

# **Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent $2 \times 2$ tables with all available data but without artificial continuity correction**

LU TIAN\*

*Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA*  
lutian@northwestern.edu

TIANXI CAI

*Department of Biostatistics, Harvard University, Boston, MA 02115, USA*

MARC A. PFEFFER

*Cardiovascular Division, Brigham & Women's Hospital, Boston, MA 02115, USA*

NIKITA PIANKOV

*Analysis Group, Boston, MA 02115, USA*

PIERRE-YVES CREMIEUX

*Department of Economics, University of Quebec at Montreal, Montreal, Quebec, Canada*  
*and Analysis Group, Boston, MA 02115, USA*

L. J. WEI

*Department of Biostatistics, Harvard University, Boston, MA 02115, USA*

## SUMMARY

Recently, meta-analysis has been widely utilized to combine information across comparative clinical studies for evaluating drug efficacy or safety profile. When dealing with rather rare events, a substantial proportion of studies may not have any events of interest. Conventional methods either exclude such studies or add an arbitrary positive value to each cell of the corresponding  $2 \times 2$  tables in the analysis. In this article, we present a simple, effective procedure to make valid inferences about the parameter of interest with all available data without artificial continuity corrections. We then use the procedure to analyze the data from 48 comparative trials involving rosiglitazone with respect to its possible cardiovascular toxicity.

*Keywords:* Continuity correction for zero events; Exact inference procedure; Odds ratio; Risk difference.

\*To whom correspondence should be addressed.

## 1. INTRODUCTION

Meta-analysis provides a framework for combining information across multiple independent, but “similar” clinical studies to make inferences about a common parameter (Egger and Smith, 1997a; Egger *and others*, 1997b). Almost all existing methods rely on large-sample approximations to the distributions of the combined point estimators. Such approximations may be inaccurate and lead to invalid conclusions when the individual study sample sizes are small, the number of studies is not large, or the event rates are low (Brown *and others*, 2001). Moreover, when the events of interest are very rare, the standard procedures either apply continuity corrections to the studies with zero events or simply exclude them from the analysis (Sankey *and others*, 1996; Bradburn *and others*, 2007). For example, recently Nissen and Wolski (2007) (N&W) performed a meta-analysis to examine whether rosiglitazone, a drug for treating type 2 diabetes mellitus, significantly increases the risk of myocardial infarction (MI) or cardiovascular disease (CVD)–related death. Of 48 trials satisfied the inclusion criteria for their analysis, 10 studies have no MI events and 25 studies have no CVD-related deaths. N&W simply excluded those studies from their analysis. On the other hand, different continuity corrections may result in different conclusions (Sweeting *and others*, 2004). An interesting yet challenging methodology question is how to use all available data without assigning an arbitrary number to the empty cells in meta-analysis.

In this article, we present a simple procedure to construct “valid” confidence intervals (CIs) for a common parameter using all available data. The new proposal does not rely on the large-sample approximation or arbitrary continuity corrections to obtain interval estimates. We reanalyze the CVD-related event data of N&W with the proposed method. A numerical study is also conducted to compare the performance of the new proposal to existing procedures.

## 2. COMBINING INDIVIDUAL CIs FOR A COMMON PARAMETER OF INTEREST

Suppose that we are interested in constructing a  $100(1 - \alpha)\%$  1-sided CI  $(a, \infty)$  for  $\Delta$ , a common parameter, based on all data from  $n$  independent studies. For a given  $\eta$ , there are  $n$  study-specific 1-sided  $\eta$ -level CIs for  $\Delta$ . Now, for any fixed value of  $\Delta$ , say 0, we examine whether 0 is the true value of  $\Delta$ . If yes, then on average, 0 should belong to at least  $100\eta\%$  of the above  $n$  intervals. The decision on whether the interval  $(a, \infty)$  should include 0 can be made easily via a simple hypothesis testing procedure. To this end, let  $y_i = 1$ , if 0 belongs to the observed  $\eta$  interval from the  $i$ th study, and  $y_i = 0$ , otherwise. Then, we include 0 in  $(a, \infty)$  if

$$t(\eta) = \sum_{i=1}^n w_i (y_i - \eta) \geq c, \quad (2.1)$$

where  $w_i$  is a study-specific positive weight,  $c$  is chosen such that  $\text{pr}\{T(\eta) < c\} \leq \alpha$ ,

$$T(\eta) = \sum_{i=1}^n w_i (B_i - \eta) \quad (2.2)$$

is the null counterpart of  $t(\eta)$ , and  $\{B_i, i = 1, \dots, n\}$  are  $n$  independent Bernoulli random variables with a “success” probability of  $\eta$ . We repeat this process with all other possible values for  $\Delta$  and obtain the final CI  $(a, \infty)$ . Here, the weight  $w_i$  may be the sample size for the  $i$ th study.

One may further improve the aforementioned interval estimate for  $\Delta$  by utilizing multiple intervals with various levels of confidence from each study. Specifically, let  $J_{ij} = (a_{ij}, \infty)$  denote the  $\eta_j$ -level 1-sided CI based on the  $i$ th study, for  $j = 1, \dots, K$ . Without loss of generality, we assume that  $0 < \eta_1 < \dots < \eta_K < 1$  and  $a_{i1} \geq \dots \geq a_{iK}$ . For any given  $\Delta$ , we would include  $\Delta$  in the final combined interval

$(a, \infty)$  if

$$\sum_{j=1}^K \tilde{w}_j t(\eta_j) \geq d, \quad (2.3)$$

where  $\tilde{w}_j$  is a positive weight for the  $\eta_j$ -level intervals and  $t(\eta_j)$  is obtained by replacing  $y_i$  and  $\eta$  in (2.1) with  $y_{ij}$  and  $\eta_j$ , respectively. Here, the critical value  $d$  is chosen such that

$$\text{pr} \left\{ \sum_{j=1}^K \tilde{w}_j T(\eta_j) < d \right\} \leq \alpha, \quad (2.4)$$

where  $T(\eta_i)$  is obtained by replacing  $B_i$  and  $\eta$  in (2.2) by  $B_{ij}$  and  $\eta_j$ , respectively, and  $\{(B_{i1}, \dots, B_{iK})', i = 1, \dots, n\}$  are  $n$  independent random vectors whose components are correlated Bernoulli variables such that  $B_{i1} \leq B_{i2} \leq \dots \leq B_{iK}$  and  $\text{pr}(B_{ij} = 1) = \eta_j$ . The final interval  $(a, \infty)$  may be obtained by repeating the process for all possible values of  $\Delta$ . We justify the validity of the proposed procedure in the Appendix. For the present case, one may let  $\tilde{w}_j$  be  $\{\eta_j(1 - \eta_j)\}^{-1}$  (Wei and Johnson, 1995; Xu and others, 2003).

Similarly, we can obtain combined  $100(1 - \alpha)\%$  1-sided CI,  $(-\infty, b)$ . It follows that  $(a, b)$  would be a  $100(1 - 2\alpha)\%$  2-sided CI for  $\Delta$ . A point estimator for  $\Delta$  may be obtained as  $\hat{\Delta}$  such that  $\hat{\Delta}$  belongs to the intersection of all nonempty 2-sided CIs for  $\Delta$ .

### 3. EXAMPLE AND NUMERICAL STUDY

First, we illustrate the new procedure with the rosiglitazone data set described in Section 1. For convenience, we provide the data from 48 studies in Table I of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>). It is important to note that N&W did not utilize the information from studies which reported zero events of interest. Thus, their analysis only included 38 studies for the MI end point and 23 studies for the CVD-related mortality.

With respect to MI, N&W obtained a 95% CI of (1.03, 1.98) with a  $p$ -value of 0.03 for the odds ratio (OR) between rosiglitazone and the control arm (in favor of the control). With respect to mortality, the 95% CI is (0.98, 2.74) with a  $p$ -value of 0.06, an almost statistically significant result in favor of the control arm. The findings by N&W are summarized in Figure I of the supplementary material available at *Biostatistics* online.

While it is unclear how to utilize studies with zero events without arbitrary continuity corrections to obtain an overall assessment for OR in meta-analysis, we are able to use all data to make inference about

Table 1. *The empirical coverage levels of 95% CIs for the OR obtained based on (i) the MH method with 0.5 imputation (MH-0.5) and (ii) Peto's procedure with event rates ranging from extremely low to moderately low. The results are based on 1000 simulated data sets consisting of 48  $2 \times 2$  tables with the observed sample sizes listed*

| OR | (a) $p_{C_i} \sim U(0, 0.01\%)$ |       | (b) $p_{C_i} \sim U(0, 0.1\%)$ |       | (c) $p_{C_i} \sim U(0, 10\%)$ |       |
|----|---------------------------------|-------|--------------------------------|-------|-------------------------------|-------|
|    | MH-0.5                          | Peto  | MH-0.5                         | Peto  | MH-0.5                        | Peto  |
| 1  | 1.000                           | 0.987 | 0.995                          | 0.958 | 0.948                         | 0.947 |
| 2  | 0.974                           | 0.968 | 0.878                          | 0.961 | 0.944                         | 0.862 |
| 5  | 0.741                           | 0.972 | 0.233                          | 0.975 | 0.957                         | 0.000 |

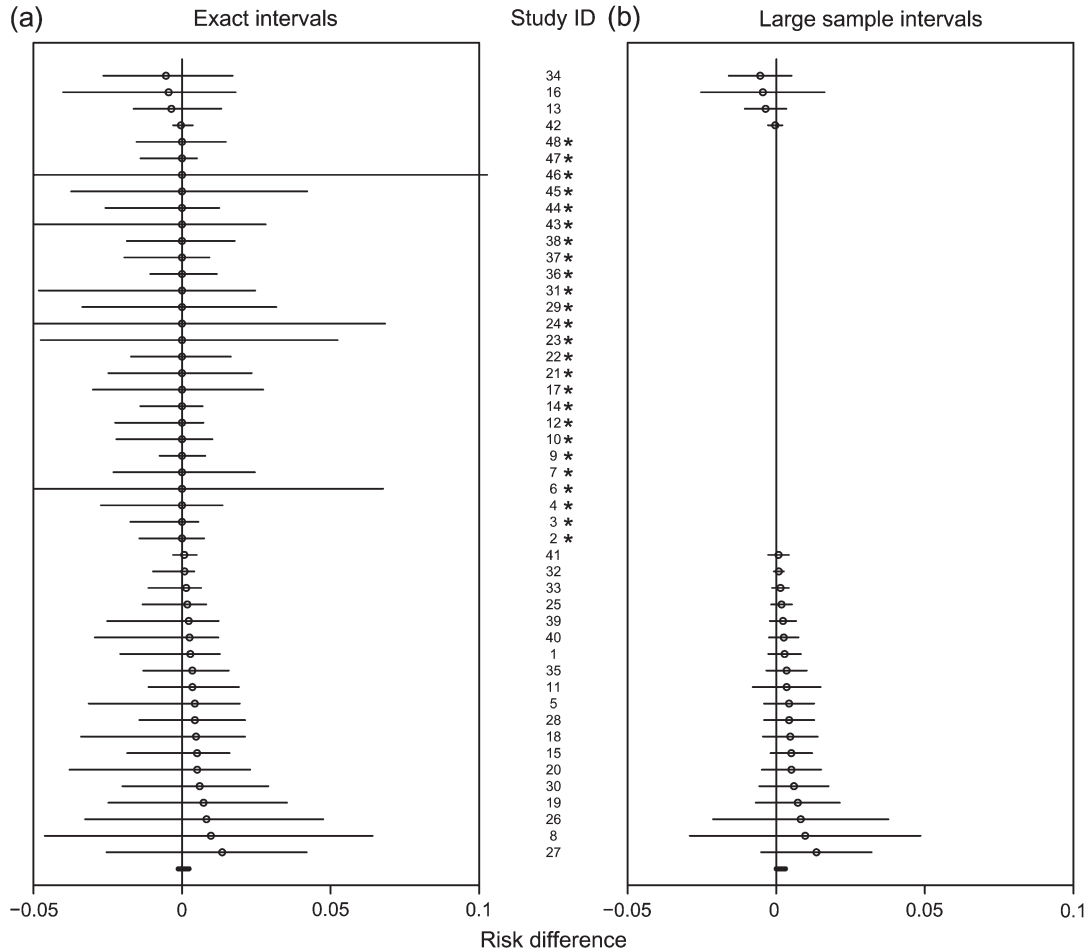


Fig. 1. The 95% CIs of the RD for CVD death (rosiglitazone minus control) with 48 studies (small circles are the observed RDs).

the risk difference (RD), as a measure of the group difference. To this end, we let  $\Delta$  be the RD (rosiglitazone minus control) in our analysis. We construct 95% CIs for  $\Delta$  based on the data from 48 studies via the proposed procedure. Here, we let  $\{\eta_j\}$  be 20 equally spaced levels from 0.1 to 0.95. Moreover, for each study, the  $\eta_j$ -level CI for  $\Delta$  is obtained via the mid- $p$ -value method (Chan and Zhang, 1999; Hwang and Yang, 2001). The cutoff point  $d$  in (2.4) is determined by randomly generating 50 000 independent samples  $\{(B_{i1}, \dots, B_{iK})\}$ . For comparison, we also obtain the corresponding Mantel–Haenszel (MH) weighted CI estimates for RD.

For the mortality end point, we present the forest tree diagram in Figure 1(a). The thick line segment right above the  $x$ -axis is the combined 95% CI  $(-0.13, 0.23)\%$ . The point estimator of RD is 0.063%, and the  $p$ -value for testing  $\text{RD} = 0$  is 0.83. On the other hand, in Figure 1(b), the 95% MH final interval for RD is  $(0.00, 0.31)\%$  with a  $p$ -value of 0.05 after excluding 25 studies with no events.

For the MI endpoint, the final combined 95% interval is  $(-0.08, 0.38)\%$ . The point estimator for RD is 0.183% with a  $p$ -value of 0.27. Again, if we exclude 10 studies with no events and use the asymptotic inference as in N&W, the final 95% MH interval is  $(0.02, 0.42)\%$  with a  $p$ -value of 0.03 (see Figure II in the supplementary material available at *Biostatistics* online).

Table 2. The empirical coverage levels and average lengths of 95% CIs for RD obtained based on (i) the MH-0 method, (ii) the MH-0.5 method, and (iii) the new exact method for cases with extremely to moderately low event rates. The results are based on 1000 simulated data sets consisting of  $48 \times 2 \times 2$  tables with the observed sample sizes listed in Table 1

| Empirical coverage level (average length $\times$ 1000) |              |              |              |                           |              |              |              |
|---------------------------------------------------------|--------------|--------------|--------------|---------------------------|--------------|--------------|--------------|
| (a) $p_c \sim U(0, 0.01\%)$                             |              |              |              | (b) $p_c \sim U(0, 10\%)$ |              |              |              |
| RD                                                      | MH-0         | MH-0.5       | Exact        | RD                        | MH-0         | MH-0.5       | Exact        |
| 0                                                       | 0.761 (0.29) | 1.000 (2.10) | 1.000 (1.45) | 0                         | 0.959 (10.3) | 0.959 (10.3) | 0.995 (12.3) |
| 0.01%                                                   | 0.896 (0.44) | 1.000 (2.13) | 1.000 (1.51) | 5%                        | 0.947 (12.2) | 0.947 (12.2) | 0.959 (14.8) |
| 0.05%                                                   | 0.930 (0.78) | 0.999 (2.22) | 0.999 (1.91) | 10%                       | 0.957 (13.7) | 0.957 (13.7) | 0.958 (16.8) |
| 0.50%                                                   | 0.955 (2.28) | 0.998 (3.69) | 0.987 (3.67) |                           |              |              |              |

Numerical studies were conducted to examine the performance of the new proposal and the standard large-sample interval estimation procedures. The current literature suggests that the large-sample MH-type combining techniques for the OR or RD are relatively more reliable than other methods (Bradburn and others, 2007). To save space, we only report the results with the MH-type procedures as the main comparators in the study.

The first part of our simulation study was designed to examine the performance of various existing interval estimation procedures for  $\Delta = \text{OR}$ . We mimicked the data structure presented in Table I of the supplementary material available at *Biostatistics* online and generated 48 studies with the observed sample sizes. Let  $p_{ci}$  and  $p_{ri}$  be the incidence rates for the control and rosiglitazone groups in the  $i$ th study, respectively. For each given OR,  $\{p_{ci}, i = 1, \dots, 48\}$  were generated randomly from a uniform distribution  $U(0, \xi)$ . Then, we let  $\text{logit}(p_{ri}) = \log(\text{OR}) + \text{logit}(p_{ci})$  to generate data for the treated group. In Table 1, we present the empirical coverage levels of 95% CIs obtained based on (i) the MH method with 0.5 imputation (MH-0.5) for studies with zero event in one arm and (ii) the Peto's method utilized in N&W. All results are based on 1000 replications. It is clear that either method may not be valid when the true OR is away from one. Through our simulation studies, we also find that the existing exact method for the OR conditional on all marginal totals of  $2 \times 2$  tables works well.

In the second part of our simulation, we use the RD as the contrast measure  $\Delta$ . We generate  $p_{ci}$  from the uniform distribution and  $p_{ri} = \text{RD} + p_{ci}$ . The results are reported in Table 2. When the underlying event rates for both groups are low, the MH method without imputation has serious coverage problems. For this case, both our exact and MH-0.5 methods are conservative. However, the average interval lengths from exact method are shorter than that of the MH-0.5 method. On the other hand, when the underlying event rates are not very low, all procedure have proper coverage levels and comparable interval lengths.

#### 4. REMARKS

Unlike the standard large-sample meta-analysis procedures, the proposed simple method provides valid exact inferences about the parameter of interest under any fixed-effects framework. It effectively utilizes all data without artificial imputation. Our method may be regarded as a generalization of the combination method of confidence distributions studied by Singh and others (2005). As other exact procedures, the proposed method can be over conservative in some cases. It would be interesting to examine if one can improve the efficiency of the proposed exact method by considering *a priori* (or random-effect) distribution for the nuisance parameter, such as the control success rate in the  $2 \times 2$  tables, as suggested by the associate editor. The computer code for implementing the procedure is available at "<http://biosun1.harvard.edu/~tcai/MetaCode.r>".

## ACKNOWLEDGMENTS

*Conflict of Interest:* None declared.

## FUNDING

National Institutes of Health (U54LM008748; R01AI052817; R01HL89778).

## APPENDIX A

## A.1 Justification of validity for the final combined interval

Assume that the  $n$  studies in our meta-analysis are realizations of a random sample from a population whose distribution is generated by a random quantity  $\Pi$ . For example, for the  $2 \times 2$  tables,  $\Pi$  consists of  $\Delta_0$ , the underlying event rate for the control arm, and possibly the sample sizes for 2 arms of the study. Let  $\pi_i$  be the realization of  $\Pi$  for the  $i$ th study,  $i = 1, 2, \dots, n$ . Note that one may further assume that the number  $n$  of studies is a random component  $N$  of  $\Pi$ . Given  $\pi_i$ , the data  $X_i$  were generated for  $i = 1, \dots, n$ . The 1-sided CI  $J_{ij} = (a_{ij}, \infty)$  for  $\Delta_0$  satisfies the condition

$$\text{pr}(\Delta_0 \in J_{ij} | \pi_i) \geq \eta_j, \quad j = 1, \dots, K, \quad (\text{A.1})$$

where the probability is generated by  $X_i$ . Now, for any given  $\Delta$ , we test the null hypothesis that  $\Delta = \Delta_0$ . Let  $Y_{ij} = 1$ , if  $\Delta \in J_{ij}$ ;  $Y_{ij} = 0$ , otherwise. Consider the test statistic  $T_0 = \sum_{i=1}^n \sum_{j=1}^K (Y_{ij} - \eta_j) w_i \tilde{w}_j$ . We then derive the critical value  $d$  in (2.3) and (2.4) from the null counterpart of  $T_0$ ,  $\tilde{T} = \sum_{i=1}^n \sum_{j=1}^K (B_{ij} - \eta_j) w_i \tilde{w}_j$ . Note that if  $\Delta = \Delta_0$ , (A.1) implies that

$$\text{pr}(Y_{ij} = 1 | \pi_i) \geq \eta_j, \quad j = 1, \dots, K. \quad (\text{A.2})$$

Under this condition, we will show that

$$\text{pr}(T_0 \geq v) \geq \text{pr}(\tilde{T} \geq v), \quad \text{for all } v. \quad (\text{A.3})$$

Equation (A.3) directly implies that conditional on any set of realizations  $\{\pi_1, \dots, \pi_n\}$ ,  $\text{pr}(T_0 < d | \Delta = \Delta_0, \pi_1, \dots, \pi_n) \leq \alpha$ , where the probability is generated by  $\{X_i, i = 1, \dots, n\}$ . It follows that  $\text{pr}(T_0 < d | \Delta = \Delta_0) \leq \alpha$ , where the probability is generated under the random pairs  $\{(X_i, \Pi_i), i = 1, \dots, n\}$ . Here,  $X_i$  is a random quantity associated with  $\Pi_i$  and  $\{\Pi_i\}$  is a random sample from the population  $\Pi$ .

To show (A.3), first let  $g(u_1, \dots, u_K) = \sum_{j=1}^K (u_j - \eta_j) \tilde{w}_j$ . Since for each individual study, the  $K$  CIs are nested,  $g(y_{i1}, \dots, y_{iK})$  or  $g(B_{i1}, \dots, B_{iK})$  can only assume  $K + 1$  possible distinct values:  $v_k = \sum_{j \geq k+1} \tilde{w}_j - \sum_{j=1}^K \eta_j \tilde{w}_j$ ,  $k = 1, \dots, K$ , where  $y$  is the observed value of  $Y$  and  $v_0 > v_1 \cdots > v_K$ . Furthermore,  $g(y_{i1}, \dots, y_{iK}) = v_s$  if and only if  $y_{ij} = I(j > s)$ . Similarly,  $g(B_{i1}, \dots, B_{iK}) = v_s$  if and only if  $B_{ij} = I(j > s)$ . It follows that for  $v \in (v_s, v_{s-1}]$ ,

$$\text{pr}\{g(Y_{i1}, \dots, Y_{iK}) \geq v\} = \text{pr}\{g(Y_{i1}, \dots, Y_{iK}) \geq v_{s-1}\} = \text{pr}(Y_{is} = 1).$$

Under (A.2),  $\text{pr}(Y_{is} = 1) \geq \eta_s = \text{pr}(B_{is} = 1) = \text{pr}\{g(B_{i1}, \dots, B_{iK}) \geq v\}$ . Consequently,  $\text{pr}\{g(Y_{i1}, \dots, Y_{iK}) \geq v\} \geq \text{pr}\{g(B_{i1}, \dots, B_{iK}) \geq v\}$  and (A.3) follows.

## REFERENCES

BRADBURN, M. J., DEEKS, J. J., BERLIN, J. AND LOCALIO, A. R. (2007). Much ado about nothing: a comparison of the performance of meta-analysis methods with rare events. *Statistics in Medicine* **26**, 53–77.

- BROWN, L., CAI, T. AND DASGUPTA, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16**, 101–103.
- CHAN, I. AND ZHANG, Z. (1999). Test-based exact confidence interval for the difference of two binomial proportions. *Biometrics* **55**, 1202–1209.
- EGGER, M. AND SMITH, G. D. (1997a). Meta-analysis: potentials and promise. *BMJ* **315**, 1371–1374.
- EGGER, M., SMITH, G. D. AND PHILLIPS, A. N. (1997b). Meta-analysis: principles and procedures. *BMJ* **315**, 1533–1537.
- HWANG, G. AND YANG, M. (2001). An optimality theory for mid  $p$ -values in  $2 \times 2$  contingency tables. *Statistical Sinica* **11**, 807–826.  
*Clinical Trials* **3**, 73–98.
- NISSEN, S. E. AND WOLSKI, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine* **357**, 1–100.
- SANKEY, S., WEISSFELD, L., FINE, M. AND KAPOOR, W. (1996). An assessment of the use of the continuity correction for sparse data in meta analysis. *Communications in Statistics: Simulation and Computation* **25**, 1031–1056.
- SINGH, K., XIE, M. AND STRAWDERMAN, W. (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics* **33**, 159–183.
- SWEETING, M. J., SUTTON, A. J. AND LAMBERT, P. C. (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* **23**, 1351–1375.
- WEI, L. J. AND JOHNSON, W. (1995). Combining dependent tests with incomplete repeated measurements. *Biometrika* **72**, 359–364.
- XU, X., TIAN, L. AND WEI, L. J. (2003). Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics* **4**, 223–229.

[Received October 31, 2007; first revision March 12, 2008; second revision May 23, 2008;  
third revision August 14, 2008; accepted for publication September 3, 2008]