BIOMETRICS 56, 304–308
March 2000

READER REACTION

Estimating Risk Difference in Multicenter Studies Under Baseline-Risk Heterogeneity

Dankmar Böhning

Department of Epidemiology, Free University Berlin, Haus 562, Fabeckstrasse 60-62, 14195 Berlin, Germany
email: boehning@zedat.fu-berlin.de

and

Jesus Sarol Jr.

Department of Epidemiology and Biostatistics, College of Public Health, University of the Philippines at Manila, Manila, Philippines
email: jsarol@nwave.net

SUMMARY. In this paper, we consider the case of efficient estimation of the risk difference in a multicenter study allowing for baseline heterogeneity. We consider the optimally weighted estimator for the common risk difference and show that this estimator has considerable bias when the true weights (which are inversely proportional to the variances of the center-specific risk difference estimates) are replaced by their sample estimates. In addition, we propose a new estimator for this situation of the Mantel–Haenszel type that is unbiased and, in addition, has a smaller variance for small sample sizes within the study centers. Simulations illustrate these findings.

KEY WORDS: Bias in conventional estimator; Cochran's estimator; Heterogeneity in baseline risks; Mantel–Haenszel estimator; Meta-analysis; Optimal summary measure; Pooling of sparse clinical trials; Risk difference.

1. Introduction

In many clinical trials, including those undertaken by large cooperative cancer groups, patients are randomized to one of two treatment groups within a center (i.e. hospital), with treatment allocation (approximately) balanced within centers. The interest in the clinical trial lies in estimating the difference in the success rates of the two treatments, sometimes called the risk difference or treatment difference. As Lipsitz et al. (1998) pointed out, the risk difference is often used as the measure of effect in practice, and it has a nice interpretation in a clinical trial. The risk difference measures the actual gain expected in terms of percentage of patients treated.

Consider k centers in which two treatments are compared and the outcome measures are binary. Let p_i be the probability for positive response in treatment 1 and q_i the probability for positive response in treatment 2 for centers $i = 1, 2, \ldots, k$. In particular, it is allowed that the p_i's are different (baseline heterogeneity). The parameter of interest is the risk difference, defined as $\tau_i = q_i - p_i$ for $i = 1, 2, \ldots, k$. Conventionally, interest lies in the hypothesis H_0: $\tau_i = \tau$ or in certain subhypotheses, such as H_{01}: $\tau = 0$.

This situation has received considerable interest, also through the work of Lipsitz et al. (1998). Lipsitz et al. were mainly interested in the issue of developing and evaluating (heterogeneity) tests for the hypothesis (of homogeneity) H_0: $\tau_i = \tau$ for $i = 1, 2, \ldots, k$ against H_1: $\tau_i \neq \tau_j$ for some $i \neq j$. The work of Lipsitz et al. (1998) was later critically discussed and extended by Lui and Kelly (1999).

We are interested here in estimating τ efficiently, a point that has been overlooked in both papers mentioned above. Estimating τ makes the most sense if there is homogeneity of the risk difference across study centers. However, if there is heterogeneity in the risk difference across centers, we assume that τ is defined as follows: Let $f(\tau')$ denote the density of the risk difference in the population of all possible clinical trials. Then define τ as the expected value with respect to $f(\tau')$, $\int_{-\infty}^{\infty} \tau' f(\tau') d\tau'$.

Let n_i be the sample size in center i for treatment 1 and m_i be the sample size for treatment 2, respectively. Also, let

Table 1
Available data in each center in CALGB study

Center	X_i	n_i	Y_i	m_i	$\hat{\tau}_i = Y_i/m_i - X_i/n_i$	$\hat{w}_i^{-1} = \widehat{\text{var}}(\hat{\tau}_i) = X_i(n_i - X_i)/n_i^3 + Y_i(m_i - Y_i)/m_i^3$
1	1	3	3	4	0.42	0.12
2	8	11	3	4	0.02	0.06
3	2	3	2	2	0.33	0.07
4	2	2	2	2	0.00	0.00
5	0	3	2	2	1.00	0.00
6	2	3	1	3	−0.33	0.15
7	2	3	2	2	0.33	0.07
8	4	4	1	5	−0.80	0.03
9	2	3	2	2	0.33	0.07
10	2	3	0	2	−0.67	0.07
11	3	3	3	3	0.00	0.00
12	0	2	2	2	1.00	0.00
13	1	5	1	4	0.05	0.08
14	2	4	2	3	0.17	0.14
15	4	6	2	4	−0.16	0.10
16	3	9	4	12	0.00	0.04
17	2	3	1	2	−0.17	0.20
18	1	4	3	3	0.75	0.05
19	2	3	1	4	−0.42	0.12
20	0	2	0	3	0.00	0.00
21	1	5	2	4	0.30	0.09

X_i and Y_i be the associated number of positive responses in center i and $\hat{\tau}_i = Y_i/m_i - X_i/n_i$ the estimated risk difference in center i, $i = 1, 2, \ldots, k$. Lipsitz et al. (1998) considered linear, unbiased estimates of the form

$$\hat{\tau}_w = \sum_{i=1}^{k} w_i \hat{\tau}_i \Big/ \sum_{i=1}^{k} w_i, \tag{1}$$

with nonrandom, nonnegative constants $w_1, w_2, \ldots, w_k$. Since it is well known that those $w_1, w_2, \ldots, w_k$ with

$$w_i^{-1} = \text{var}(\hat{\tau}_i) = p_i(1 - p_i)/n_i + q_i(1 - q_i)/m_i,$$

$i = 1, 2, \ldots, k$, minimize the variance of (1), these weights are used in (1). Note that this implies that there is no other estimator of the form $\Sigma_{i=1}^k w_i^* \hat{\tau}_i / \Sigma_{i=1}^k w_i^*$ with smaller variance. However, the estimator (1) cannot be used in practice since p_i and q_i are unknown. Therefore, it has become common practice to replace them by their sample estimates X_i/n_i and Y_i/m_i, leading to

$$\hat{\tau}_{\hat{w}} = \frac{\sum_{i=1}^{k} \hat{w}_i \hat{\tau}_i}{\sum_{i=1}^{k} \hat{w}_i}, \tag{2}$$

with $\hat{w}_i^{-1} = X_i(n_i - X_i)/n_i^3 + Y_i(m_i - Y_i)/m_i^3$. This estimator is suggested in several textbooks of epidemiology, such as that by Kleinbaum, Kupper, and Morgenstern (1982, p. 359) or in textbooks of meta-analysis, such as that by Petitti (1994, p. 103). This replacement of the true weights by their sample estimates causes considerable problems, which superficially appear as technical difficulties, though they point to deeper problems of statistical inference. Note first that a weight in (2) is not defined in the occurrence of any of the four cases $X_i = 0$ or $X_i = n_i$ in combination with $Y_i = 0$ or $Y_i = m_i$. Lipsitz et al. (1998) remove that center from the pool for which such a case has occurred. Second, the estimator defined in (2) is not necessarily unbiased. This is primarily due to the occurrence of the product terms $\hat{w}_i \hat{\tau}_i$, for which the expected value is not necessarily equal to $w_i \tau_i$ since $\hat{w}_i$ and $\hat{\tau}_i$ are not independent in this case.

We will show that the estimator (2) is no longer efficient. Even of greater concern is the appearance of considerable bias when the study sizes in the centers are small. In fact, we will develop a new estimator of the Mantel–Haenszel type that is unbiased and has smaller variance for small sample sizes in the centers. This result remains persistent even if the number of centers gets large.

2. A New Estimator

We consider again $\hat{\tau}_i = Y_i/m_i - X_i/n_i$, which evidently can be written as

$$\hat{\tau}_i = \frac{Y_i n_i - X_i m_i}{n_i m_i}.$$

Now, instead of taking $(1/k)\Sigma_{i=1}^k \hat{\tau}_i$, we consider the ratio of sums

$$\hat{\tau}_{\text{MH}} = \frac{\sum_{i=1}^{k} (Y_i n_i - X_i m_i)}{\sum_{i=1}^{k} n_i m_i.} \tag{3}$$

We think that this estimator is in line with Mantel–Haenszel because of its similarity in first taking sums and

then ratios. Note that $\hat{\tau}_{MH}$ is a weighted average of the $\hat{\tau}_i$'s, i.e., it is of the form $\hat{\tau}_{MH} = \Sigma_{i=1}^{k} w_i\hat{\tau}_i/\Sigma_{i=1}^{k} w_i$ with $w_i = n_i m_i$. Note that these weights are nonrandom. Consequently, $\hat{\tau}_{MH}$ is unbiased. In addition, its variance is readily available as

$$\text{var}(\hat{\tau}_{MH}) = \frac{\sum_{i=1}^{k}\{n_i^2 m_i q_i(1-q_i) + m_i^2 n_i p_i(1-p_i)\}}{\left(\sum_{i=1}^{k} m_i n_i\right)^2}, \quad (4)$$

from which an estimated variance can be easily derived as

$$\widehat{\text{var}}(\hat{\tau}_{MH}) = \frac{\sum_{i=1}^{k}\{n_i^2 Y_i(m_i - Y_i)/m_i + m_i^2 X_i(n_i - X_i)/n_i\}}{\left(\sum_{i=1}^{k} m_i n_i\right)^2}. \quad (5)$$

Note that a strong advantage of $\hat{\tau}_{MH}$ is that it is defined in all data constellations, in particular, if $X_i = 0$ (or $X_i = n_i$) and $Y_i = 0$ (or $Y_i = m_i$).

3. An Application

We return to the data considered previously by Lipsitz et al. (1998). The data are from the Cancer and Leukemia Group B (CALGB) randomized clinical trial comparing two chemotherapy treatments with respect to survival (lived/died by the end of the study) in patients with multiple myeloma (Cooper et al., 1993). A total of 156 eligible patients was accrued in the 21 centers. The data are presented in Table 1. Note first that there are five centers with an estimated variance of zero for their risk difference estimator $\hat{\tau}_i$. These are centers 4, 5, 11, 12, and 20. Consequently, these five centers are deleted when computing $\hat{\tau}_{\hat{w}}$. This implies that 24 patients are lost in the analysis through the statistical procedure. Second, for the five centers that are deleted from the analysis, the risk difference is nonnegative. This clearly shows that $\hat{\tau}_{\hat{w}}$ is more negative than $\hat{\tau}_{MH}$ in this case. Indeed, we find that $\hat{\tau}_{\hat{w}} = -0.0181$ (0.00467) and $\hat{\tau}_{MH} = 0.0199$ (0.00694). The numbers in parentheses are the estimated variances according to $1/\Sigma_{i=1}^{k}\hat{w}_i$ for the estimator $\hat{\tau}_{\hat{w}}$ and (5) for the estimator $\hat{\tau}_{MH}$.

4. A Simulation Study

To compare $\hat{\tau}_{MH}$ with the conventional estimator $\hat{\tau}_{\hat{w}}$, a simulation study was done following the design of Lipsitz et al. (1998). Baseline risks $p_1, p_2, \ldots, p_k$ were generated from a uniform distribution on 0 to 0.8. To mimic variation in the sample sizes, n_i and m_i were generated from a Poisson distribution with parameter n for $i = 1, \ldots, k$. (If for small parameter values of n, values for $n_i = m_i = 0$ or 1 were sampled as sample sizes, then these were replaced by the sample size of two.) Binomial variates X_i with parameters n_i and p_i and binomial variates Y_i with parameters m_i and $q_i = p_i + \tau = p_i + 0.1$ were drawn for each center $i, i = 1, \ldots, k$. Both estimates $\hat{\tau}_{MH}$ and $\hat{\tau}_{\hat{w}}$ were then computed. The procedure was replicated 10,000 times. From these replicates, bias and variance were computed. The sample sizes in the centers varied as $n = 4, 8, 16, 32$ and the number of centers as $4, 8, 16, 32, 64$.

Table 2
Bias and standard error as a function of number of centers (k)

Sample size n	Bias($\hat{\tau}_{MH}$)	SE($\hat{\tau}_{MH}$)	Bias($\hat{\tau}_{\hat{w}}$)	SE($\hat{\tau}_{\hat{w}}$)
		$k = 4$		
4	0.0016844	0.172589	0.0171581	0.223894
8	0.0019042	0.120244	0.0092945	0.139923
16	−0.0007421	0.082643	0.0022685	0.086147
32	0.0007376	0.057654	0.0017826	0.057724
		$k = 8$		
4	−0.0008645	0.123719	0.0126881	0.163317
8	−0.0020579	0.085822	0.0060327	0.103213
16	−0.0002045	0.058414	0.0020076	0.060898
32	0.0000662	0.040782	0.0011932	0.040903
		$k = 16$		
4	−0.0001277	0.086965	0.0137672	0.119625
8	0.0007735	0.059676	0.0074448	0.074363
16	0.0002429	0.041131	0.0024659	0.043126
32	0.0001952	0.028723	0.0013385	0.028934
		$k = 32$		
4	0.0006703	0.061514	0.0153914	0.084838
8	−0.0005359	0.042436	0.0057000	0.052207
16	−0.0003399	0.029611	0.0012639	0.031472
32	−0.0001426	0.020489	0.0010254	0.020616
		$k = 64$		
4	0.0002278	0.043309	0.0153416	0.060391
8	−0.0001824	0.030054	0.0065411	0.038046
16	0.0003329	0.020753	0.0019735	0.021963
32	−0.0000531	0.014406	0.0010678	0.014517

A total of 20 constellations were studied. The results are provided in Table 2. Note the considerable bias of $\hat{\tau}_{\hat{w}}$ for small n. To demonstrate that this is not due to the sampling error caused by the simulation study, we have provided the estimated bias for $\hat{\tau}_{MH}$ as well. Note that this trend is persistent even if the number of centers gets large (Table 3). This might indicate that the estimator $\hat{\tau}_{\hat{w}}$, though consistent in center sample size n, might be inconsistent in the number of centers k. For all sample sizes up to $n = 32$, $\hat{\tau}_{MH}$ has smaller variance than $\hat{\tau}_{\hat{w}}$. (See Figure 1.)

5. Discussion

5.1 *Estimating Optimal Weights*

The results of this work shed some light on commonly believed efficient estimators. We have demonstrated that replacing the

Table 3
Bias and standard error; sample size (n) in each center = 4

Number of centers, k	Bias($\hat{\tau}_{MH}$)	SE($\hat{\tau}_{MH}$)	Bias($\hat{\tau}_{\hat{w}}$)	SE($\hat{\tau}_{\hat{w}}$)
4	0.0016844	0.172589	0.0171581	0.223894
8	−0.0008645	0.123719	0.0126881	0.163317
16	−0.0001277	0.086965	0.0137672	0.119625
32	0.0006703	0.061514	0.0153914	0.084838
64	0.0002278	0.043309	0.0153416	0.060391

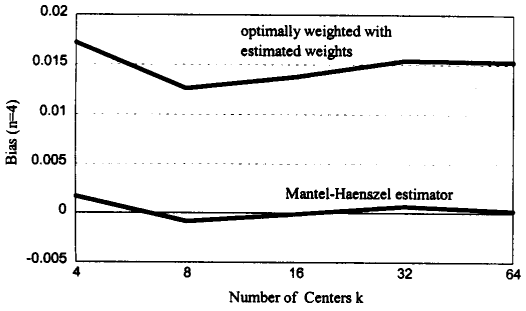

Figure 1. Bias of the two estimators versus number of centers.

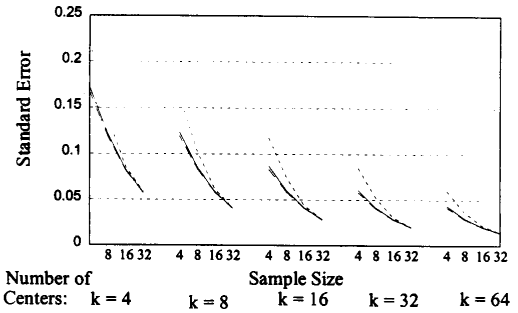

Figure 3. Standard deviation of $\hat{\tau}_{\mathrm{MH}}$ (solid), $\hat{\tau}_{\mathrm{coc}}$ (dashed), and $\hat{\tau}_{\hat{w}}$ (dotted).

true weights by their sample estimates leads to a loss in efficiency of the overall estimator. The reasons for this loss can be traced. If we consider $\hat{\tau}_{\hat{w}}$, we notice in forming $\mathrm{E}(\hat{\tau}_{\hat{w}})$ the appearance of product terms $\hat{w}_i\hat{\tau}_i/(\Sigma_j\,\hat{w}_j)$, which complicate considerably the computation of moments of $\hat{\tau}_{\hat{w}}$. We have to note that estimated center-specific weight and estimated center-specific differences are not necessarily independent. This assumption of independence might be appropriate for normally distributed outcome measures in the center since mean and variance-based estimates can be considered independent. In the case of binomial proportions or rates, however, we have a strong binding of mean and variance, so this assumption appears not to be justified, at least not for small sample sizes.

If the sample sizes for each treatment arm coincide across centers, $\hat{\tau}_{\mathrm{MH}}$ reduces to the simple mean of the risk differences $\hat{\tau}_{\mathrm{MH}} = (1/k)(\Sigma_{i=1}^{k}(Y_i/m - X_i/n)) = \bar{Y}/m - \bar{X}/n$, which seems appropriate since none of the centers is especially pronounced with respect to the sample size.

5.2 *Cochran's Weights*

Combining evidence based on the risk difference from several studies hints at the direction of meta-analytic procedures. However, combining the risk difference is rarely studied in the literature. More frequently, the pooling of several log-relative risks or log-odds ratios is studied. One of the few papers dealing with the meta-analytic handling of the risk difference can be found in the *Handbook of Research Synthesis* (Cooper and Hedges, 1994). Shadish and Haddock (1994) discuss combining differences between risks. They suggest using the estimator (2) under ignorance of the problems we have discussed here. However, they give some warnings against pooled estimators using weights that put too much weight on large sample sizes. Looking for estimators with nonrandom weights that put less weight on large studies leads to the weights suggested by Cochran (1954), $w_i = n_i m_i/(n_i + m_i)$. These weights occur when we use optimal pooling according to (1) and there is complete homogeneity, e.g., $p_i = q_j$ for all $i = 1,\ldots,k$ and $j = 1,\ldots,k$. Figure 2 clearly demonstrates that the Mantel–Haenszel weights put more weight on studies with large sample sizes in comparison with the Cochran weights.

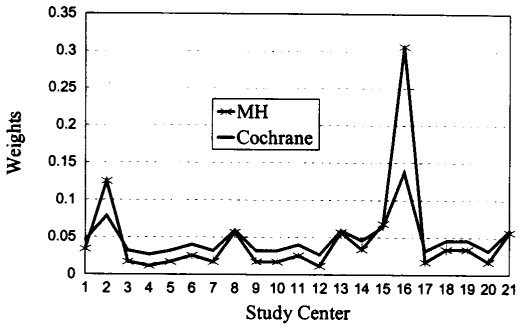

Figure 2. Comparison of Mantel–Haenszel and Cochran weights for the 21 centers of the CALGB study.

Let us define formally the estimator based on the weights by Cochran as $\hat{\tau}_{\mathrm{coc}} = \Sigma_{i=1}^{k} w_i\hat{\tau}_i/\Sigma_{i=1}^{k} w_i$, with $w_i = n_i m_i/(n_i + m_i)$. A comparison of both (unbiased) estimators in terms of their variance based on the same simulation study of the last section is shown in Figure 3. (It is clear that these variances need not be simulated since they can be calculated directly. Nevertheless, we have done so in order to achieve a better comparability to the simulated variance of the conventional estimator.) There is some benefit in using $\hat{\tau}_{\mathrm{coc}}$, though this benefit is rather small.

5.3 *Optimal Weights Under Between-Study Homogeneity*

Now suppose that there is homogeneity in the risks across centers for each treatment arm, e.g., $p_i = p$ and $q_i = q$ for all $i = 1,\ldots,k$. This implies that $w_i^{-1} = \mathrm{var}(\hat{\tau}_i) = p_i(1-p_i)/n_i + q_i(1-q_i)/m_i = p(1-p)/n_i + q(1-q)/m_i$. Estimating $\hat{p} = \Sigma_{i=1}^{k} X_i/\Sigma_{i=1}^{k} n_i$ and $\hat{q} = \Sigma_{i=1}^{k} Y_i/\Sigma_{i=1}^{k} m_i$ in a pooled manner leads to $\hat{w}_i^{-1} = \hat{p}(1-\hat{p})/n_i + \hat{q}(1-\hat{q})/m_i$, and further to $\hat{\tau}_{\mathrm{hom}} = \Sigma_{i=1}^{k}\hat{w}_i\hat{\tau}_i/\Sigma_{i=1}^{k}\hat{w}_i$. The estimators $\hat{\tau}_{\mathrm{hom}}$ and $\hat{\tau}_{\mathrm{coc}}$ behaved very similarly in the simulations study, so the results are not reported here.

Acknowledgements

The authors are grateful to the editor and Dr Jörg Kaufmann for helpful comments. This research is done under support of the German Research Foundation.

RÉSUMÉ

Dans cette étude, on s'intéresse à l'estimation efficiente de la différence des risques dans une étude multicentrique en tenant compte de l'hétérogénéité initiale. Si on considère l'estimateur à pondération optimale pour la différence de risque commune, on montre que cet estimateur est fortement biaisé quand les vraies pondérations (qui sont inversement proportionnelles aux variances des différences de risque spécifique dans chaque centre) sont remplacées par leurs estimations d'échantillonnage. Aussi, on propose un nouvel estimateur pour des situations de type Mantel-Haenszel qui n'est pas biaisé et qui, en outre, a une variance plus petite quand les échantillons des études multicentriques sont petits. Les résultats sont illustrés à partir de simulations.

REFERENCES

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Cooper, H. and Hedges, L. (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Cooper, M. R., Dear, K. B. G., McIntyre, O. R., Ozer, H., Ellerton, J., Cannellos, G., Duggan, B., and Schiffer, C. (1993). A randomized clinical trial comparing Melphalan/Prednisone with and without α-2b interferon in newly-diagnosed patients with multiple myeloma: A cancer and leukemia group B study. *Journal of Clinical Oncology* **11,** 155–160.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, California: Lifetime Learning.

Lipsitz, S. R., Dear, K. B. G., Laird, N. M., and Molenberghs, G. (1998). Tests for homogeneity of the risk difference when data are sparse. *Biometrics* **54,** 148–160.

Lui, K.-J. and Kelly, C. (2000). A revisit on tests for homogeneity of the risk difference. *Biometrics,* **56,** 197–203.

Petitti, D. B. (1994). *Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis. Methods for Quantitative Synthesis in Medicine.* Oxford: Oxford University Press.

Shadish, W. R. and Haddock, C. K. (1994). Combining estimates of effect size. In *The Handbook of Research Synthesis,* H. Cooper and L. Hedges (eds), 261–281. New York: Russell Sage Foundation.

Received April 1999. *Revised June* 1999. *Accepted July* 1999.