Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 79

Nonparametric Inference Procedure For Percentiles of the Random Effect Distribution in Meta Analysis

Rui Wang* Lu Tian†
Tianxi Cai‡ L. J. Wei**

*Harvard University, rwang@hsph.harvard.edu
†Northwestern University, lutian@northwestern.edu
‡Harvard University, tcai@hsph.harvard.edu
**Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper79

Nonparametric Inference Procedure for Percentiles of the Random Effect Distribution in Meta Analysis

Rui Wang

Department of Biostatistics, Harvard University, Boston, MA 02115

rwang@hsph.harvard.edu

Lu Tian

Department of Preventive Medicine, Northwestern University, Chicago, IL 60611

lutian@northwestern.edu

Tianxi Cai and L.J. Wei

Department of Biostatistics, Harvard University, Boston, MA 02115

tcai@hsph.harvard.edu *wei@hsph.harvard.edu*

SUMMARY. Suppose that the random effect distribution of the parameter of interest in meta analysis is completely unspecified. We propose a simple nonparametric interval estimation procedure for making inferences about the percentiles, for example, the median, of this distribution. Regardless of the number of studies involved, the new proposal is theoretically valid provided that the individual study sample sizes are large. Based on an extensive numerical study, we find that the procedure performs well even with moderate study sample sizes. The new method can be implemented with study-level summary statistics. The proposal is illustrated with the data from a recent meta analysis to investigate the potential treatment-related toxicity from erythropiesis-stimulating agents (ESAs) for treating cancer patients with anemia (Bennett et al., 2008).

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

1. Introduction

Meta analysis techniques have been utilized frequently to make inferences about the distribution of an unobservable random parameter. Under the fixed effect 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). It is important to note that these methods are valid asymptotically, that is, 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. (2008) proposed a general exact interval estimation procedure under the fixed effect model, which combines study-specific exact confidence intervals instead of point estimates across the studies.

Often the fixed effect modeling assumption may not be realistic. The study-specific parameter values may vary markedly across sub-groups of the population. Under the random effect model, the procedure for estimating the *mean* of the random effect distribution proposed by DerSimonian and Laird (DL) (1986) is commonly used in practice. Their method utilizes a linear combination of study-specific estimates of the parameter 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 sample sizes and the number of studies (Brockwell and Gordon, 2001; Bohning et al., 2002; Sidik and Jonkman, 2007; Viechtbauer, 2007). Recently, various novel alternatives or modifications to the DL method have been proposed, for example, by Hardy and Thompson (1996), Biggerstaff and Tweedie (1997), Hartung (1999), Hartung and Knapp (2001a, 2001b) and DerSimonian and Kacker (2007). Their validity, however, is not clear when the number of studies is not large or the parametric assumption for the random effect is violated. An excellent review on meta analysis with the random

effect model is given by Sutton and Higgins (2008).

In this article, we propose a simple interval estimation procedure for the percentiles of the random effect distribution based on study-level data without assuming a parametric form of the distribution. When the random effect distribution is not symmetric, its median may be a better measure as the center than the mean for which the DL and its generalizations try to estimate. Moreover, our inference procedure may provide information beyond the center of the distribution of the random parameter if the number of studies is not too small. Regardless of the number of studies involved in the analysis, the new proposal is theoretically valid when the sample sizes of individual studies are large. An extensive numerical study was conducted to examine the performance of the new proposal under various practical settings. We find that our interval estimation procedure for percentiles of the random effect distribution has correct coverage level. Moreover, when the empirical level of the DL method is close to its nominal counterpart, our corresponding interval estimate has comparable length. We illustrate the method with the data from a recent meta analysis for evaluating potential treatment-related toxicity of erythropiesis-stimulating agents (ESAs) for treating cancer patients with anemia (Bennett et al., 2008). For some cases, our results are markedly different from those reported in Bennett et al.

It is important to note that 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 asymptotical procedures for mixed effects models assume that

the number of studies is large.

2. Interval Estimates for Percentiles of the Random Effect 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 $\{\Pi_k, X_k\}, k = 1, \cdots, K$, be K independent copies of $\{\Pi, X\}$. The problem is how to make inferences, for instance, about the median μ_0 of $G(\cdot)$ with $\{X_k, k = 1, \cdots, K\}$. 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 kth table. Here, the nuisance parameter Λ_k consists of the underlying event rate for the "control" group and the sample size for the kth study.

If we can observe $\{\Theta_k, k = 1, \cdots, K\}$, a simple nonparametric estimator for μ_0 is the sample median. Exact confidence intervals for μ_0 can be obtained by inverting a sign test statistic

$$T(\mu) = \sum_{k=1}^{K} \{I(\Theta_k < \mu) - 1/2\}, \tag{1}$$

for testing the null hypothesis that the median μ_0 is μ, where $I(\cdot)$ is the indicator function. The null distribution of $T(\mu_0) + K/2$ is a binomial with size K and "success" probability of $1/2$.

Suppose that given Π_k, $\hat{\Theta}_k$ is a consistent estimator for Θ_k based on the data X_k. Also, suppose that the distribution of $\hat{\Theta}_k$ is approximately normal with mean Θ_k and variance $\hat{\sigma}_k^2$, $k = 1, \cdots, K$. To obtain confidence intervals for μ_0 without observing Θ_k directly, one may replace Θ_k with $\hat{\Theta}_k$ in the test statistic $T(\cdot)$. Let the corresponding test statistic be denoted by $\tilde{T}(\cdot)$. When the sample size n_k for the kth study, $k = 1, \cdots, K$, is large, the unconditional null distribution of $\tilde{T}(\mu_0) + K/2$ is approximately binomial with size K and

success rate of 0.5. Unfortunately the variable $I(\hat{\Theta}_k < \mu)$ may not be a good surrogate for $I(\Theta_k < \mu)$ especially when the variance $\hat{\sigma}_k^2$ of $\hat{\Theta}_k$ is large with respect to the distance between the unobserved realized Θ_k and μ. For this case, the chance of observing the event, $\{\hat{\Theta}_k < \mu\}$ can be very close to 1/2 (like tossing a fair coin), and the noise generated from such an unstable variable $\{I(\hat{\Theta}_k < \mu) - 1/2\}$ may well out-weight its added value to the power of the test based on $\tilde{T}(\mu)$. An alternative approach is to replace the indicator $I(\hat{\Theta}_k < \mu)$ with a measure of likelihood for the event, $\Theta_k < \mu$, for example, the observed coverage level of the interval $(-\infty, \mu)$ for the realized Θ_k. Under the present setting, this coverage level is approximately $\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k)$, where $\Phi(\cdot)$ is the distribution function of the standard normal. The resulting test statistic is

$$\hat{T}(\mu) = \sum_{k=1}^{K} \{\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2\}. \tag{2}$$

Note that if for the kth study, the data X_k are quite informative for the event, $\Theta_k < \mu$, that is, the coverage level of $(-\infty, \mu)$ for Θ_k is close to either one or zero, this study carries heavy weight in the statistic $\hat{T}(\mu)$.

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

$$\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - I(\Theta_k < \mu) \to 0, \quad \text{as} \quad n_k \to \infty. \tag{3}$$

Therefore, for fixed K, for large $n_k, k = 1, \cdots, K$, the distribution of $\hat{T}(\mu)$ can be approximated by that of $T(\mu)$. This approximation, however, is rather discrete and for moderate sample sizes, the resulting confidence intervals for μ_0 do not have adequate coverage levels based on our extensive numerical study discussed in Section 4. Moreover, this limiting null distribution is generated by weighting all K studies equally. Now, from (3), asymptotically, $\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k)$ is symmetric around 1/2. This motivates us to consider the following

procedure to generate an approximation to the null distribution of $\hat{T}(\mu)$. First, let

$$T^*(\mu) = \sum_{k=1}^{K} |\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2|(2\Delta_k - 1), \tag{4}$$

where $\{\Delta_k, k = 1, \cdots, K\}$ is a random sample, independent of the data, from a Bernoulli with success probability of 1/2. Then, conditional on the data, we repeatedly generate M realizations of $\{\Delta_k\}$ to obtain the null distribution. A relatively large or small observed value of $\hat{T}(\mu)$ compared with this generated reference distribution suggests a rejection of the hypothesized median value μ of $G(\cdot)$. In the Appendix, we justify the asymptotic validity of the test based on (2) and (4). Confidence intervals for μ_0 can be obtained by inverting this test. In contrast to the existing methods in the literature, the new proposal is valid with any fixed K, the number of studies involved in the analysis. In the Section 4, we show empirically that the resulting interval estimation procedure performs well even when the sample sizes $\{n_k\}$ are not large.

When the number of studies is not small, the above proposal can be easily generalized to the case for making inferences about certain percentiles of the distribution $G(\cdot)$. Specifically, let us hypothesize that the $100p^{th}$ percentile μ_0 is μ. The test statistic is then

$$\hat{T}_p(\mu) = \sum_{k=1}^{K} \{\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2\}. \tag{5}$$

Note that we use 1/2 in the above $\{\cdot\}$ instead of p. The rationale is that for the kth study, the term $\Phi(\cdot)$ is a measure of likelihood that the unknown fixed Θ_k is less than μ. If this likelihood is around 1/2, we give this study very little weight in (5). Unconditionally, $\hat{T}_p(\mu_0) + K/2$ is asymptotically binomial with mean of Kp and size of K. Again, like the case with median, this limiting distribution puts equal weights for all individual studies and therefore may not be a good approximation to the null distribution of (5) for finite sample

sizes $\{n_k, k = 1, \cdots, K\}$. Instead, conditional on the data, an approximation to the null distribution can be generated via the random quantity

$$\hat{T}_p(\mu) = \sum_{k=1}^{K} |\Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - 1/2|(2\epsilon_k - 1), \tag{6}$$

where $\{\epsilon_k, k = 1, \cdots, K\}$ is a random sample from a Bernoulli with success probability of p. Confidence intervals for the $100p$th percentile can then be obtained by inverting this conditional permutation test accordingly. For inference about the pth percentile, we let the counterpart of $\tilde{T}(\cdot)$ be denoted by $\tilde{T}_p(\cdot)$

Note that the likelihood measure of the event $\Theta_k < \mu$ does not have to be based on the large sample normal approximation to the distribution of $\hat{\Theta}_k$. For example, in dealing with multiple 2×2 tables with rare events, often there are studies without any events. For this case, the variance of $\hat{\Theta}_k$, say, the risk difference estimator, cannot be estimated without using continuity correction. On the other hand, for each study, one may construct the exact confidence intervals for Θ and use the exact level for the interval $(-\infty, \mu)$ to replace $\Phi(\cdot)$ in (2), (4), (5) and (6) for testing the hypothesized percentile value. Lastly, to avoid the conservativeness of the test, for example, based on $\tilde{T}(\cdot)$, due to its discrete null distribution, we used the standard mid p-value method to implement the test procedure.

3. Examples

We use two examples to illustrate the new proposal. The first one is with survival time observations and the second example deals with multiple 2×2 tables. In a recent meta analysis, Bennett et al. (2008) examined whether the erythropoiesis-stimulating agents (ESAs) for treating cancer-related anemia would increase the patient's risk of mortality with the study-level data from 51 phase III comparative trials (ESA vs. placebo or standard of care). In Table 1 (copied from Figure 2 of Bennett et al.), we list their two-sample hazard ratio

point and 95% interval estimates. Here, $K = 51$, and for the kth study, Θ_k is the log-hazard ratio and $\hat{\Theta}_k$ is its estimate. Since the patient-level data are not available, we approximate 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), $k = 1, \cdots, K$. The 95% confidence interval for the median of the distribution of the random hazard ratio ($\exp(\Theta)$) are $(0.94, 1.21)$ based on the test statistic $\hat{T}(\cdot)$ and (4). The corresponding interval based on the indicator functions $\{I(\hat{\Theta}_k < \mu)\}$ via $\tilde{T}(\mu)$ is $(0.90, 1.26)$, which is wider than the above interval. Note that the 95% confidence interval for the *mean* of the random effects distribution reported in Bennett et al. (2008) using DL method is $(1.01, 1.20)$, which excludes the null value one, indicating that mortality for ESAs is significantly higher than that for the control group. On the other hand, our results are not statistically significant.

The 95% intervals for the 25th and 75th percentiles based on (5) and (6) are $(0.70, 0.99)$ and $(1.18, 1.48)$, respectively. The counterparts based on $\tilde{T}_p(\cdot)$ are $(0.49, 0.93)$ and $(1.25, 1.72)$. Again, the intervals based on $\hat{T}_p(\cdot)$ are shorter than those with $\tilde{T}_p(\cdot)$.

The cancer-related anemia case was also evaluated by Bennett et al. (2008) with six studies (see the top portion of Table 1). The 95% confidence intervals based on $\hat{T}(\cdot)$, $\tilde{T}(\cdot)$ and DL method are $(0.55, 1.94), (0.50, 3.96)$ and $(1.00, 1.67)$, respectively. Note that the DL interval is quite different from ours. The null value well sits in our interval, but not in theirs. 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, $K = 40$).

Bennett et al. (2008) also investigated whether ESAs would increase the risk of venous thromboembolism event (VTE) from 38 comparative phase III trials (Table 2, copied from Figure 3 of Bennett et al.). Note that there are 41 studies listed in Table 2. Since three studies do not have any VTE events, as Bennett et al. did, our analysis is based on the study-level data from 38 trials. Here, we let Θ be the log-relative-risk (not the relative-risk). Again,

we use their 38 sets of point and 95% interval estimates for the relative risk to construct 95% confidence intervals for the median of the random relative risk based on $\hat{T}(\cdot)$ and $\tilde{T}(\cdot)$. These intervals are $(1.54, 2.53)$ and $(1.52, 3.00)$ respectively. The corresponding interval based on the DL method is $(1.31, 1.87)$. The 95% intervals for the 25th and 75th percentiles for the relative risk distribution based on $\hat{T}_p(\cdot)$ in (5) are $(1.05, 1.72)$ and $(2.09, 3.97)$, respectively. The counterparts based on the indicators $\tilde{T}_p(\cdot)$ are $(0.75, 1.84)$ and $(2.80, 4.93)$.

4. Numerical Study to Evaluate Performance of the New Proposal

We conducted an extensive numerical study to examine the performance of the proposed interval estimation procedure for the percentiles of the random effect model under various practical settings. For comparisons, we included the DL interval estimation method and the one based on $\tilde{T}(\cdot)$. In our study we considered cases with binary or continuous responses, various symmetric or asymmetric random effect 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 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.

More specifically, in one part of our numerical study, we considered meta analysis for multiple 2×2 tables under the settings similar to that presented in Table 2. In Table 2 for the VTE rate comparisons, 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 kth 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 2 (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 $\{(P_{0k}, P_{1k})', k = 1, \cdots, K\}$, 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 2. The results are summarized in Table 3 (left half). The average coverage levels for our method range from 0.94 to 0.95. On the other hand, the average empirical coverage level for the DL method can be as low as 0.86. The median lengths of our method are uniformly shorter than those of the procedure using $\tilde{T}(\cdot)$. In Table 4 (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 effect distribution. For example, we consider a bivariate beta distribution for $\{(P_{0k}, P_{1k})', k = 1, \cdots, 40\}$ 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 right half portions of Tables 3

and 4. Again, the new procedure performs well. The DL 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 were also lower than the nominal level 95% in this setting. For example, when $K = 40$, the coverage for the mean is only 64%.

Although our method is developed assuming that the random effect distribution is continuous, we also considered cases with fixed effect models in our numerical study. For example, we let $(P_{0k}, P_{1k}) = (0.1, 0.2), k = 1, \cdots, K$. The results are summarized in Table 5. 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 the interval length. We also studied the performance of our method for $\Theta_k = P_{1k} - P_{0k}$, the risk difference for the kth study. The results were very similar to those for the relative risk.

5. Remarks

In this article, we present a simple nonparametric interval estimation procedure for percentiles of the random effect model. In contrast to existing methods, the new proposal does not require that the number of studies is large. Moreover, if the random effect distribution is symmetric and the *exact* distribution of $\hat{\Theta}_k$, $k = 1, \cdots, K$, 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 regardless the sizes of the individual studies or the number of studies. 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 $\hat{\Theta}_k$ is symmetric around Θ_k under rather mild conditions (Lehmann, 1975 , p. 86). If the unspecified random effect distribution is symmetric around μ_0, one can use our procedure to obtain exact confidence intervals for μ_0.

Note that if the fixed effect model is approximately correct, the existing interval proce-

dures for the common parameter value μ_0 may be more efficient than those developed under the random effect model. The standard heterogeneity tests genearlly do not have power to detect inadequacy of the fixed effect modeling assumption. Therefore, in practice, sensitivity analysis with both random and fixed effect models is highly recommended.

REFERENCES

Bennett, C.L., Silver, S.M., Djulbegovie, 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* **229**, 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 Measure ments 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. (2008). 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*, in Revision.

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 $T^*(\cdot)$

Let $D_k = \Phi((\mu - \hat{\Theta}_k)/\hat{\sigma}_k) - I(\Theta_k < \mu)$. We show that D_k goes to 0, in probability, as $n_k \to \infty$. Here, the probability is generated by the random element (X_k, Π_k). For any fixed positive constant c, first we show that $\text{pr}(|D_k| \geq c \mid \Pi_k) \to 0$ 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 $n_k \to \infty$, $(\mu - \Theta_k)/\hat{\sigma}_k \to \infty$ in probability, and $(\Theta_k - \hat{\Theta}_k)/\hat{\sigma}_k \to N(0,1)$ 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, $\text{pr}(|D_k| \geq c \mid \Pi_k) \to 0$. Similarly if $\Theta_k > \mu$, we can show that $\text{pr}(|D_k| \geq c \mid \Pi_k) \to 0$ as $n_k \to \infty$. Therefore, $\text{pr}(|D_k| \geq C \mid \Pi_k) \to 0$ for any Π_k such that $\Theta_k \neq \mu$. These, coupled with the fact that $G(\cdot)$ is continuous, implies that $\text{pr}(|D_k| \geq c) = \text{E}_{\Pi_k}\{\text{pr}(|D_k| \geq c \mid \Pi_k)\} \to 0$ for any c by the dominate convergence theorem. Therefore, $D_k \to 0$ in probability as $n_k \to \infty$. It follows that $|\hat{T}(\mu) - \sum_{k=1}^{K}(I(\Theta_k < \mu) - 1/2)| \to 0$, in probability, as $\min\{n_1, \cdots, n_K\} \to \infty$. Similarly, since

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

one can show that $T^*(\mu) - \sum_{k=1}^{K} |I(\Theta_k < \mu) - 0.5|(2\Delta_k - 1) \to 0$, in probability as $\min\{n_1, \cdots, n_K\} \to \infty$, where Δ_k is a random sample from a Bernoulli with a success

probability of 1/2. Therefore, for any t and positive c,

$$\mathrm{pr}_{\{(X_k,\Pi_k)_{k=1,\cdots,K}\}}\left(\left|\ \mathrm{pr}(T^*(\mu)\le t|(X_k,\Pi_k)_{k=1,\cdots,K})-\right.\right.$$

$$\left.\left.\mathrm{pr}(0.5\sum_{k=1}^{K}(2\Delta_k-1)\le t\ \right|\ge c\ \right)\le c,$$

when $\min\{n_1,\cdots,n_K\}$ is large. This, coupled with the fact that $\sum_{k=1}^{K}(I(\Theta_k<\mu)-1/2)\sim 0.5\sum_{k=1}^{K}(2\Delta_k-1)$ under the null hypothesis, implies that one can approximate the null distribution of $\hat{T}(\mu)$ by the distribution of $T^*(\mu)$ 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,[22] 2005	0.50	(0.05-4.99)
	Gordon et al,[17] 2006	0.67	(0.23-2.00)
	Abels,[28] 1993	0.89	(0.41-1.93)
	Charu et al,[15] 2007	1.38	(0.44-4.33)
	Glaspy et al,[19] 2007	1.43	(1.06-1.92)
	Smith et al,[52] 2003	3.96	(0.29-54.12)
Treatment-Related Anemia	Throuvalas et al,[55] 2000	0.13	(0-332.66)
	Dunphy et al, [35] 1999	0.14	(0-6.88)
	Vadhan-Raj et al, [56] 2004	0.15	(0-415.90)
	Dammacco et al, [33] 2001	0.32	(0.11-0.95)
	Del Mastro et al, [34] 1997	0.36	(0.05-2.56)
	Cazzola et al, [31] 1995	0.37	(0.06-2.27)
	P-174,[11] 2004	0.41	(0.03-5.76)
	Thatcher et al, [54] 1999	0.49	(0.03-8.71)
	Kotasek et al, [39] 2003	0.55	(0.11-2.71)
	Oberhoff et al, [44] 1998	0.61	(0.24-1.55)
	Blohmer et al, [24] 2003 (AGO/NOGG)	0.67	(0.34-1.33)
	Henry and Abels, [38] 1994	0.75	(0.28-2.01)
	Vansteenkiste et al,[2] 2002	0.78	(0.60-1.01)
	Littlewood et al, [41] 2001	0.81	(0.62-1.06)
	Taylor et al, [21] 2005 (DA 232)	0.85	(0.45-1.60)
	EPO-CAN-17,[12] 2007	0.88	(0.49-1.59)
	Amgen DA 145, [13] 2007	0.93	(0.82-1.05)
	Razzouk et al, [49] 2004	0.98	(0.14-6.90)
	Savonije et al, [51] 2004	0.98	(0.36-2.67)
	ten Bokkel Huinink et al, [53] 1998	1.01	(0.19-5.31)
	Osterborg et al, [46] 1996	1.02	(0.51-2.04)
	Coiffier et al, [32] 2001	1.02	(0.38-2.73)
	Debus et al, [16] 2007(EPO-GER-22)	1.02	(0.60-1.74)
	Osterborg et al, [47] 2005	1.04	(0.80-1.35)
	EPO-GBR-7, [12] 2007	1.07	(0.73-1.57)
	Case et al, [30] 1993	1.08	(0.44-2.66)
	Witzig et al, [57] 2005	1.09	(0.83-1.43)
	Moebus et al, [18] 2007	1.14	(0.77-1.69)
	Strauss et al, [23] 2007	1.16	(0.69-1.95)
	Thomas et al, [36] 2007 (GOG-191)	1.25	(0.65-2.41)
	Thatcher et al, [54] 1999	1.26	(0.24-6.60)
	Overgaard et al, [14] 2007 (DAHANCA 10)	1.28	(0.97-1.69)
	Hedenus et al, [37] 2003	1.36	(0.98-1.89)
	Leyland-Jones et al, [40] 2005 (INT-76)	1.37	(1.07-1.75)
	Henke et al, [5] 2003	1.39	(1.05-1.84)
	Machtay et al, [42] 2007 (RTOG 99-03)	1.41	(0.80-2.49)
	PREPARE, [48] 2007	1.50	(0.96-2.34)
	Grote et al, [43] 2005 (N93-004)	1.53	(0.65-3.61)
	INT-3, [11] 2004	1.56	(0.42-5.79)
	INT-1, [11] 2004	1.58	(0.32-7.82)
	Rose et al, [50] 1994	1.68	(0.66-4.29)
	Bamias et al, [29] 2003	1.80	(0.53-6.12)
	Wright et al, [58] 2007 (EPO-CAN-20)	1.84	(1.01-3.35)
	EPO-CAN-15, [11] 2004	2.70	(1.17-6.23)
	Wilinson et al, [20] 2006	4.54	(0.40-51.20)
	O'Shaughnessy et al, [45] 2005	7.39	(0.15-366.10)

Table 2: Event Rates and Study-Level Summary Statistics for VTE for Cancer Studies with ESAs vs Control from Bennett et al. (2008)

	No. of VETevents/No. of patients		Relative Risk	
Study	ESA	Control	Point Estimate	95% CI
Bamias et al, [29] 2003	0/72	1/72	0.33	(0.01-8.05)
Smith et al, [52] 2003	1/64	1/22	0.34	(0.02-5.27)
Case et al, [30] 1993	2/81	3/76	0.63	(0.11-3.64)
Henry and Abels, [38] 1994	6/67	8/65	0.73	(0.27-1.98)
Wright et al, [58] 2007 (EPO-CAN-20)	2/33	3/37	0.75	(0.13-4.20)
Grote et al, [43] 2005 (N93-004)	24/109	26/115	0.97	(0.60-1.59)
EPO-CAN-17, [12] 2007	19/175	14/175	1.36	(0.70-2.62)
Littlewood et al, [41] 2001	14/251	5/124	1.38	(0.51-3.75)
Vansteenkiste et al, [2] 2002	7/155	5/159	1.44	(0.47-4.43)
INT-1, [11] 2004	3/164	1/80	1.46	(0.15-13.85)
Leyland-Jones et al, [40] 2005 (INT-76)	36/448	25/456	1.47	(0.89-2.40)
Witzig et al, [57] 2005	9/168	6/165	1.47	(0.54-4.05)
Osterborg et al, [46] 1996	1/48	0/24	1.53	(0.06-36.23)
Amgen DA 145, [13] 2007	66/298	43/298	1.53	(1.08-2.18)
Henke et al, [5] 2003	10/180	6/171	1.58	(0.59-4.26)
ten Bokkel Huinink et al, [53] 1998	2/45	0/17	1.96	(0.10-38.79)
Debus et al, [16] 2007 (EPO-GER-22)	20/108	10/107	1.98	(0.97-4.03)
Charu et al, [15] 2007	Not Available		2.36	(0.13-43.20)
Rose et al, [50] 1994	9/142	2/79	2.50	(0.55-11.30)
Thatcher et al, [54] 1999	2/44	0/22	2.56	(0.13-51.05)
Osterborg et al, [46] 1996	2/47	0/25	2.71	(0.14-54.32)
Abels, [28] 1993	1/65	0/59	2.73	(0.11-65.68)
Throuvalas et al, [55] 2000	1/28	0/26	2.79	(0.12-65.66)
GOG-191, [11] 2007	9/58	3/55	2.84	(0.81-9.96)
Razzouk et al, [49] 2004	6/112	2/110	2.95	(0.61-14.28)
Welch et al, [61] 1995	1/15	0/15	3.00	(0.13-74.41)
Osterborg et al, [47] 2005	1/170	0/173	3.05	(0.13-74.41)
Gordon et al, [17] 2006	4/162	0/56	3.15	(0.17-57.55)
ten Bokkel Huinink et al, [53] 1998	4/42	0/16	3.56	(0.20-62.58)
Vadhan-Raj et al, [56] 2004	7/29	2/31	3.74	(0.85-16.56)
INT-3, [11] 2004	8/135	1/65	3.85	(0.49-30.15)
Wilkinson et al, [20] 2006	12/173	1/59	4.09	(0.54-30.80)
Savonije et al, [51] 2004	9/211	1/104	4.44	(0.57-34.55)
Machtay et al, [42] 2007 (RTOG 99-03)	2/71	0/70	4.93	(0.24-100.89)
EPO-GBR-7, [12] 2007	5/151	1/149	4.93	(0.58-41.73)
Dammacco et al, [33] 2001	5/69	1/76	5.51	(0.66-45.98)
EPO-CAN-15, [11] 2004	16/53	2/53	8.00	(1.93-33.09)
Rosenzweig et al, [60] 2004	4/14	0/13	8.40	(0.50-142.27)
Cascinu et al, [59] 1994	0/50	0/50		
P-174, [11] 2004	0/33	0/12		
Thatcher et al, [54] 1999	0/42	0/22		

Table 3: 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 Bivariate Logit-Normal Or A Bivariate Beta Distribution For The Two Underlying Random Event Rates

	Bivariate Logit-Normal			Bivariate Beta		
Number of	DL	$\hat{T}(\cdot)$	$\tilde{T}(\cdot)$	DL	$\hat{T}(\cdot)$	$\tilde{T}(\cdot)$
Studies K	ECL, ML	ECL, ML	ECL, ML	ECL, ML	ECL, ML	ECL, ML
40	0.86, 0.62	0.94, 0.72	0.95, 0.90	0.87, 0.40	0.95, 0.52	0.96, 0.65
30	0.88, 0.71	0.94, 0.83	0.95, 1.03	0.88, 0.46	0.95, 0.61	0.96, 0.75
20	0.88, 0.85	0.94, 1.00	0.95, 1.23	0.90, 0.55	0.96, 0.75	0.96, 0.91
10	0.88, 1.18	0.95, 1.54	0.97, 2.15	0.91, 0.76	0.96, 1.10	0.98, 1.56
6	0.91, 1.57	0.95, 2.29	0.97, 2.89	0.88, 1.00	0.95, 1.58	0.97, 2.10

Table 4: 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	
Number of Studies K	$\hat{T}_p(\cdot)$ ECL, ML	$\tilde{T}_p(\cdot)$ ECL, ML	$\hat{T}_p(\cdot)$ ECL, ML	$\tilde{T}_p(\cdot)$ ECL, ML	$\hat{T}_p(\cdot)$ ECL, ML	$\tilde{T}_p(\cdot)$ ECL, ML	$\hat{T}_p(\cdot)$ ECL, ML	$\tilde{T}_p(\cdot)$ ECL, ML
40	0.95, 0.86	0.86, 1.16	0.95, 0.81	0.92, 0.92	0.96, 0.48	0.93, 0.55	0.96, 0.73	0.92, 0.96
35	0.96, 0.91	0.88, 1.21	0.96, 0.86	0.90, 1.02	0.96, 0.52	0.95, 0.61	0.96, 0.78	0.93, 1.04
30	0.96, 1.00	0.90, 1.37	0.96, 0.94	0.91, 1.12	0.95, 0.56	0.94, 0.64	0.96, 0.85	0.93, 1.07
25	0.96, 1.12	0.90, 1.49	0.97, 1.06	0.92, 1.23	0.96, 0.62	0.93, 0.65	0.96, 0.94	0.92, 1.10
20	0.96, 1.24	0.92, 1.52	0.97, 1.16	0.92, 1.32	0.96, 0.72	0.95, 0.80	0.96, 1.37	0.95, 1.37

Table 5: 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	DL	$\hat{T}(\cdot)$	$\tilde{T}(\cdot)$
Studies K	ECL, ML	ECL, ML	ECL, ML
$K = 40$	0.92, 0.24	0.95, 0.27	0.96, 0.35
$K = 30$	0.94, 0.26	0.95, 0.30	0.96, 0.39
$K = 20$	0.95, 0.30	0.95, 0.35	0.97, 0.45
$K = 10$	0.97, 0.47	0.96, 0.57	0.98, 0.84
$K = 6$	0.96, 0.75	0.95, 1.03	0.97, 1.34

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