

NONPARAMETRIC INFERENCE PROCEDURE FOR PERCENTILES OF THE
RANDOM EFFECTS DISTRIBUTION IN META ANALYSIS

BY RUI WANG¹, LU TIAN², TIANXI CAI³, AND L.J. WEI⁴

Harvard University and Massachusetts General Hospital, Stanford University, Harvard
University and Harvard University

To investigate whether treating cancer patients with erythropoiesis-stimulating agents (ESAs) would increase the mortality risk, recently Bennett et al. (2008) conducted a meta analysis with the data from 52 phase III trials comparing ESAs with placebo or standard of care. With a standard parametric random effects modeling approach, the study concluded that ESA administration was significantly associated with increased *average* mortality risk. In this article, we present a simple *nonparametric* inference procedure for the *distribution* of the random effects. We re-analyzed the ESA mortality data with the new method. Our results about the center of the random effects distribution were markedly different from those reported in Bennett et al. Moreover, our procedure, which estimates the distribution of the random effects, as opposed to just a simple population average, suggests that the ESA may be beneficial for approximately a quarter of the study populations with respect to mortality. This new meta analysis technique can be implemented with study-level summary statistics. In contrast to existing methods for parametric random effects models, the validity of our proposal does not require the number of studies involved to be large. From the results of an extensive numerical study, we find that the new procedure performs well even with moderate individual study sample sizes.

Keywords: Bivariate beta; Conditional permutation test; Erythropoiesis-stimulating agents; Logit-normal, Two-level hierarchical model.

¹Supported by NIH grants R37 AI24643 and T32 AI007358.

²Supported by NIH grant R01 HL089778

³Supported by NIH grants R37 AI24643 and U54 LM008748

⁴Supported by NIH grants R01 AI052817 and U54 LM008748

1. Introduction

Conventional meta analysis techniques have been utilized frequently to make inferences about a single parameter, for example, the center of the distribution of the random or fixed effects. Under the random effects model, the procedure for estimating the *mean* of the random effects proposed by DerSimonian and Laird (DL) (1986) is routinely used in practice. Their method utilizes a linear combination of study-specific point estimates with the weights depending on the within- and among-study variance estimates. This procedure is simple to implement and does not require patient-level data. The validity of the procedure, however, heavily depends on the individual study sample sizes and the number of studies [Brockwell and Gordon (2001), Bohning et al. (2002), Sidik and Jonkman (2007), Viechtbauer (2007)]. In addition, this and other related methods for random effects models in meta analysis do not provide inferences about the distribution function of the random effects. Estimation of this distribution function or its quantile counterpart provides valuable information for the complex risk-benefit decision on a new drug or device.

In a recent meta analysis, Bennett et al. (2008) examined whether the erythropoiesis-stimulating agent (ESA) for treating anemia of cancer patients would increase the patient's risk of mortality with the data from 52 phase III comparative trials (ESA vs. placebo or standard of care). In Table 1, we present their two-sample study-specific hazard ratio point and 95% interval estimates. Bennett et al. (2008) concluded that administration of ESAs was significantly associated with increased mortality. Using the aforementioned DL method for the mortality data, the resulting 0.95 confidence interval for the *mean* of the random hazard ratios (treated vs. untreated with ESA) across the studies was (1.01 – 1.20). Note that the lower bound of the above interval is barely over one. Furthermore, it is known in the literature that the DL method sometimes can produce liberal confidence interval estimates, that is, the true coverage level of the DL interval estimator tends to be smaller (sometimes substantially) than its nominal counterpart [Hardy and Thompson (1996), Brockwell and

Gordon (2001, 2007), Sidik and Jonkman (2002)]. Therefore, it is possible that the interval estimates reported in Bennett et al. are “too tight”. Moreover, from Table 1, it appears that the study-specific hazard ratio estimates for half of the trials are less than one, suggesting that even if the average hazard ratio is more than one, the ESA may not be harmful uniformly across all study populations. Lastly, since the DL method is based on a weighted average of hazard ratio estimates, the resulting interval estimates may be sensitive to outliers.

In this article, we propose a simple inference procedure for the percentiles of the random effects distribution based on study-level data without assuming a parametric form of the distribution. We re-analyzed the mortality data reported in Bennett et al. (2008). The resulting 0.95 confidence interval for the *median* of the random hazard ratios was (0.94, 1.26). The 0.95 confidence interval for the lower quartile of the random hazard ratio was (0.70, 0.99), indicating that, in approximately a quarter of the study populations, patients on average may benefit from ESA treatment with respect to mortality. In contrast to all existing methods that can only handle inference problems for the center of the random effects distribution, the validity of the new proposal does not require the number of studies to be large. The new proposal is theoretically valid when the sample sizes of individual studies are large. Through an extensive numerical study, we find that the new method performs well even with moderate individual study sample sizes. On the other hand, the commonly used DL method tends to give liberal confidence interval estimators, that is, their coverage levels can be markedly smaller than their nominal counterparts.

2. Interval Estimates for Percentiles of the Random Effects Distribution

Consider a typical two-level hierarchical model. Let $\Pi' = (\Theta, \Lambda')$ be a row vector of random parameters, where Θ is a univariate parameter of interest and Λ is a finite- or infinite-dimensional vector of nuisance parameters. Let $G(\cdot)$ be the continuous, completely unspecified distribution function of Θ . Given an *unobservable* realization Π , a data set X is

generated. Let $\{\Pi_k, X_k\}, k = 1, \dots, K$, be K independent copies of $\{\Pi, X\}$. The problem is how to make inferences, for instance, about the median μ of $G(\cdot)$ with $\{X_k, k = 1, \dots, K\}$. As an example, consider the case with K 2×2 tables and let Θ_k be the log-risk-ratio or risk difference for the k th table. Here, the nuisance parameter Λ_k consists of the underlying event rate for the “control” group and the sample size for the k th study n_k .

If we can observe $\{\Theta_k, k = 1, \dots, K\}$, a simple nonparametric estimator for μ is the sample median. Exact confidence intervals for μ can be obtained by inverting a sign test for testing the null hypothesis that the median is μ_0 . Under $H_0 : \mu = \mu_0$, consider

$$T(\mu_0) = \sum_{k=1}^K B_k, \quad (1)$$

where $B_k = I(\Theta_k < \mu_0) - I(\Theta_k > \mu_0)$ and $I(\cdot)$ is the indicator function. The null distribution of $T(\mu_0)$ can be generated by

$$T^* = \sum_{i=1}^K \Delta_k, \text{ where } \Delta_k = \begin{cases} 1 & \text{with prob. 0.5} \\ -1 & \text{otherwise.} \end{cases} \quad (2)$$

Suppose that given Π_k , $\hat{\Theta}_k$ is a consistent estimator for Θ_k based on the data X_k . To test H_0 , one may replace Θ_k in (1) with $\hat{\Theta}_k$. This results in a test statistic

$$\tilde{T}(\mu_0) = \sum_{k=1}^K \hat{B}_k = \sum_{k=1}^K \{I(\hat{\Theta}_k < \mu_0) - I(\hat{\Theta}_k > \mu_0)\}. \quad (3)$$

When the sample size n_k for each individual study, $k = 1, \dots, K$, is large, we can make inferences about the median by comparing the observed value of (3) to the distribution of (2).

Now, the test based on (3) does not take into account the precision of the estimator $\hat{\Theta}_k$. It gives equal weight to each individual study. Consider the k th study, suppose that

the variance $\hat{\sigma}_k^2$ of $\hat{\Theta}_k$ is large with respect to the distance between Θ_k and μ_0 . Then the likelihood of the unobservable $\Theta_k < \mu_0$ can be quite close to $1/2$ (like tossing a fair coin). Therefore, the noise generated from such an unstable variable \hat{B}_k may well outweigh its added value to the power of the test based on $\tilde{T}(\mu_0)$. On the other hand, if $\hat{\sigma}_k^2$ is small and $\hat{\Theta}_k < \mu_0$, the likelihood of $\Theta_k < \mu_0$ would be closer to 1.

This motivates us to consider a simple modification of test statistic (3) by putting a weight w_k for \hat{B}_k in (3). Here, the weight w_k is a measure of likelihood for the event, $\Theta_k < \mu_0$, for example, the observed coverage level of the interval $(-\infty, \mu_0)$ for the realized Θ_k . When the individual study sizes $n_k, k = 1, \dots, K$, are large, and the distribution of $\hat{\Theta}_k$ conditional on Π_k is approximately normal with mean Θ_k and variance $\hat{\sigma}_k^2$, where $n_k \hat{\sigma}_k^2$ converges to a constant, this coverage level is approximately $\Phi((\mu_0 - \hat{\Theta}_k)/\hat{\sigma}_k)$, where Φ is the distribution function of the standard normal. Let the resulting test statistic be

$$\hat{T}(\mu_0) = \sum_{k=1}^K |\Phi((\mu_0 - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2| \hat{B}_k. \quad (4)$$

In the Appendix we show that in probability, for any given μ ,

$$|\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2| \hat{B}_k - B_k/2 \rightarrow 0, \quad \text{as } n_k \rightarrow \infty. \quad (5)$$

It follows that for fixed K , for large $n_k, k = 1, \dots, K$, the distribution of $\hat{T}(\mu_0)$ can be approximated by that of $T(\mu_0)$. This approximation, however, is rather discrete and for moderate sample sizes, the resulting confidence intervals for μ do not have adequate coverage levels based on the results from our extensive numerical study presented in Section 4. An alternative way to generate an approximation to the null distribution of $\hat{T}(\mu_0)$ is to use:

$$\hat{T}^*(\mu_0) = \sum_{k=1}^K |\Phi((\mu_0 - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2| \Delta_k. \quad (6)$$

Here, $\{\Delta_k\}$'s are the only random quantities and are analogous to the random multipliers used in the wild bootstrap [Wu (1986)]. The weight from the k -th study is multiplied by Δ_k , which is 1 or -1 with probability 0.5 and is generated by the analyst independently of the observed data. In the Appendix, we also justify the asymptotic validity of the test based on (4) and (6). Confidence intervals for μ can be obtained by inverting this test. In contrast to the existing methods in the literature, the validity of the new proposal does not require the number of studies involved (K) in the analysis to be large. In the Section 4, we show empirically that the new interval estimation procedure performs well even when the sample sizes (n_k) are not large.

The above proposal can be generalized easily to make inferences about certain percentiles of the distribution $G(\cdot)$. Specifically, let us hypothesize that the $100p^{th}$ percentile μ is μ_0 . As for the median case, define $B_k = I(\Theta_k < \mu_0) - I(\Theta_k > \mu_0)$, and \hat{B}_k is obtained by replacing Θ in B_k with $\hat{\Theta}_k$. The test statistic is given by

$$\hat{T}_p(\mu_0) = \sum_{k=1}^K |\Phi((\mu_0 - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2| \hat{B}_k. \quad (7)$$

and the reference distribution is generated by

$$\hat{T}_p^*(\mu_0) = \sum_{k=1}^K |\Phi((\mu_0 - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2| \Delta_k. \quad (8)$$

where $\Delta_k = 1$ with probability p , and $= -1$ with probability $1 - p$. Let the resulting test statistic corresponding to (3) be denoted by $\tilde{T}_p(\mu_0)$. Confidence intervals for the $100p$ th percentile can then be obtained by inverting the conditional test accordingly.

3. Safety Meta Analysis of Erythropoiesis-Stimulating Agents

We re-analyzed the data reported in Bennett et al.(2008) using the new proposal. Here, $K = 52$, and for the k th study, Θ_k was the log-hazard ratio and $\hat{\Theta}_k$ was its estimate. Since

the patient-level data were not available, we approximated the standard error estimate of $\hat{\Theta}_k$ by one fourth of the reported length of the 95% confidence interval (converted to the log-scale), $k = 1, \dots, K$. The 95% confidence interval for the median of the distribution of the random hazard ratio ($\exp(\Theta)$) was (0.94, 1.21) based on the test statistic $\hat{T}(\cdot)$ and (6). The corresponding interval based on the indicator functions $\{I(\hat{\Theta}_k < \mu)\}$ via $\tilde{T}(\cdot)$ was (0.90, 1.26), which was wider than the above interval. The 95% confidence interval for the *mean* of the random effects distribution reported in Bennett et al. (2008) using DL method was (1.01, 1.20). In the next section, we show that the empirical coverage levels of the DL method can be substantially lower than their nominal counterparts even when the number of studies is not that small (say, $K = 40$).

The 95% intervals for the 25th and 75th percentiles based on (7) and (8) were (0.70, 0.99) and (1.18, 1.48), respectively. The counterparts based on $\tilde{T}_p(\cdot)$ were (0.49, 0.93) and (1.25, 1.72). Again, the intervals based on $\hat{T}_p(\cdot)$ were shorter than those with $\tilde{T}_p(\cdot)$. Note that the upper bound of the 95% interval for the 25th percentile was small than 1, which suggested that approximately for a quarter of the study populations, their average hazard ratios for the ESA versus the control were most likely less than one. That is, on average, the patients in these study populations may benefit from taking ESA with respect to mortality.

Further investigation to identify characteristics of these trials would be informative for identifying future cancer patients who would benefit from the ESAs with respect to reduction of blood cell transfusions and quality of life. On the other hand, it is crucial to identify future patients who would have unacceptable toxicity risks.

Bennett et al. (2008) also evaluated the cancer-related anemia with six studies separately (see the top portion of Table 1) and investigated whether ESAs would increase the risk of venous thromboembolism event (VTE) from 38 comparative phase III trials. The results

obtained using the new proposal are reported in the supplemental article [Wang et al. (2009)].

4. Numerical Studies to Evaluate Performance of the New Proposal

We conducted extensive numerical studies to examine the performance of the proposed interval estimation procedure for the *percentiles* of the random effects model under various practical settings. The current existing random-effects methods for meta-analysis have focused on making inferences about the *mean* of the random effects distribution. To the best of our knowledge, there are no methods in the literature that address the same issue as our proposed procedure does. In our numerical studies, we included the commonly used DL interval estimation method, the method proposed by Sidik and Jonkman (2002) (SJ), and the one based on $\tilde{T}(\cdot)$ for comparisons. We considered cases with binary or continuous responses, various symmetric or asymmetric random effects distributions, and a wide range of study sample sizes and number of studies. Based on the results of our numerical investigation, we find that the new proposal performs well with respect to the confidence interval coverage level and length. The DL (or SJ) method tends to be liberal, that is, the empirical coverage levels can be markedly lower than their nominal counterparts. The procedure based on the test statistic $\tilde{T}(\cdot)$ produces confidence intervals whose average lengths are uniformly wider than those with our method. When we deal with percentiles other than median, the method based on $\tilde{T}_p(\cdot)$ may have the under-coverage problem.

Specifically, in our numerical studies, we first considered meta analysis for multiple 2×2 tables under the settings similar to the meta-analysis dataset for VTE rate comparisons in Figure 3 of Bennett et al. (2008). There are 41 studies listed and the raw data are available for 40 studies. We let $\Theta_k = \log(P_{1k}/P_{0k})$ be the log-relative risk for the k th study, where P_{1k} and P_{0k} are the underlying event rates for the ESA and control groups, respectively. We then assumed that the random vector $(\text{logit}(P_{0k}), \text{logit}(P_{1k}))'$ was a random sample with size K from a bivariate normal, whose mean η and variance-covariance matrix

Σ were estimated by their sample counterparts via the observed rates in Table I of the online supplemental article [Wang et al. (2009)]. Note that we used the conventional 0.5 continuity correction for studies with zero cells. The resulting sample means and variance-covariance matrix are $(-3.56, -2.86)'$ and $\begin{pmatrix} 0.90 & 0.62 \\ 0.62 & 1.10 \end{pmatrix}$, respectively. The density of Θ is given in Figure 1 (panel (a)), which appears to be quite symmetric. For each realization $\{(P_{0k}, P_{1k})', k = 1, \dots, K\}$, we generated the corresponding set of 2×2 tables. We then used three aforementioned methods with this realized data set to construct three 95% confidence intervals for the median of the log-relative risk random parameter Θ . For each realized dataset, we excluded studies with 0-0 cells (that is, no events occurred in either group), and used the 0.5 continuity correction for studies with one zero cell. The average empirical coverage levels and the median interval lengths were obtained with 2000 realized data sets. Under the same setting, we repeated this process with different K , the number of studies in our simulated meta analysis. For each K , the sample sizes were chosen from the first K studies listed in Table I of the online supplemental article [Wang et al. (2009)]. The results are summarized in Table 2 (top half). The average coverage levels for our method range from 0.94 to 0.95. On the other hand, the average empirical coverage level can be as low as 0.86 for the DL method, and 0.88 for the SJ method. The median lengths of the intervals obtained via $\hat{T}(\cdot)$ are uniformly shorter than those of the procedure using $\tilde{T}(\cdot)$. In Table 3 (left half), we report the results for the 25th and 75th percentiles. Again our proposal behaves well, but the one with $\tilde{T}_p(\cdot)$ may not have correct coverage level.

We also considered cases with rather asymmetric random effects distribution. For example, we considered a bivariate beta distribution for $\{(P_{0k}, P_{1k})', k = 1, \dots, 40\}$ via three independent gamma random variables which have a common unit scale parameter and shape parameters of 2, 8, and 10, respectively [Olkin and Liu (2003)]. The density function of the random parameter Θ , the log-relative risk, is given in Figure 1 (panel (b)). Under the same

setting as the previous simulation, the results are reported in the bottom half portion of Tables 2 and the right half of Table 3. Again, the new procedure performs well. The DL (or SJ) method still has coverage problem. Note that the DL method provides confidence interval estimates for the mean of $G(\cdot)$, not the median. We then investigated the coverage properties of the DL method for the mean and found that the empirical coverage of the DL intervals was also lower than the nominal level 95% in this setting. For example, when $K = 40$, the coverage for the mean was only 64%.

Although our method is developed assuming that the random effects distribution is continuous, we also considered cases with fixed effects models in our numerical study. For example, we let $(P_{0k}, P_{1k}) = (0.1, 0.2), k = 1, \dots, K$. The results are summarized in Table 4. For this case, the DL method has correct coverage level for most scenarios under which our interval estimation procedure is comparable with the DL method with respect to efficiency, which is reflected in the interval length. We also studied the performance of our method for $\Theta_k = P_{1k} - P_{0k}$, the risk difference for the k th study. The results were very similar to those for the relative risk.

Our numerical studies with continuous responses yield similar results. We summarize the study settings and the results in the supplemental article [Wang et al. (2009)]. We expect similar results for censored time to event observations where hazard ratios are used for treatment effect measurements.

5. Discussion

In this article, we present a simple nonparametric interval estimation procedure for percentiles of the random effects distribution. Random effects meta-analysis are frequently employed in medical research. However, the validity of the most popular method (DL) and its variations [Hardy and Thompson (1996), Biggerstaff and Tweedie (1997), Hartung (1999), Hartung and Knapp (2001a, 2001b), DerSimonian and Kacker (2007)] is not clear

when the number of studies is not large or the parametric assumption for the random effects is violated. An excellent review on meta analysis with the random effects model is given by Sutton and Higgins (2008). In contrast to existing methods, the new proposal does not require that the number of studies is large. The new proposal is valid provided the individual study sample sizes are large.

In addition, if the random effects distribution is symmetric and the *exact* distribution of $\hat{\Theta}_k$, $k = 1, \dots, K$, conditional on Π_k , is symmetric around the unknown fixed realized Θ_k , it is easy to show that the resulting interval estimators based on $\hat{T}(\cdot)$ for the median (or mean) are valid without requiring the sizes of the individual studies or the number of studies to be large. For instance, under the usual two-sample location shift model with continuous response variable, let Θ be the location shift parameter of interest. Then, the two-sample rank estimator $\hat{\Theta}$ is symmetric around Θ under rather mild conditions [Lehmann (1975), p. 86]. If the unspecified random effects distribution is symmetric around μ , one can use our procedure to obtain exact confidence intervals for μ . We conducted a simulation study to examine the performance of the method in this setting and the study was described in detail in the supplemental article [Wang et al. (2009)].

The proposed procedure can be implemented with study level summary statistics. When patient level data are available, various novel procedures have been studied for the mixed effects regression models for continuous, discrete or censored event time observations [Laird and Ware (1982), Hougaard (1995), Hogan and Laird (1997), Henderson et al. (2000), Lam, Lee and Leung (2002), Nelder, Lee and Pawitan (2006), Cai, Cheng and Wei (2002), Zeng and Lin (2007), Zeng, Lin and Lin (2008)]. To the best of our knowledge, all the existing asymptotic procedures for mixed effects models assume that the number of studies is large.

Under the current practice of conducting meta analysis, inferences are made only for the “center” of the random effects distribution. A conclusion on the risk or benefit from an intervention solely based on an estimated center of the random effects distribution provides

limited information and is usually not sufficient. If the number of studies involved is not small, we highly recommend estimating this distribution or its percentiles as proposed in this article.

Under the fixed effects modeling setting, this distribution has a single unknown mass point. The standard estimation procedure for such a fixed parameter value utilizes a weighted average of study-specific point estimates. For analyzing multiple 2×2 tables, the most commonly used procedures are Mantel-Haenszel [Mantel and Haenszel (1959)] and Peto methods [Yusuf et al. (1985)]. These methods are valid when the number of studies and each individual study sample size are large. Moreover, for the case that the event rate is small, these standard methods may not perform well. Recently, Tian et al. (2009) proposed a general exact interval estimation procedure under the fixed effects model, which combines study-specific exact confidence intervals instead of point estimates across the studies. If the fixed effects model is approximately correct, the existing interval procedures for the common parameter value μ may be more efficient than those developed under the random effects model. The standard heterogeneity tests generally do not have power to detect violation of the fixed effects modeling assumption. Therefore, in practice, sensitivity analyses with both random and fixed effects models are highly recommended.

Acknowledgements. We thank Professor Michael Newton and a referee for their comments which have led to an improved version of the paper.

REFERENCES

- Bennett, C.L., Silver, S.M., Djulbegovic, B., Samaras, A.T., Blau, C.A., Gleason, K.J., Barnate, S.E., Elverman, K.M., Courtney, D.M., MeKoy, J.M., Edwards, B.J., Tigue, C.C., Raisch, D.W., Yarnold, P.R., Dorr, D.A., Kuzel, T.M., Tallman, M.S., Trifilio, S.M., West, D.P., Lai, S.Y., Henke, M. (2008). Venous Thromboembolism and Mortality Associated with Recombinant Erythropoietin and Darbepoetin Administration for the Treatment of Cancer-Associated Anemia, *Journal of the American Medical Association* **299**, 914-924.
- Biggerstaff, B.J., and Tweedie, R.L. (1997), Incorporating Variability in Estimates of Heterogeneity in the Random Effects Model in Meta-analysis. *Statistics in Medicine* **16**, 753-768.
- Bohning, D., Malzahn, U., Dietz, E., and Schlattmann, P. (2002). Some General Points in Estimating Heterogeneity Variance with the DerSimonian-Laird Estimator. *Biostatistics* **3**, 445-457.
- Brockwell, S.E., and Gordon, I.R. (2001). A Comparison of Statistical Methods for Meta-Analysis. *Statistics in Medicine* **20**, 825-840.
- (2007). A Simple Method for Inference on an Overall Effect in Meta-analysis. *Statistics in Medicine* **26**, 4531-4543.
- Cai, T., Cheng, S. C., and Wei, L. J. (2002). Semiparametric Mixed-effects Models for Clustered Failure Time Data. *Journal of the American Statistical Association* **97**, 514-522.
- DerSimonian, R., and Laird, N.M. (1986). Meta-analysis in Clinical Trials, *Controlled Clinical Trials* **7**, 177-188.
- DerSimonian, R., and Kacker, R. (2007). Random-effects Model for Meta-Analysis of Clinical Trials: An Update. *Contemporary Clinical Trials* **28**, 105-114.
- Hardy, R.J., and Thompson, S.G. (1996). A Likelihood Approach to Meta-Analysis with Random Effects. *Statistics in Medicine* **15**: 619-629.

- Hardy, R.J., and Thompson, S.G. (1998). Detecting and Describing Heterogeneity in Meta-analysis. *Statistics in Medicine* **17**, 841-856.
- Hartung, J. (1999). An Alternative Method for Meta-Analysis. *Biometrical Journal* **8**, 901-916.
- Hartung, J., and Knapp, G. (2001a). A Refined Method for the Meta-Analysis of Controlled Clinical Trials with Binary Outcome. *Statistics in Medicine* **20**, 3875-3889.
- (2001b). On Tests of the Overall Treatment Effect in Meta-Analysis with Normally Distributed Responses. *Statistics in Medicine* **20**, 1771-1782.
- Henderson, R., Diggle, P., and Dobson, A. (2000). Joint Modelling of Longitudinal Measurements and Event Time Data, *Biostatistics (Oxford)*, **1**: 465-480.
- Hogan, J.W. and Laird, N.M. (1997). Mixture Models For The Joint Distribution of Repeated Measures And Event Times, *Statistics in Medicine*, **16**: 239-257.
- Hougaard, P. (1995), Frailty models for survival data, *Lifetime Data Analysis*, **1**, 255-273.
- Laird, N.M. and Ware, J.H. (1982). Random Effects Models for Longitudinal Data. *Biometrics* **38**, 963-974.
- Lam, K. F., Lee, Y. W., and Leung, T. L. (2002). Modeling Multivariate Survival Data by a Semiparametric Random Effects Proportional Odds Model, *Biometrics*, **58**: 316-323.
- Lehmann, E.L. (1975). Nonparametrics: Statistical Methods Based on Ranks. San Francisco, Holden-Day.
- Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of National Cancer Institution* **22**, 719-748.
- Nelder, J.A., Lee, Y., and Pawitan, Y. (2006). Generalized Linear Models with Random Effects: A Unified Approach via H-likelihood, London, Chapman and Hall.
- Olkin, I., and Liu, R. (2003). A Bivariate Beta Distribution. *Statistics & Probability Letters* **62**, 407-412.

- Sidik K, and Jonkman J.N. (2002). A Simple Confidence Interval for Meta-analysis. *Statistics in Medicine* **21**, 3153-3159.
- (2007). A Comparison of Heterogeneity Variance Estimators in Combining Results of Studies. *Statistics in Medicine* **26**, 1964-1981.
- Sutton, A.J. and Higgins, J.P. (2008). Recent Developments in Meta-Analysis. *Statistics in Medicine* **27**, 625-650.
- Tian L., Cai T., Pfeffer M.A., Piankov N., Cremieux P., and Wei L.J. (2009). Exact and Efficient Inference Procedure for Meta-Analysis and Its Application to the Analysis of Independent 2×2 Tables with All Available Data But Without Artificial Continuity Correction. *Biostatistics*, **10**, 275-281.
- Wang, R., Tian L., Cai T., Wei, L.J. (2009). Supplement to “Nonparametric Inference Procedure for Percentiles of the Random Effects Distribution in Meta Analysis”.
- Wu, C.F.J. (1986). Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis. *The Annals of Statistics* **14**, 1261-1295.
- Yusuf, S., Peto, R., Lewis, J., Collins, R., Sleight, P. et al (1985). Beta Blockade During and After Myocardial Infarction: An Overview of the Randomised Trials. *Progress in Cardiovascular Diseases* **27**, 335-371.
- Zeng, D., and Lin, D. Y. (2007). Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **69**, 507-564.
- Zeng, D., Lin D.Y., Lin X. (2008). Semiparametric Transformation Models with Random Effects for Clustered Failure Time Data. *Statistica Sinica* **18**, 355-377.

APPENDIX

Justification for the validity of the conditional permutation test $\hat{T}(\cdot)$ based on the approximation generated by $\hat{T}^*(\cdot)$

Let $D_k = |\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2| \hat{B}_k - B_k/2$. We show that D_k goes to 0, in probability, as $n_k \rightarrow \infty$. Here, the probability is generated by the random element (X_k, Π_k) . For any fixed positive constant c , first we show that $\text{pr}(|D_k| \geq c \mid \Pi_k) \rightarrow 0$ for any given Π_k with $\Theta_k \neq \mu$. To this end, consider two cases. First, if $\Theta_k < \mu$, then conditional on Π_k ,

$$|D_k| = |\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1| = 1 - \Phi((\mu - \Theta_k)/\hat{\sigma}_k + (\Theta_k - \hat{\Theta}_k)/\hat{\sigma}_k).$$

As $n_k \rightarrow \infty$, $(\mu - \Theta_k)/\hat{\sigma}_k \rightarrow \infty$ in probability, and $(\Theta_k - \hat{\Theta}_k)/\hat{\sigma}_k \rightarrow N(0, 1)$ in distribution. Therefore, for any $c > 0$, we can find N such that when $n_k > N$, $\text{pr}((\mu - \Theta_k)/\hat{\sigma}_k + (\Theta_k - \hat{\Theta}_k)/\hat{\sigma}_k \leq \Phi^{-1}(1 - c)) < c$, which is equivalent to $\text{pr}(\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) < 1 - c) = \text{pr}(|D_k| \geq c) < c$. Therefore, $\text{pr}(|D_k| \geq c \mid \Pi_k) \rightarrow 0$. Similarly if $\Theta_k > \mu$, we can show that $\text{pr}(|D_k| \geq c \mid \Pi_k) \rightarrow 0$ as $n_k \rightarrow \infty$. Therefore, $\text{pr}(|D_k| \geq C \mid \Pi_k) \rightarrow 0$ for any Π_k such that $\Theta_k \neq \mu$.

These, coupled with the fact that $G(\cdot)$ is continuous, implies that $\text{pr}(|D_k| \geq c) = \text{E}_{\Pi_k} \{\text{pr}(|D_k| \geq c \mid \Pi_k)\} \rightarrow 0$ for any c by the dominate convergence theorem. Therefore, $D_k \rightarrow 0$ in probability as $n_k \rightarrow \infty$. It follows that $|\hat{T}(\mu) - \sum_{k=1}^K B_k/2| \rightarrow 0$, in probability, as $\min\{n_1, \dots, n_K\} \rightarrow \infty$.

Similarly, since

$$\left| |\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2| \Delta_k - |I(\Theta_k < \mu) - 1/2| \Delta_k \right| \leq |D_k|,$$

one can show that $\hat{T}^*(\mu) - \sum_{k=1}^K |I(\Theta_k < \mu) - 1/2| \Delta_k \rightarrow 0$, in probability as $\min\{n_1, \dots, n_K\} \rightarrow \infty$.

∞ , where

$$\Delta_k = \begin{cases} 1 & \text{with prob. } p \\ -1 & \text{with prob. } 1 - p, \end{cases}$$

for the 100 p th percentile and is independent of the data. Therefore, for any t and positive c ,

$$\Pr_{\{(X_k, \Pi_k)_{k=1, \dots, K}\}} \left(\left| \Pr(\hat{T}^*(\mu) \leq t | (X_k, \Pi_k)_{k=1, \dots, K}) - \Pr\left(\sum_{k=1}^K \Delta_k/2 \leq t\right) \right| \geq c \right) \leq c,$$

when $\min\{n_1, \dots, n_K\}$ is large. This, coupled with the fact that $\sum_{k=1}^K B_k/2 \sim \sum_{k=1}^K \Delta_k/2$ under the null hypothesis that the 100 p th percentile of Θ_k is μ , implies that one can approximate the null distribution of $\hat{T}(\mu)$ by the distribution of $\hat{T}^*(\mu)$ conditional on the observed data.

Table 1: Study-level Summary Statistics for Mortality for Cancer Studies with ESAs vs Control from Bennett et al (2008)

	Study	Two Sample Hazard Ratio	
		Point Estimate	95% Confidence Interval
Anemia of Cancer	Mystakidou et al, 2005	0.50	(0.05-4.99)
	Gordon et al, 2006	0.67	(0.23-2.00)
	Abels, 1993	0.89	(0.41-1.93)
	Charu et al, 2007	1.38	(0.44-4.33)
	Glaspy et al, 2007	1.43	(1.06-1.92)
Treatment-Related Anemia	Smith et al, 2003	3.96	(0.29-54.12)
	Throuvalas et al, 2000	0.13	(0-332.66)
	Dunphy et al, 1999	0.14	(0-6.88)
	Vadhan-Raj et al, 2004	0.15	(0-415.90)
	Dammacco et al, 2001	0.32	(0.11-0.95)
	Del Mastro et al, 1997	0.36	(0.05-2.56)
	Cazzola et al, 1995	0.37	(0.06-2.27)
	P-174, 2004	0.41	(0.03-5.76)
	Thatcher et al, 1999	0.49	(0.03-8.71)
	Kotasek et al, 2003	0.55	(0.11-2.71)
	Oberhoff et al, 1998	0.61	(0.24-1.55)
	Blohmer et al, 2003 (AGO/NOGG)	0.67	(0.34-1.33)
	Henry and Abels, 1994	0.75	(0.28-2.01)
	Vansteenkiste et al, 2002	0.78	(0.60-1.01)
	Littlewood et al, 2001	0.81	(0.62-1.06)
	Taylor et al, 2005 (DA 232)	0.85	(0.45-1.60)
	EPO-CAN-17, 2007	0.88	(0.49-1.59)
	Amgen DA 145, 2007	0.93	(0.82-1.05)
	Razzouk et al, 2004	0.98	(0.14-6.90)
	Savonije et al, 2004	0.98	(0.36-2.67)
	ten Bokkel Huinink et al, 1998	1.01	(0.19-5.31)
	Osterborg et al, 1996	1.02	(0.51-2.04)
	Coiffier et al, 2001	1.02	(0.38-2.73)
	Debus et al, 2007(EPO-GER-22)	1.02	(0.60-1.74)
	Osterborg et al, 2005	1.04	(0.80-1.35)
	EPO-GBR-7, 2007	1.07	(0.73-1.57)
	Case et al, 1993	1.08	(0.44-2.66)
	Witzig et al, 2005	1.09	(0.83-1.43)
	Moebus et al, 2007	1.14	(0.77-1.69)
	Strauss et al, 2007	1.16	(0.69-1.95)
	Thomas et al, 2007 (GOG-191)	1.25	(0.65-2.41)
	Thatcher et al, 1999	1.26	(0.24-6.60)
	Overgaard et al, 2007 (DAHANCA 10)	1.28	(0.97-1.69)
Hedenus et al, 2003	1.36	(0.98-1.89)	
Leyland-Jones et al, 2005 (INT-76)	1.37	(1.07-1.75)	
Henke et al, 2003	1.39	(1.05-1.84)	
Machtay et al, 2007 (RTOG 99-03)	1.41	(0.80-2.49)	
PREPARE, 2007	1.50	(0.96-2.34)	
Grote et al, 2005 (N93-004)	1.53	(0.65-3.61)	
INT-3, 2004	1.56	(0.42-5.79)	
INT-1, 2004	1.58	(0.32-7.82)	
Rose et al, 1994	1.68	(0.66-4.29)	
Bamias et al, 2003	1.80	(0.53-6.12)	
Wright et al, 2007 (EPO-CAN-20)	1.84	(1.01-3.35)	
EPO-CAN-15, 2004	2.70	(1.17-6.23)	
Wilinson et al, 2006	4.54	(0.40-51.20)	
O'Shaughnessy et al, 2005	7.39	(0.15-366.10)	

Table 2: Empirical Coverage Levels (ECL) And Median Lengths (ML) For 0.95 Interval Estimates For Median Based On DerSimonian-Laird (DL), Sidik and Jonkman (SJ), $\hat{T}(\cdot)$ and $\tilde{T}(\cdot)$ With A Bivariate Logit-Normal Or A Bivariate Beta Distribution For The Two Underlying Random Event Rates

Bivariate Logit-Normal				
Number of	DL	SJ	Proposed Method ($\hat{T}(\cdot)$)	$\tilde{T}(\cdot)$
Studies K	ECL, ML	ECL, ML	ECL, ML	ECL, ML
40	86%, 0.62	88%, 0.65	94%, 0.72	95%, 0.90
30	88%, 0.71	91%, 0.75	94%, 0.83	95%, 1.03
20	88%, 0.85	91%, 0.90	94%, 1.00	95%, 1.23
10	88%, 1.18	94%, 1.36	95%, 1.54	97%, 2.15
6	91%, 1.57	97%, 2.06	95%, 2.29	97%, 2.89
Bivariate Beta				
	DL	SJ	Proposed Method ($\hat{T}(\cdot)$)	$\tilde{T}(\cdot)$
Studies K	ECL, ML	ECL, ML	ECL, ML	ECL, ML
40	87%, 0.40	89%, 0.42	95%, 0.52	96%, 0.65
30	88%, 0.46	90%, 0.48	95%, 0.61	96%, 0.75
20	90%, 0.55	92%, 0.59	96%, 0.75	96%, 0.91
10	91%, 0.76	93%, 0.89	96%, 1.10	98%, 1.56
6	88%, 1.00	94%, 1.30	95%, 1.58	97%, 2.10

Table 3: Empirical Coverage Levels (ECL) and Median Lengths (ML) for 0.95 Confidence Intervals For The 25th And 75th Percentiles Based On $\hat{T}_p(\cdot)$ and $\tilde{T}_p(\cdot)$ With A Bivariate Logit-Normal Or A Bivariate Beta Distribution For The Two Underlying Random Event Rates

	Bivariate Logit-Normal				Bivariate Beta			
	p25		p75		p25		p75	
	$\hat{T}_p(\cdot)$	$\tilde{T}_p(\cdot)$	$\hat{T}_p(\cdot)$	$\tilde{T}_p(\cdot)$	$\hat{T}_p(\cdot)$	$\tilde{T}_p(\cdot)$	$\hat{T}_p(\cdot)$	$\tilde{T}_p(\cdot)$
Number of Studies K	ECL, ML	ECL, ML	ECL, ML	ECL, ML	ECL, ML	ECL, ML	ECL, ML	ECL, ML
40	95%, 0.86	86%, 1.16	95%, 0.81	92%, 0.92	96%, 0.48	93%, 0.55	96%, 0.73	92%, 0.96
35	96%, 0.91	88%, 1.21	96%, 0.86	90%, 1.02	96%, 0.52	95%, 0.61	96%, 0.78	93%, 1.04
30	96%, 1.00	90%, 1.37	96%, 0.94	91%, 1.12	95%, 0.56	94%, 0.64	96%, 0.85	93%, 1.07
25	96%, 1.12	90%, 1.49	97%, 1.06	92%, 1.23	96%, 0.62	93%, 0.65	96%, 0.94	92%, 1.10
20	96%, 1.24	92%, 1.52	97%, 1.16	92%, 1.32	96%, 0.72	95%, 0.80	96%, 1.37	95%, 1.37

Table 4: Empirical Coverage Levels (ECL) And Median Lengths (ML) For 0.95 Interval Estimates For Median Based On DerSimonian-Laird (DL), $\hat{T}(\cdot)$ and $\tilde{T}(\cdot)$ With A Fix Effect Model (The Underlying Event Rates Are 0.1 and 0.2)

Number of Studies K	DL ECL, ML	Proposed Method ($\hat{T}(\cdot)$) ECL, ML	$\tilde{T}(\cdot)$ ECL, ML
$K = 40$	92%, 0.24	95%, 0.27	96%, 0.35
$K = 30$	94%, 0.26	95%, 0.30	96%, 0.39
$K = 20$	95%, 0.30	95%, 0.35	97%, 0.45
$K = 10$	97%, 0.47	96%, 0.57	98%, 0.84
$K = 6$	96%, 0.75	95%, 1.03	97%, 1.34

Figure 1: The True Density Functions For The Random Log-Relative-Risk Parameter For The Simulation Study

