum_{k=1}^K \left[\text{Cov} \left(\theta_k, \frac{p_{kt}-q_{kt}}{n_{kt} p_{kt} q_{kt}} \right) - \text{Cov} \left(\theta_k, \frac{p_{kc}-q_{kc}}{n_{kc} p_{kc} q_{kc}} \right) \right] = 0, \end{cases} \quad (6)$$

where ε is an arbitrarily small positive number. Obviously, the SA estimator with $a = \frac{1}{2}$ cannot achieve the minimum MSE for large n except for the case

$\sum_{k=1}^K \text{Cov}(\theta_k, \frac{p_{kt} - q_{kt}}{n_{kt} p_{kt} q_{kt}}) = \sum_{k=1}^K \text{Cov}(\theta_k, \frac{p_{kc} - q_{kc}}{n_{kc} p_{kc} q_{kc}})$, where the choice of a does not matter at all (i.e., any number in $(0, 1]$ would minimize the approximate MSE).

4 A Simulation-based Comparative Study

Under our general RE framework, we conducted simulation to address the following research questions for meta-analysis of rare binary events; (i) whether the superiority of SA_0.5, as shown in Bhaumik et al.¹, remains valid; and (ii) how various existing methods perform in terms of both bias and MSE, especially for small sample sizes. To do so, we first proposed a new RE model that accommodates groups with unequal variability. Using data simulated from this model, we compared the performance of various methods in estimating the overall treatment effect θ for both large and small samples. Here, a large-sample case means that the $O(n^{-2})$ term can be ignored compared to $O(n^{-1})$; otherwise, we call it a small-sample case. To evaluate the performance of each estimator, we examined both bias and MSE.

As pointed out by one of the reviewers, the empirical performance of these methods on hypothesis testing and interval estimation is important as well. Thus, we also examined the type I error under the null hypothesis $H_0: \theta = 0$ and the coverage of the corresponding 95% confidence interval (CI) for each method.

4.1 A new RE model allowing for unequal group variability

We consider the following RE model:

$$\text{logit}(p_{kc}) = \mu_k - w \cdot \theta_k; \quad (7)$$

$$\text{logit}(p_{kt}) = \mu_k + (1 - w) \cdot \theta_k,$$

where $\mu_k \stackrel{iid}{\sim} N(\mu, \tau_\mu^2)$, $\theta_k \stackrel{iid}{\sim} N(\theta, \tau_\theta^2)$ and w is a constant which belongs to $[0, 1]$. All μ_k s and θ_k s are assumed to be independent. Based on the RE model, we have:

$$V[\text{logit}(p_{kc})] = \tau_\mu^2 + w^2 \cdot \tau_\theta^2;$$

$$V[\text{logit}(p_{kt})] = \tau_\mu^2 + (1 - w)^2 \cdot \tau_\theta^2;$$

$$\text{Cov}[\text{logit}(p_{kc}), \text{logit}(p_{kt})] = \tau_\mu^2 - w(1-w) \cdot \tau_\theta^2.$$

So we can control the sign of the difference in the variance of the logit between the control and treatment groups by the constant w : $0 \leq w < \frac{1}{2}$ means the variance of the logit in the treatment is greater than that in the control; $w = \frac{1}{2}$ means equal variances; and $\frac{1}{2} < w \leq 1$ means the variance in the treatment is less than that in the control. Meanwhile, the covariance between $\text{logit}(p_{kc})$ and $\text{logit}(p_{kt})$ also depends on the constant w , and it can achieve the maximum value τ_μ^2 when $w = 0$ or 1 , and achieve the minimum value $\tau_\mu^2 - \frac{\tau_\theta^2}{4}$ when $w = \frac{1}{2}$.

This new RE model fits in the general framework outlined in Section 2; and it is more flexible than the RE model (1) in Bhaumik et al.¹ because it allows for unequal group variability without assuming any specific direction. Further, when $w = \frac{1}{2}$, (7) is reduced to the model (2) used in Smith et al.¹².

4.2 Methods to be compared

In our comparison, we first include all the four methods evaluated in Bhaumik et al.¹'s simulation: SA_0.5, MH, EL and DSL. According to (6), to achieve smaller MSE, we include the SA estimator with other correction factors $a = 1$ and $a = 0.25$, denoted by SA_1 and SA_0.25, respectively. Note that $a = 0.25$ was recommended by Hitchcock¹⁴ and Gart and Zweifel¹⁵ for estimating the log odds ratio in an individual study, which was proposed based on the minimum logit χ^2 method. Also included is another popular estimator based on generalized linear mixed models (GLMM¹⁶), which has been reported to work well in a wide range of applications including rare binary events^{17;18}. In addition, we consider an integrative median unbiased estimator, labeled iMUE, which is described below.

For any binomial random variable x , based on a sample of size m with parameter p , Parzen et al.¹⁹ proposed a median unbiased estimator (MUE) $\tilde{p}$ to estimate p for rare events:

$$\tilde{p}(x, m) = \begin{cases} \frac{1 - 0.5^{\frac{1}{m}}}{2} & \text{if } x = 0; \\ \frac{F^{-1}(0.5 | \alpha = x, \beta = m - x + 1) + F^{-1}(0.5 | \alpha = x + 1, \beta = m - x + 2)}{2} & \text{if } 0 < x < m; \\ \frac{1 + 0.5^{\frac{1}{m}}}{2} & \text{if } x = m. \end{cases}$$

where $F(\cdot | \alpha, \beta)$ is the cumulative distribution function of a beta distribution with parameters α and β . Based on the MUE estimator of p , $\hat{\theta}_{MUE,k}$ can be defined as the estimator of the treatment effect θ_k in the k th study:

$$\hat{\theta}_{MUE.k} = \ln \frac{\tilde{p}_{kt}}{1 - \tilde{p}_{kt}} - \ln \frac{\tilde{p}_{kc}}{1 - \tilde{p}_{kc}},$$

where $\tilde{p}_{kt} = \tilde{p}(x_{kt}, n_{kt})$ is the MUE of p_{kt} and $\tilde{p}_{kc} = \tilde{p}(x_{kc}, n_{kc})$ is the MUE of p_{kc} . Note that $\hat{\theta}_{MUE.k}$ always exists. In our meta-analysis of rare binary events, it is natural to define iMUE by $\hat{\theta}_{iMUE} = \sum_{k=1}^K \hat{\theta}_{MUE.k} / K$.

4.3 Simulation settings

We consider meta-analysis of $K \in \{10, 20, 50\}$ independent studies of rare binary events. We generated the responses x_{kc} and x_{kt} from $\text{Binomial}(n_{kc}, p_{kc})$ and $\text{Binomial}(n_{kt}, p_{kt})$, respectively, for $k = 1, \dots, K$, where $n_{kt} = R_k \cdot n_{kc}$ and (p_{kc}, p_{kt}) satisfies (7). We set $\theta \in \{-1.5, -1.0, \dots, 1.5\}$, $\mu = -5$ (i.e., 0.0067 in the probability magnitude), and $\tau_\mu^2 = \tau_\theta^2 = 0.5$. We further set $w \in \{0, 0.5, 1.0\}$, representing larger/equal/smaller variability in the treatment, compared to that in the control. To allow different allocation ratios across studies, we set $\log_2 R_k \sim N(\log_2 R, \sigma_R^2)$, where $R \in \{1, 2, 4\}$ and $\sigma_R^2 = 0.5$. Note that the value of σ_R^2 is selected to approximately equal the sample standard deviation of the sample ratios in real data analysis in Section 5. Here, we restrict our attention to $R > 1$, because the two settings (R, θ, w) and $(1/R, -\theta, 1 - w)$ are the same except for group label switching, and so have exactly the same performance.

We generated n_{kc} s from $\text{Uniform}[2000, 3000]$ to examine large-sample performance, and from $\text{Uniform}[20, 1000]$ to examine small-sample performance, and rounded random draws to the nearest integers. Note that the above ranges for large and small samples were chosen based on preliminary simulation outcomes as well as theoretical arguments. Based on results in Gart et al.¹³, we know that the $O(n^{-2})$ term can be ignored when $\min(n_{kc}p_{kc}, n_{kt}p_{kt})_{k=1}^K$ is not small, and at least it should be larger than 1 since $O(n^{-2})$ may be larger than $O(n^{-1})$ if it is not. Thus, for small sample sizes, we choose $[20, 1000]$ so that the empirical means of $\min(n_{kc}p_{kc}, n_{kt}p_{kt})_{k=1}^K$ in all our settings are below one. Besides, in real meta-analysis datasets^{8;20;9} that are treated as small-sample cases, most sample sizes of individual studies are small, but a few sample sizes are around 1000. That's why we choose a large upper bound 1000 for small-sample sizes. For large sample sizes, we find that further increasing either the upper or lower bound would lead to exactly the same conclusions as reported below.

4.4 Results based on bias and MSE

For each setting, we simulated 1000 datasets to compute the empirical bias and MSE of all the methods. We further calculated log₂ relative efficiency (Re) based on MSE, where we set GLMM as the benchmark. For a specific method, the Re is defined as the MSE of GLMM divided by the MSE of the method, where log₂ Re > 0 means that it is more efficient than GLMM.

We use nested loop plots for visualization, which can efficiently present results from a large number of simulation scenarios²¹. We first need to figure out a reasonable lexicographical order of all $P = 4$ parameters, including the mean treatment effect θ , the constant w , the sample size ratio R between the treatment and control, and the number of studies K , among the $P! = 24$ possible orders. We find that the influences of θ and w are much larger than those of R and K . So the lexicographical order we choose is to sort (1) θ , (2) w , (3) R , and (4) K .

Large-sample results—In the Supplementary Material, Figure S1 reports the bias and the MSE for different methods over all the 189 simulation scenarios with large sample sizes. As the smallest unit of these figures shows, the no. of studies K does not affect the choice of methods in most scenarios. Thus, we use $K = 50$ as an example to compare the bias and MSE of the eight different methods.

Figure 1(a) shows bias comparison using results from the 63 simulation scenarios for $K = 50$ and large sample sizes. We find that the effect of the ratio R on comparing bias of different methods is marginal. Overall, SA_0.5 works the best in terms of bias, which is consistent with our theoretical result that SA_0.5 is asymptotically unbiased under the general framework. The second and third best methods are iMUE and SA_0.25, and both of them yield estimates that are only slightly biased. They are followed by GLMM and SA_1, which form the middle group. The bottom group contains DSL, EL and MH, which are plotted separately from the others for better visualization, whose bias values are typically much larger. Among them, MH severely overestimates θ for cases of $w = 0$ and underestimates θ for $w = 1$; but it performs well for $w = 0.5$ (i.e., equal group variability). DSL underestimates θ when $\theta > 0$ and overestimates θ when $\theta < 0$ except for cases with $\theta > 0$ and $w = 0$ or cases with $\theta < 0$ and $w = 1$, where it is nearly unbiased.

In Figure 1(b), we compare MSE for $K = 50$ and large sample sizes based on $\log_2 \text{Re}$, where the yellow horizontal line at $y = 0$ represents GLMM. We find that EL and MH are the worst two methods, typically having much larger MSE than the other methods. DSL can be bad for cases with large positive θ and $w = 1$ or cases with large negative θ and $w = 0$. Overall, SA_0.5 seems to be the most efficient estimator because it is the best in some settings (e.g., $\theta > 0$ and $w = 1$), and is always among the top group otherwise. In addition, the performance of iMUE and SA_0.25 is pretty close to the top in nearly all the settings; and SA_1 is the most efficient when $\theta = 0$ and -0.5 .

Small-sample results—Bias and MSE results for all the 189 scenarios with small sample sizes are shown in Figure S2 of the Supplementary Material. Again, K seems not to affect the choice of the methods, and so we use the results for $K = 50$ for comparison.

As shown in Figure 2(a), the ratio R has little impact on the relative performance of the different methods based on bias. Thus, we summarize the least biased method(s) by w and θ in Table 1. We note that SA_0.25 performs consistently well in all the settings. Although it is not chosen as the best for settings with $w = 0$, its performance is quite close to the best method(s) and much better than some of the other methods.

Figure 2(b) displays MSE comparison for small sample sizes based on $\log_2 \text{Re}$. Unlike bias comparison, we find that R plays a role in evaluating the different methods. So in Table 2, we report the most efficient method(s) by θ , w and R , where the method(s) in the parenthesis is the second best and has performance very close to the best one.

4.5 Results based on type I error and coverage of 95% CIs

We examined the type I error of each method for testing the null hypothesis $H_0: \theta = 0$ at the significance level 0.05 by simulating data from various null settings (with $\theta = 0$). As in Bhaumik et al.¹, we used the following statistic for testing $H_0: T = \hat{\theta}/\widehat{\text{SE}}(\hat{\theta})$, where $\hat{\theta}$ stands for an estimator of θ , and $\widehat{\text{SE}}(\hat{\theta})$ is the estimated standard error (SE) of $\hat{\theta}$. We also used the methods described in Bhaumik et al.¹ to estimate the SEs of SA_0.5, DSL, EL and MH. For SA_0.25, SA_1 and iMUE, we used the delete- d jackknife method²² with $d = 2$ to estimate the SEs. For GLMM, we used the output for the SE from an R package named *lme4*²³. We then performed a t-test with the number of degrees of freedom $df = K - 1$ for all the methods, except for SA_0.5, for which we performed a z-test instead. This is because when K is not large, the t-test works better than the z-test for all the methods but SA_0.5. The type I error rate of each method under each setting was empirically computed as the proportion of rejecting H_0 among 2000 replicates.

To further evaluate the coverage rate of 95% CIs for each method, we computed the proportion of the CIs that contain the true θ value among 2000 replicates under each setting. The CIs of SA_0.5 were computed using the z-test, i.e., $\hat{\theta} \pm 1.96\widehat{\text{SE}}(\hat{\theta})$; and the CIs of the other methods were computed using the t-test, i.e., $\hat{\theta} \pm t_{0.025, K-1}\widehat{\text{SE}}(\hat{\theta})$, where $t_{0.025, K-1}$ is the critical value of the t distribution with $df = K - 1$.

Due to the limited space, tables and figures for results based on type I error and coverage of 95% CIs are put in Sections S2 and S3 of Supplementary Material, along with a detailed description and a thorough discussion. Below we provide a brief summary about the performance for large and small samples, respectively.

Large-sample results—As shown in Table S1 of Supplemental Material, SA_0.5, SA_0.25, SA_1 and iMUE perform better than the other methods for large sample sizes in terms of the type I error; and their rates are close to the nominal level 0.05. As shown in Figures S3–S5, SA_0.5, SA_0.25 and iMUE form the top performing group in terms of the 95% CI coverage, and they provide coverage rates close to 95%. The performance of EL and MH is consistently poor in terms of both measures.

Small-sample results—From Table S2 and Figures S6–S8, it appears that SA_0.25 is the best and iMUE is the second best, followed by SA_0.5, for small sample sizes in terms of both type I error and CI coverage; and they can provide rates pretty close to the targeted level (0.05 or 95%) in most settings. Also, SA_1, DSL, EL and MH do not perform well in general, among which EL and MH seem to be the worst two; and GLMM does not perform well when $w = 1$.

5 Example: Rosiglitazone Meta-Analysis

Rosiglitazone was a widely used drug to treat hyperglycemia in patients with type 2 diabetes mellitus. However, the use of rosiglitazone therapy has decreased dramatically as it might increase the risks of myocardial infarction (MI) and cardiovascular death (CVD). In order to evaluate potential adverse effects of rosiglitazone, Nissen and Wolski²⁰ conducted a meta-analysis of 48 randomized trials (the data depicted in Table S1 of the Supplementary Material), including the 42 studies they analyzed as well as 6 additional studies excluded from their analysis because of all zero counts for both MI and CVD events. In our meta-analysis, we applied the eight methods, namely, SA_0.25, SA_0.5, SA_1, iMUE, GLMM, DSL, EL and MH, to synthesize all the 48 trials, which include 29,818 patients in total. In the rosiglitazone groups, there are 86 MI and 39 CVD cases reported while in the control groups, there are 72 MI and 22 CVD cases.

5.1 MI data

Table 3 summarizes results from the different methods for the MI data, including the estimated treatment effect $\hat{\theta}$, estimated odds ratio (OR), and 95% CI of OR. We used the same methods described in Section 4.4 to compute the CI of each method. From Table 3, we find that all the methods except for SA_1 report positive estimates of the mean treatment effect θ , and all except for MH provide CIs of OR that include 1, and so conclude that the effect of rosiglitazone on MI is not statistically significant.

To understand the above results better, we conducted an exploratory analysis of the MI data. We estimated the log odds for each group of study k using $\pi_{kt} = \ln(x_{kt} + 0.5)/(n_{kt} - x_{kt} + 0.5)$ and $\pi_{kc} = \ln(x_{kc} + 0.5)/(n_{kc} - x_{kc} + 0.5)$ and find the sample variance of π_{kt} s is 0.947 and the sample variance of π_{kc} s is 0.703. We then conducted the robust Brown-Forsythe Levene-type test for equal variances²⁵, which does not require the normality assumption. The test does not reject the null at the significance level $\alpha = 0.05$. We also computed the average sample size ratio between the treatment and control and find $\bar{R} = 1.64$. Thus, we believe MI data may fit in one of the simulation settings for small sample sizes with $w = 0.5$, $R = 2$, $\theta = 0$ or 0.5 and $K = 50$. Note that for the adverse effect, θ should be nonnegative since a negative θ actually means the effect is beneficial instead, which is difficult to interpret.

According to Figure 2(a), SA_1 consistently underestimate θ when $\theta = 0$, which might explain its negative estimate for the MI data. Based on Table 1, regardless of the θ values, SA_0.25 and MH are the least biased for $w = 0.5$. However, there is a small chance that $w < 0.5$. Although the Levene's test does not reject equal variances, the treatment appears to have heavier tails and so a larger variance than the control, as shown in the left panel of Figure 3, which plots the empirical density curves of the estimated log odds for MI data. As MH severely overestimates θ when $w < 0.5$, we leave out MH by noting MH does give the largest positive estimate among all; and we select SA_0.25 as a much safer choice because SA_0.25 performs consistently well in terms of bias regardless of the θ , w values. As to MSE, for $R = 2$ and $\theta = 0, 0.5$, iMUE and DSL are the most efficient for $w = 0.5$ and $w = 0$, respectively. Besides, SA_0.25 and iMUE are the best two methods in terms of type I error and 95% CI coverage for these settings. The three estimators, SA_0.25, iMUE and DSL, all

give small positive estimates for θ , with CIs covering 1, which supports the conclusion that the (small positive) effect of rosiglitazone on MI is not statistically significant.

5.2 CVD data

As shown in Table 4 for the CVD data, SA_0.5 and SA_1 give slightly negative estimates while the other methods report positive estimates; and all the CIs cover 1, which indicates that all the different methods unanimously conclude the effect of rosiglitazone on CVD is not statistically significant.

The right panel of Figure 3 shows that the empirical density curve of the log odds for the treatment has much wider tails than that for the control. The sample variance of $\pi_{k\delta}$ is 0.823 and that of $\pi_{k\epsilon}$ is 0.474. The Brown-Forsythe test for equal variances returns a p-value 0.040, suggesting that the variance of the treatment is larger. Thus, we believe CVD data may fit in one of the simulation settings for small sample sizes with $w = 0$, $R = 2$, $\theta = 0$ or 0.5 and $K = 50$. For these settings, Table 1 suggests that GLMM and iMUE are the least biased; Table 2 suggests that DSL is the most efficient; and Table S2 and Figure S6 suggest that SA_0.25, iMUE, GLMM and DSL work well in terms of the type I error and 95% CI coverage. All the four methods report positive estimates, and iMUE, DSL and SA_0.25 all suggest that the effect is pretty close to zero. Thus, we believe there is no strong evidence to support the existence of the rosiglitazone effect on CVD.

6 Discussion

In the context of meta-analysis of rare binary events, we have proposed a general framework for RE models, which requires minimal assumptions. We have conducted a comprehensive simulation study to compare the performance of various estimators in terms of bias, MSE, type I error and CI coverage. We find that for large samples, SA_0.5 performs very well: it is the least biased in nearly all the settings examined; it is the most efficient in some settings and always among the most efficient group otherwise; and it has good performance on the type I error and coverage rate as well. We also find iMUE and SA_0.25 perform quite well for large samples in terms of all the four measures. In addition, SA_1 can achieve the minimum MSE in some of the settings, which confirms our theoretical results in Section 3.2.

For small samples, the variability direction (i.e., which group has larger variability) plays an important role in selecting the least biased method(s), as indicated by Table 1. In situations when there is no or vague information about the direction, we recommend SA_0.25 because it is always among the least biased group if it is not the best. As to MSE, the choice of the most efficient methods depends on the variability direction and the sample size ratio between the treatment and control, as well as the effect size, as summarized in Table 2. When such information is not available or partially available, we should select a safe choice (e.g., SA_0.25) and avoid methods (e.g., EL, MH, SA_1) that can perform well only in some specific conditions but have very large MSE in neighboring conditions. In addition, our finding that SA_0.25 and iMUE perform well in terms of type I error and CI coverage for small sample sizes can also help us make a wise decision.

Through a data example involving rosiglitazone meta-analysis, we have illustrated how to choose appropriate methods according to the above guidelines, combined with exploratory analysis. It is worth mentioning that our conclusions about the rosiglitazone effect on MI and CVD support the action that the U.S. Food and Drug Administration (FDA) lifted its earlier restrictions on rosiglitazone in 2013.

Our finding that the variability direction has a large impact on the small-sample performance may indicate a future research direction. For the purpose of simulation, we have proposed a flexible RE model (7). It would be interesting to develop novel statistical methods for estimating and testing the treatment effect based on this RE model, which can formally account for unequal group variability to improve the inference for small samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. 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): 555–567. [PubMed: 23734068]
2. Normand SLT. Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in medicine*. 1999; 18(3):321–359. [PubMed: 10070677]
3. Hedges, LV, Olkin, I. *Statistical methods for meta-analysis*. Academic press; 2014.
4. Hemminki E, McPherson K. Impact of postmenopausal hormone therapy on cardiovascular events and cancer: pooled data from clinical trials. *Bmj*. 1997; 315(7101):149–153. [PubMed: 9251544]
5. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies. *J natl cancer inst*. 1959; 22(4):719–748. [PubMed: 13655060]
6. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials*. 1986; 7(3):177–188. [PubMed: 3802833]
7. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in cardiovascular diseases*. 1985; 27(5):335–371. [PubMed: 2858114]
8. Sweeting M, J Sutton A, C Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in medicine*. 2004; 23(9):1351–1375. [PubMed: 15116347]
9. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine*. 2007; 26(1):53–77. [PubMed: 16596572]
10. 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. [PubMed: 17768699]
11. Cai T, Parast L, Ryan L. Meta-analysis for rare events. *Statistics in medicine*. 2010; 29(20):2078–2089. [PubMed: 20623822]
12. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in medicine*. 1995 Dec.14:2685–2699. [PubMed: 8619108]
13. Gart JJ, Pettigrew HM, Thomas DG. The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika*. 1985; 72(1):179–190.
14. Hitchcock SE. A note on the estimation of the parameters of the logistic function, using the minimum logit χ^2 method. *Biometrika*. 1962; 49(1/2):250–252.

15. Gart JJ, Zweifel JR. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*. 1967;181–187. [PubMed: 6049534]
16. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 1993; 88:9–25.
17. Agresti, A. *Categorical Data Analysis*. 3. John Wiley & Sons; 2013.
18. Bai O, Chen M, Wang X. Bayesian Estimation and Testing in Random Effects Meta-Analysis of Rare Binary Adverse Events. *Statistics in biopharmaceutical research*. 2016; 8(1):49–59. [PubMed: 27127551]
19. Parzen M, Lipsitz S, Ibrahim J, Klar N. An estimate of the odds ratio that always exists. *Journal of Computational and Graphical Statistics*. 2002; 11(2):420–436.
20. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*. 2007; 356(24):2457–2471. [PubMed: 17517853]
21. Rücker G, Schwarzer G. Presenting simulation results in a nested loop plot. *BMC medical research methodology*. 2014; 14(1):1. [PubMed: 24383436]
22. Efron, B, Tibshirani, RJ. *An introduction to the bootstrap*. CRC press; 1994.
23. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015; 67(1):1–48.
24. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954; 10(1): 101–129.
25. Brown MB, Forsythe AB. Robust tests for the equality of variances. *Journal of the American Statistical Association*. 1974; 69(346):364–367.

Figure 1. Comparison of large-sample performance based on $K = 50$ studies; (a) Bias comparison. For better visualization, we separate the methods into two panels, where the bottom panel includes methods with typically worse performance including DSL, EL and MH. (b) MSE comparison using $\log_2 \text{Re}$.

Figure 2.

Comparison of small-sample performance based on $K = 50$ studies; (a) Bias comparison. For better visualization, we separate the methods into two panels, where the bottom panel includes methods with typically worse performance including SA_1, DSL and EL. (b) MSE comparison using $\log_2 \text{Re}$.

Figure 3.
Data example of rosiglitazone meta-analysis: empirical density curves of estimated log odds for the treatment and control.

Table 1

Bias comparison: the least biased methods for small sample sizes based on simulation

θ	$w = 0$	$w = 0.5$	$w = 1$
< 0	GLMM	SA_0.25 & MH	SA_0.25 & iMUE
0	GLMM & iMUE	SA_0.25 & MH	SA_0.25

Table 2
MSE comparison: the most efficient methods for small sample sizes based on simulation

θ	$w = 0$				$w = 0.5$				$w = 1$			
	$R = 1$	$R = 2$	$R = 4$	$R = 1$	$R = 2$	$R = 4$	$R = 1$	$R = 2$	$R = 1$	$R = 2$	$R = 4$	$R = 1$
-1.5	GLMM	GLMM	GLMM	SA_1	MH (SA_0.25)	iMUE (SA_0.25)	SA_1 (DSL)	iMUE	DSL	SA_1	SA_1	SA_1
-1	GLMM	GLMM	GLMM	SA_1	SA_0.25 (MH)	DSL	DSL (SA_1)	DSL	DSL	SA_1	SA_1 (SA_0.5)	iMUE
-0.5	GLMM	SA_1	DSL	DSL	SA_1	SA_1	iMUE	SA_1	SA_1	iMUE	iMUE (SA_0.25)	iMUE
0	SA_1	DSL	GLMM (DSL)	SA_1	iMUE (SA_0.5)	iMUE	iMUE	SA_1	SA_1	iMUE	iMUE (SA_0.25)	iMUE
0.5	DSL (SA_1)	DSL	GLMM	SA_0.5 (iMUE)	iMUE	iMUE	iMUE	iMUE (SA_0.25)	SA_0.25	SA_0.25	SA_0.25	SA_0.25
1	DSL	GLMM	GLMM	iMUE (SA_0.25)	SA_0.25 (iMUE)	SA_0.25 (iMUE)	SA_0.25 (iMUE)	SA_0.25	SA_0.25	SA_0.25	SA_0.25	SA_0.25
1.5	GLMM	GLMM	GLMM	MH (SA_0.25)	SA_0.25 (MH)	SA_0.25 (MH)	SA_0.25 (MH)	SA_0.25 (MH)	SA_0.25 (MH)	SA_0.25 (MH)	SA_0.25 (MH)	SA_0.25 (MH)

Table 3

Results for MI data in rosiglitazone meta-analysis. Note that the results of DSL and EL are the same because the Q statistic²⁴ of this study is 0.

Para	SA_0.5	SA_0.25	SA_1	iMUE	GLMM	DSL	EL	MH
$\hat{\theta}$	0.021	0.179	-0.109	0.070	0.299	0.180	0.180	0.356
OR	1.021	1.196	0.896	1.073	1.348	1.198	1.198	1.427
95% CI	(0.653,1.598)	(0.841,1.701)	(0.737,1.090)	(0.804,1.432)	(0.963,1.886)	(0.887,1.619)	(0.887,1.619)	(1.021,1.995)

Results for CVD data example in rosiglitazone meta-analysis. Note that the results of DSL and EL are the same because the Q statistic²⁴ of this study is 0.

Table 4

Para	SA_0.5	SA_0.25	SA_1	iMUE	GLMM	DSL	EL	MH
$\hat{\theta}$	-0.037	0.107	-0.153	0.005	0.439	0.087	0.087	0.529
OR	0.964	1.113	0.858	1.005	1.551	1.092	1.092	1.698
95% CI	(0.583,1.594)	(0.857,1.446)	(0.732,1.005)	(0.810,1.248)	(0.897,2.682)	(0.724,1.647)	(0.724,1.647)	(0.970,2.972)