An integrative shrinkage estimator for random-effects meta-analysis of rare binary events

Lie Li, Ou Bai, Xinlei Wang*

Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, P O Box 750332, Dallas, TX 75275-0322, USA

Abstract

Meta-analysis has been a powerful tool for inferring the treatment effect between two experimental conditions from multiple studies of rare binary events. Recently, under a random-effects (RE) model, Bhaumik et al. developed a simple average (SA) estimator and showed that with the continuity correction factor 0.5, the SA estimator was the least biased among a set of commonly used estimators. In this paper, under various RE models that allow for treatment groups with equal and unequal variability (in either direction), we develop an integrative shrinkage (iSHRI) estimator based on the SA estimator, which aims to improve estimation efficiency in terms of mean squared error (MSE) that accounts for the bias-variance tradeoff. Through simulation, we find that iSHRI has better performance in general when compared with existing methods, in terms of bias, MSE, type I error and confidence interval coverage. Data examples of rosiglitazone meta-analysis are provided as well, where iSHRI yields competitive results.

1. Introduction

In medical research, when events of interest are rare, a single randomized study rarely has sufficient power to provide reliable information regarding the treatment effect between two experimental conditions (say, treatment vs. control). Therefore, meta-analysis is often used to combine information from multiple studies of rare binary events. In the past, various meta-analysis approaches have been developed to estimate the overall treatment effect, based on either fixed-effect (FE) models [e.g., 18] or random-effects (RE) models [e.g., 10]. Note that FE models assume a common treatment effect across different studies while RE models allow the treatment effects to vary from study to study.

When dealing with rare binary events, under the FE assumption, commonly used methods for estimating the treatment effect include the Mantel-Haenszel [MH, 18] with a constant zero-cell correction factor 0.5, Peto [30], and inverse variance methods [10,14]. A previous study [3] has compared the performance of twelve FE methods. The general recommendation is to use the MH method with some appropriate continuity corrections, which is consistent with the findings of Sweeting et al. [25]. In practice, RE models seem to be less restrictive, especially for clinical trials, because doses and follow-up time can be different in individual studies. Through a meta-analysis of multiple rosiglitazone studies, Shuster et al. [22] pointed out that the performance based on RE models is superior to that based on FE models. In this paper, we focus on meta-analysis of rare binary events based on RE models.

For RE meta-analysis, Shuster [21] showed that the widely used DerSimonian and Laird [DSL, 10] method can be highly biased for rare events, and suggested to use the simple (unweighted) average of estimates from individual studies. Recently, Bhaumik et al. [2] formally proposed a simple average (SA) estimator based on a RE model, which is the unweighted average of estimated log odd ratios (with a positive continuity correction factor σ) in individual studies. Bhaumik et al. [2] showed that, when σ = 1, the SA estimator (SA_0.5) is asymptotically unbiased and has superior bias performance when compared with existing estimators, including MH, empirical logit [EL, 10], and DSL methods. However, Li and Wang [17] pointed out that the RE model they assumed is restrictive in the sense that it forces the variability in the treatment group to be no less than that in the control group, and more importantly, SA_0.5 fails to minimize the mean squared error (MSE), which is an established measure of estimation performance that takes into account the bias-variance tradeoff.

Based on the SA estimator, we aim to develop a shrinkage estimator with smaller MSE to improve estimation efficiency. Shrinkage methods, which shrink some “standard” estimator toward zero or any other fixed value, have been widely used in various fields [24,26–29]. Many shrinkage methods [7,8,29] were developed under a rigorous Bayesian framework via empirical Bayes (EB) approaches, which shrink a point estimate from the sample to the prior. By contrast, the others were derived based on statistical decision theory directly (by minimizing...
some loss function), of which an example is the famous James—Stein estimator [6,16]. A recent example is that, to estimate the failure rate of rare events from multiple heterogeneous systems, Xiao and Xie [28] developed a plug-in shrinkage estimator based on Poisson distributions, which shrinks the maximum likelihood estimator (MLE) of the failure rate toward a predetermined point in the parameter space, and so lifts degenerated zero estimates to some reasonable positive values. This estimator can be expressed as a weighted average of the MLE and a data-dependent point.

In the context of meta-analysis of rare binary events, we develop an integrative shrinkage estimator (iSHRI) by using the SA method in Bhaumik et al. [2] to combine shrinkage estimators from individual studies, which intends to shrink the estimated log odd ratios with $a = \frac{1}{2}$ toward a predetermined point. The shrinkage factor is obtained from minimizing the MSE and then plug in estimates from the data. We thoroughly compare the bias and MSE of the proposed method with existing methods via simulation. We further examine their performance on hypothesis testing and interval estimation using the type I error and coverage probability of confidence intervals (CIs). Besides, two data examples of rosiglitazone meta-analysis are provided for further comparison.

2. An integrative shrinkage estimator

Consider a meta-analysis consisting of $K$ randomized studies. In the $k$th study, let $n_k$ and $a_k$ be the number of subjects in the treatment and control groups, respectively. Assume $x_k \sim \text{Binomial}(n_k, p_k)$ and $x_k' \sim \text{Binomial}(n_k, p_k')$, where $x_k$ is the number of observed events of interest in the $k$th study, and $p_k(p_k')$ is the probability of observing an event in the treatment (control) group. Let $q_k \equiv 1 - p_k$, and $q_k' \equiv 1 - p_k'$. To measure the treatment effect in study $k$, the log odd ratio is used throughout this paper, i.e., $\theta_k \equiv \ln\frac{x_k}{n_k - x_k} + \ln\frac{a_k}{n_k - a_k}$. We denote the mean treatment effect across component studies by $\theta$, and the between-study heterogeneity among individual treatment effects by $\tau^2$, satisfying $\tau^2 > 0$. Below we consider three RE models, to accommodate realistic situations where treatment and control groups can have equal variability or unequal variability (in either direction). All the three models involve random terms $\theta_k$’s and $\mu_k$’s, where $\theta_k \overset{iid}{\sim} N(\theta, \tau^2)$, $\mu_k \overset{iid}{\sim} N(\mu, \tau^2)$, and any two components of $(\mu_1, \ldots; \mu_K; \theta_1, \ldots, \theta_K)$ are assumed to be independent. Here, $\mu_k$ can represent the log odds of the control or treatment group or the average of the two groups, depending on model specification below.

1. Model I is the one used in Bhaumik et al. [2], which implicitly assumes that the variance of $\logit(p_k)$ is greater than that of $\logit(p_k')$:
   \[
   \logit(p_k) = \mu_k, \quad \logit(p_k') = \mu_k + \theta_k.
   \]

2. Model II is the one used in Smith et al. [23], which assumes equal variances between the two groups:
   \[
   \logit(p_k) = \mu_k - \frac{\theta_k}{2}, \quad \logit(p_k') = \mu_k + \frac{\theta_k}{2}.
   \]

3. Model III assumes the variance in the treatment is less than that in the control:
   \[
   \logit(p_k) = \mu_k - \theta_k, \quad \logit(p_k') = \mu_k.
   \]

Although originally developed under Model I, the SA estimator $\hat{\theta}_k$ can be used with any of the three models to estimate the mean treatment effect $\theta$, given by:

\[
\hat{\theta}_k = \frac{1}{K} \sum_{i=1}^{K} \hat{\theta}_{ki},
\]

where $\hat{\theta}_{ki}$ is an estimator of the individual treatment effect $\theta_k$ in the $k$th study.

$$
\hat{\theta}_k = \ln \frac{x_k + a}{n_k - x_k + a} - \ln \frac{x_k' + a}{n_k - x_k' + a}.
$$

and $a$ is a positive continuity correction factor. Bhaumik et al. [2] proved that $\hat{\theta}_k$ is asymptotically unbiased when $a = \frac{1}{2}$ under Model I; that is, $E(\hat{\theta}_k^2) = \theta + O(n^{-2})$, where $n \equiv \min\{(n_k, n_k'), n_k, n_k'\}$ is the overall minimum number of subjects. It can be shown that the asymptotic unbiasedness of $\hat{\theta}_k$ also holds under Models II and III. This theoretical property ensures that SA_0.5 (i.e., $\hat{\theta}_k$) performs well in terms of bias for large sample sizes. However, Li and Wang [17] proved that SA_0.5 is suboptimal in terms of MSE, and showed via simulation that it can have poor MSE performance especially for small sample sizes.

Motivated by Xiao and Xie [28], in the $k$th study, we consider a shrinkage estimator of $\hat{\theta}_k$ based on $\hat{\theta}_{2k}$, denoted by $\hat{\theta}_{2k}$, where $\bar{\theta}_0$ is a fixed point in the parameter space of $\theta$, and $c$ is a shrinkage factor. Then our integrative shrinkage estimator for the overall treatment effect $\theta$ can be given by:

$$
\hat{\theta}_{iSHRI} = \frac{\sum_{k=1}^{K} \hat{\theta}_{2k}}{K} = c\hat{\theta}_k + (1 - c)\bar{\theta}_0.
$$

Clearly, the iSHRI estimator shrinks SA_0.5 toward the predetermined point $\bar{\theta}_0$.

Based on the asymptotic unbiasedness of $\hat{\theta}_k$, it is easy to show

Bias($\hat{\theta}_{iSHRI}$) = $(1 - c)\theta - \bar{\theta}_0 + O(n^{-2})$.

Further, we can show that the variance of $\hat{\theta}_{iSHRI}$ is given by

$$
V(\hat{\theta}_{iSHRI}) = c^2 V(\hat{\theta}_k) = \frac{c^2}{K} \sum_{k=1}^{K} \left[ E_P \left( \frac{1}{n_k p_k q_k} \right) + E_{P_k'} \left( \frac{1}{n_k p_k q_k} \right) \right] + \frac{c^2}{K} \tau^2 + O(n^{-2}).
$$

Next, if the model specification, $\hat{\theta}_{iSHRI}$, has the following MSE:

$$
\text{MSE}(\hat{\theta}_{iSHRI}) = (1 - c)^2 \theta^2 + \frac{c^2}{K} \sum_{k=1}^{K} \left[ E_P \left( \frac{1}{n_k p_k q_k} \right) + E_{P_k'} \left( \frac{1}{n_k p_k q_k} \right) \right] + O(n^{-2}).
$$

To minimize the asymptotic MSE of $\hat{\theta}_{iSHRI}$, we ignore the term $O(n^{-2})$ and set the shrinkage factor $c$ to

$$
c = \frac{(\theta - \bar{\theta}_0)^2 + \frac{\tau^2}{K}} {\theta^2 + \frac{\tau^2}{K}} \sum_{k=1}^{K} \left[ E_P \left( \frac{1}{n_k p_k q_k} \right) + E_{P_k'} \left( \frac{1}{n_k p_k q_k} \right) \right].
$$

satisfying $c \leq 1$. The minimized asymptotic MSE of iSHRI, denoted by mAMSE, is always less than or equal to the asymptotic MSE of SA_0.5:

$$
mAMSE(\hat{\theta}_{iSHRI}) = c^2 \left[ \frac{\tau^2}{K} \sum_{k=1}^{K} \left[ E_P \left( \frac{1}{n_k p_k q_k} \right) + E_{P_k'} \left( \frac{1}{n_k p_k q_k} \right) \right] \right] + O(n^{-2}).
$$

To be able to calculate the shrinkage factor $c^*$ in (3), we need to estimate $\theta$, $\tau^2$, $E_P \left( \frac{1}{n_k p_k q_k} \right)$ and $E_{P_k'} \left( \frac{1}{n_k p_k q_k} \right)$ from the data as well as choosing an appropriate value of $\bar{\theta}_0$. We estimate $\theta$ using SA_0.5. Based on some preliminary simulation, we find that how to estimate $\tau^2$ (e.g., the popular DSL estimator or other estimators introduced in Bhaumik et al. 2), has not much effect on the performance of iSHRI. Thus, we set $\tau^2 = 0$ for simplicity. We also adopt the estimators in Gart et al. [12] for $E_P \left( \frac{1}{n_k p_k q_k} \right)$ and $E_{P_k'} \left( \frac{1}{n_k p_k q_k} \right)$:
Further, we use $\hat{\Delta}_4$ as a surrogate of the fixed point $\delta_0$, due to the following reasons. First of all, our simulation results (not reported here) show that the obvious choice $\delta_0 = 0$ does not work well. There seems to be no other plausible constant in $(0,1)$ that can be used to set $\delta_0$. Second, iSHRI is a weighted average of SA_0.5 and $\delta_0$. Thus, it would be reasonable to find a good estimator of $\theta$ to replace $\delta_0$. Hitchcock [15] and Gart and Zweifel [13] recommended $a = \frac{1}{2}$ as an alternative continuity correction to estimate the log odds ratio, based on the minimum logit $\chi^2$ method. Further, Li and Wang [17] show that SA_0.25 performs well in general. Note that using $\hat{\delta}_4$ to replace $\delta_0$ would lead to an increase in the variance of $\hat{\delta}_{iSHRI}$, and so the derivation of the optimal shrinkage factor $c^*$ would become analytically difficult. To avoid this difficulty, we directly replace $\delta_0$ by $\hat{\delta}_2$ in (3). This plug-in strategy was also adopted by Xiao and Xie [28], which turns out to work well in our context. We will show via simulation that it works well in the context of meta-analysis of rare binary events, too.

Based on these estimators, the data-dependent shrinkage factor $c^*$ is given by

$$c^* = \left( \frac{\hat{\Delta}_2 - \hat{\Delta}_4}{\hat{\Delta}_2} \right)^2 + \frac{1}{2} \sum_{k=1}^{K} \left( \frac{n_k + 1}{n_k} \frac{1}{\tau_k + \frac{1}{6}} + \frac{1}{\tau_k - \frac{1}{6}} + \frac{1}{n_k - \frac{1}{6}} + \frac{1}{n_k + \frac{1}{6}} \right).$$

From now on, we refer to the estimator $\hat{c}^* \hat{\Delta}_2 + (1 - \hat{c}^*) \hat{\Delta}_4$ as the iSHRI estimator.

3. Simulation

We conducted a simulation study to compare the performance of the proposed method iSHRI with various existing methods including SA_0.5, DSL, EL, MH, and GLMM [4], among which the first four were included in the comparison done by Bhaumik et al. [2] and the last one, GLMM, stands for the popular estimator based on generalized linear mixed models, which has been reported to work well for analysis of sparse binary data. For the purpose of point estimation, we examined the bias and MSE of these methods; and for the purpose of hypothesis testing and interval estimation, we evaluated the type I error rate for testing $H_0: \delta = 0$ and per cent coverage of 95% CIs.

In our simulation, we set the number of studies $K \in \{10, 20, 50\}$, the mean treatment effect $\delta \in [-1.5, -1.25, ..., 1.5]$, $\mu = -5$ for rare binary events (i.e., 0.0067 in the probability magnitude), and $\tau_0^2 = \tau_1^2 = 0.5$. We simulated $n_k$ and $n_0$ from uniform[20,1000], to mimic realistic situations where sample sizes of the two groups can be similar or very different. In this way, for the rare events simulated with $\mu = -5$, $n_0$ would not be large enough for SA_0.5 to become unbiased in most simulated cases. We generated the responses $x_{0k}$ and $x_{0k}$ from Binomial$(n_0, \mu_0)$ and Binomial$(n_0, \mu_0)$, respectively, for $k = 1, ..., K$, where $(\mu_0, \mu_0)$’s were computed from Models I, II and III, respectively. For each setting, we simulated 1000 datasets to compute the empirical bias, MSE, type I error rate and per cent coverage of all the methods.

3.1. Comparing bias

The results for bias comparison are shown in Fig. 1. The number of studies $K$ seems not to affect the relative performance of the different methods while the model specification matters a lot. GLMM is the least biased under Model I, which assumes that the variability in the treatment is larger than that in the control. MH is the best under Model II, in which equal variability is assumed. Under Model III that assumes the variability in the treatment is smaller than that in the control, iSHRI is the least biased except for large positive $\theta$, where MH and GLMM outperform iSHRI; however, MH and GLMM are the worst two for many other $\theta$ values. Overall, iSHRI works the best in terms of bias because iSHRI is the best or always among the top performing group in all the settings. Note that the performance of SA_0.5 is consistently worse than that of iSHRI. This is because in these simulated settings, SA_0.5 is not asymptotically unbiased due to generally small $n$, as mentioned before.

Another interesting observation is that for Model I, bias is large for $\delta < 0$; and for Model III, bias is large for $\delta > 0$. This is because in these cases, the number of events of interest can be extremely low in one group (treatment or control) because $E(\logit(\mu_0)) = \mu + \delta$ is very small for Model I with $\delta < 0$ and $E(\logit(\mu_0)) = \mu - \delta$ is very small for Model III with $\delta > 0$. From Fig. 1, we can find that the bias of all methods is large when the number of events is extremely low. As will be shown later, the poor bias performance leads to large MSE in Fig. 2 and low CI coverage rates in Fig. 3 for these cases.

3.2. Comparing MSE

Fig. 2 shows results for MSE comparison. Clearly, there is no single estimator that beats the others in all the settings. Under Model I, iSHRI seems to be the best for $K = 10$ or 20, and GLMM seems to be the best for $K = 50$. Under Model II, the performance of the different methods is less heterogeneous and iSHRI performs well when $\theta$ moves away from zero. Under Model III, iSHRI offers either the best or close to the best MSE performance in all cases. Overall, iSHRI has very strong performance in terms of MSE.

3.3. Evaluating type I error

For testing $H_0$: $\delta = 0$, we used the statistic $T = \hat{\delta}/\sqrt{SE(\hat{\delta})}$, where $\hat{\delta}$ is an estimator of $\theta$, and $SE(\hat{\delta})$ stands for the standard estimated SE of $\hat{\delta}$. The methods described in Bhaumik et al. [2] were used to estimate the SEs of SA_0.5, DSL, EL and MH. For iSHRI, it is hard to get a closed form formula for the SE, and so we used the delete-$d$ jackknife method [11] with $d = 2$ to estimate the SEs. For GLMM, we used an R package named lme4 [1] to estimate the SEs. In our preliminary simulation, we find that when $K$ is not large, based on the statistic $T$, a $t$-test with the number of degrees of freedom $df = K - 1$ works better than a $z$-test for all the methods except for SA_0.5; and when $K$ is large, the two tests are equivalent. Thus, we performed the $z$-test for SA_0.5, and the $t$-test for other methods. Using data generated under $H_0$, we computed empirical type I error rates at the significance level 0.05, and report the results in Table 1. Unlike other methods, the rates of SA_0.5 are consistently smaller than the target level 0.05, meaning that SA_0.5 is quite conservative in rejecting $H_0$. Among all, the rates of iSHRI are closest to 0.05 in nearly all the settings, although they are slightly above 0.05. By contrast, GLMM has inflated rates under Models II and III; DSL have inflated rates under Models I and III; and the rates of EL and MH are highly inflated for all the cases.

3.4. Comparing CI coverage

To compute the 95% CIs, we used $\hat{\delta} \pm 1.96 \sqrt{SE(\hat{\delta})}$ for SA_0.5, and $\hat{\delta} \pm t_{0.025,K-1} \sqrt{SE(\hat{\delta})}$ for the other methods, where $t_{0.025,K-1}$ is the critical value of the $t$ distribution with $df = K - 1$. Then for each method considered, we computed the empirical proportion of the 95% CIs that contain the true $\theta$ value among 1000 replicates in each setting. To clearly show differences among the estimators especially in the neighborhood of the nominal level 95% (represented by the black solid line), we focus on coverage rates from 50% to 100% in Fig. 3. Under Model I, GLMM appears to be the best and its coverage is quite close to 95%; and iSHRI perform as well as GLMM when $\theta$ is large positive or close to zero. Under Model II, iSHRI is the best for $K = 10$ and 20; and it is also

\[ \hat{\theta}_P\left(\frac{1}{n_0/n_0/n_0}\right) = \frac{n_0 + 1}{n_0} \left( \frac{1}{\tau_k + \frac{1}{6}} + \frac{1}{n_k - \frac{1}{6}} \right). \]

\[ \hat{\theta}_P\left(\frac{1}{n_0/n_0/n_0}\right) = \frac{n_0 + 1}{n_0} \left( \frac{1}{\tau_k + \frac{1}{6}} + \frac{1}{n_k - \frac{1}{6}} \right). \]
Fig. 1. Comparison of bias.

Fig. 2. Comparison of MSE.
the best for $K = 50$ except for large positive/negative $\theta$. Under Model III, iSHRI is the best except for some cases with large positive $\theta$, where SA_0.5 provides coverage closer to 95%. Note that although GLMM performs well under Model I, it performs poorly under Model III; and the other methods including DSL, EL and MH do not perform well in general. Thus, the performance of iSHRI is competitive in terms of CI coverage.

4. Data examples

Nissen and Wolski [19] conducted a meta-analysis of 48 randomized trials to evaluate potential adverse effects of rosiglitazone therapy, including the risks of myocardial infarction (MI) and cardiovascular death (CVD). Later, Nissen and Wolski [20] updated the analysis with 56 randomized trials including the previous 48 trials. In this section, we applied the six methods, namely, SA_0.5, iSHRI, GLMM, DSL, EL and MH, to evaluate adverse effects of rosiglitazone with respect to MI and CVD based on these two studies.

4.1. MI data

Table 2 summarizes results from the different methods for meta-analysis of the MI data of the 48 and 56 trials, respectively, including the estimated treatment effect $\hat{\theta}$, estimated odds ratio (OR), and 95% CI of the OR. We used the same methods described in Section 3 to compute the CIs of each method. The results of DSL and EL are the same because the $Q$ statistic [9] is 0 in both analyses. From Table 2, we find that, for the 48 and 56 trials, all the methods report positive estimates of $\theta$ that belong to the interval $[0,0.5]$; and only MH provide CIs of OR that do not include 1, meaning that all the other methods do not reject $H_0: \theta = 0$ except for MH.

We conducted the Brown-Forsythe Levene-type (BFL) test for equal variances [5] using sample log odds (with the correction factor 0.5); and for the MI data in both studies, the test did not reject the null at the significance level $\alpha = 0.05$. Thus, both MI studies may fit in simulation settings with $K = 50$ and $\theta \in [0,0.5]$ under Model I. Based on our simulation results in Section 3, for such settings, iSHRI and MH are the least biased, and all the methods achieve similar small MSE values. Meanwhile, iSHRI performs the best in maintaining the nominal type I error as well as the nominal CI coverage, suggesting iSHRI is the best choice for the purpose of testing. Here, even for the purpose of estimation, we believe ISHRI is a much safer choice than MH. Although the BFL test did not reject variance equality, there is still a small chance that the variance in the treatment is larger than that in the control based on their variance estimates; and for $\theta \in [0,0.5]$ under Model I, MH severely overestimates the treatment effect and its MSE is also the largest among all. Note that MH does give the largest positive estimates.
for both MI studies. In addition, between the two studies, we find iSHRI gives the most consistent estimates of $\theta$ (0.107 vs. 0.108) while MH gave estimates with the largest difference (0.356 vs. 0.250).

To summarize, we decide to rely on the results from iSHRI, and conclude that the (small) positive effect of rosiglitazone on MI is not statistically significant.

### 4.2. CVD data

As shown in Table 3, for the CVD data, SA_0.5 gives slightly negative estimates in both studies, and DSL and EL return slightly negative estimates in the updated study but positive estimates in the first study, while the other methods report both positive estimates. Comparing estimates of $\theta$ between the two studies, SA_0.5 and iSHRI have the smallest changes while GLMM and MH change a lot, even though their sign does not change. Note that for any adverse effect, $\theta$ is expected to be nonnegative, otherwise the “adverse” effect is actually beneficial. All the CIs cover 1, suggesting the action of not rejecting $H_0: \theta = 0$.

The BFL test for equal variances in the study of the 48 trials returns a p-value of 0.040, suggesting Model I; and it returns a p-value of 0.246 in the study of the 56 trials, suggesting Model II. Thus, the CVD data of the 56 trials returns $\hat{\theta}$ close to zero, and there is no strong evidence to support its existence.

### 5. Discussion

For meta-analysis of rare binary events, an integrative shrinkage estimator of treatment effects, called iSHRI, has been developed based on the SA estimator proposed in Bhaumik et al. [2]. We have shown that iSHRI should have smaller MSE than SA_0.5 if individual sample size is sufficiently large. Further, the performance of iSHRI has been compared with commonly used methods in terms of bias, MSE, type I error, and CI coverage under various settings simulated to mimic realistic situations. We find that for relatively small sample sizes, the performance of a meta-analysis method depends largely on the variability direction (i.e., which group has larger variability) as well as the size and direction of the mean treatment effect $\theta$. However, iSHRI has generally satisfactory performance in both estimation and testing, which seems to be better than or comparable to other top methods in practice. The performance of a meta-analysis method depends largely on the variability direction (i.e., which group has larger variability) as well as the size and direction of the mean treatment effect $\theta$. However, iSHRI has generally satisfactory performance in both estimation and testing, which seems to be better than or comparable to other top methods in practice.

We end this paper by pointing out several major differences between our iSHRI method and that proposed in Xiao and Xie [28]. Firstly, the underlying research problems are different, though both involve rare events. Xiao and Xie [28] consider the problem of estimating the failure rate for highly reliable systems while we consider the problem of estimating the treatment effect from multiple independent studies in the context of meta-analysis of rare binary events. Thus, the application scopes of the two methods are completely different. Second, due to the different problem set-ups, Xiao and Xie [28] rely on Poisson distributions while we rely on binomial-normal hierarchical models for statistical inference. In addition, in terms of technical detail, Xiao and Xie [28] shrink the maximum likelihood estimator of the failure rate, and we shrink the simple average estimator of the treatment effect that is nonparametric; and so our derivation is different from theirs.

### Acknowledgements

This work was supported by the NIH grant R15GM113157 (PI: Xinlei Wang).

### References

[1] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, J. Stat. Software 67 (1) (2015) 1–48.
[2] D.K. Bhaumik, A. Amatya, S.-L.T. Normand, J. Greenhouse, E. Kaizar, B. Neelon, R.D. Gibbons, Meta-analysis of rare binary adverse event data, J. Am. Stat. Assoc. 107 (498) (2012) 555–567.
[3] M.J. Bradburn, J.J. Deeks, J.A. Berlin, A. Russell Localio, Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events, Stat. Med. 26 (1) (2007) 53–77.
[4] N.E. Breslow, D.G. Clayton, Approximate inference in generalized linear mixed models, J. Am. Stat. Assoc. 88 (1993) 9–25.
[5] M.B. Brown, A.B. Fosythe, Robust tests for the equality of variances, J. Am. Stat. Assoc. 69 (346) (1974) 364–367.
[6] N.A. Butler, M.C. Denham, The peculiar shrinkage properties of partial least squares regression, J. Roy. Stat. Soc. B 62 (3) (2000) 585–595.
[7] B.P. Garlin, T.A. Louis, Bayes and Empirical Bayes Methods for Data Analysis vol. 17, Chapman & Hall/CRC Boca Raton, FL, 2000.
[8] M. Clyde, E.I. George, Flexible empirical bayes inference for wavelets, J. Roy. Stat. Soc. B 62 (4) (2000) 681–698.
[9] W.G. Cochran, The combination of estimates from different experiments, Biometrics 10 (1) (1954) 101–129.
[10] R. DerSimonian, N. Laird, Meta-analysis in clinical trials, Contr. Clin. Trials 7 (3) (1986) 177–188.
[11] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, CRC press, 1994.
[12] J.J. Gart, H.M. Pettigrew, D.G. Thomas, The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses, Biometrika 72 (1) (1985) 179–190.
[13] J.J. Gart, J.R. Zweifel, On the bias of various estimators of the logit and its variance with application to quantal bioassay, Biometrika (1967) 181–187.
[14] S. Greenland, J.M. Rothstein, Estimation of a common effect parameter from sparse follow-up data, Biometrics 41 (1985) 55–68.
[15] S.E. Hitchcock, A note on the estimation of the parameters of the logistic function, using the minimum logit $\gamma$ 2 method, Biometrika 49 (1/2) (1962) 250–252.
[16] W. James, C. Stein, Estimation with quadratic loss, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1961, pp. 361–379.
[17] L. Li, X. Wang, Meta-analysis of rare binary events in treatment groups with unequal variance, Stat. Med. Med. Res. (2007), http://dx.doi.org/10.1177/0962822907212146 In Press.
[18] N. Mantel, W. Haenszel, Statistical aspects of the analysis of data from retrospective studies, J. Natl. Cancer Inst. 22 (4) (1959) 719–748.
[19] S.E. Nissen, K. Wolski, Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes, N. Engl. J. Med. 356 (2007) 2457–2471.
[20] S.E. Nissen, K. Wolski, Rosiglitazone revisited: an updated meta-analysis of risk for myocardial infarction and cardiovascular mortality, Arch. Intern. Med. 170 (14) (2010) 1191–1201.
[21] J.J. Shuster, Empirical vs natural weighting in random effects meta-analysis, Stat. Med. 29 (12) (2010) 1259–1265.
[22] J.J. Shuster, L.S. Jones, D.A. Salmon, Fixed vs random effects meta-analysis in rare

### Table 3

Results for CVD data in rosiglitazone meta-analysis.

| Study            | Para   | SA_0.5 | iSHRI | GLMM | DSL | EL   | MH   |
|------------------|--------|--------|-------|------|-----|------|------|
| Nissen and Wolski [19] | $\hat{\theta}$ | -0.037 | 0.054 | 0.439 | 0.087 | 0.087 | 0.529 |
|                  | OR     | 0.964  | 1.055 | 1.551 | 1.092 | 1.092 | 1.698 |
|                  | 95% CI | (0.583,1.594) | (0.859,1.297) | (0.897,2.681) | (0.724,1.647) | (0.724,1.647) | (0.970,2.972) |
| Nissen and Wolski [20] | $\hat{\theta}$ | -0.065 | 0.011 | 0.009 | -0.057 | -0.057 | 0.026 |
|                  | OR     | 0.937  | 1.011 | 1.099 | 0.944 | 0.944 | 1.026 |
|                  | 95% CI | (0.590,1.487) | (0.831,1.233) | (0.759,1.341) | (0.729,1.222) | (0.729,1.222) | (0.779,1.367) |
event studies: the rosiglitazone link with myocardial infarction and cardiac death, Stat. Med. 26 (2007) 4375–4385.

[23] T.C. Smith, D.J. Spiegelhalter, A. Thomas, Bayesian approaches to random-effects meta-analysis: a comparative study, Stat. Med. 14 (1995) 2665–2699.

[24] J.H. Stock, M.W. Watson, Generalized shrinkage methods for forecasting using many predictors, J. Bus. Econ. Stat. 30 (4) (2012) 481–493.

[25] M. Sweeting, J. Sutton A, C. Lambert P, What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data, Stat. Med. 23 (9) (2004) 1351–1375.

[26] T. Tong, Y. Wang, Optimal shrinkage estimation of variances with applications to microarray data analysis, J. Am. Stat. Assoc. 102 (477) (2007) 113–122.

[27] J.K. Vaurio, K.E. Jänkälä, Evaluation and comparison of estimation methods for failure rates and probabilities, Reliab. Eng. Syst. Saf. 91 (2) (2006) 209–221.

[28] X. Xiao, M. Xie, A shrinkage approach for failure rate estimation of rare events, Qual. Reliab. Eng. Int. 32 (1) (2016) 123–132.

[29] S. Xu, An empirical bayes method for estimating epistatic effects of quantitative trait loci, Biometrics 63 (2) (2007) 513–521.

[30] S. Yusuf, R. Peto, J. Lewis, R. Collins, P. Sleight, Beta blockade during and after myocardial infarction: an overview of the randomized trials, Prog. Cardiovasc. Dis. 27 (5) (1985) 335–371.