BIOMETRICS 55, 732–737
September 1999

Valid Inference in Random Effects Meta-Analysis

Dean A. Follmann* **and Michael A. Proschan**

Office of Biostatistics Research, National Heart Lung and Blood Institute,
2 Rockledge Center MSC 7938, Bethesda, Maryland 20892-7938, U.S.A.
**email:* follmann@helix.nih.gov

SUMMARY. The standard approach to inference for random effects meta-analysis relies on approximating the null distribution of a test statistic by a standard normal distribution. This approximation is asymptotic on k, the number of studies, and can be substantially in error in medical meta-analyses, which often have only a few studies. This paper proposes permutation and *ad hoc* methods for testing with the random effects model. Under the group permutation method, we randomly switch the treatment and control group labels in each trial. This idea is similar to using a permutation distribution for a community intervention trial where communities are randomized in pairs. The permutation method theoretically controls the type I error rate for typical meta-analyses scenarios. We also suggest two *ad hoc* procedures. Our first suggestion is to use a t-reference distribution with $k-1$ degrees of freedom rather than a standard normal distribution for the usual random effects test statistic. We also investigate the use of a simple t-statistic on the reported treatment effects.

KEY WORDS: Clinical trials; Permutation test; Randomization test; t-test.

1. Introduction

Meta-analysis is an important tool used in medical research to quantitatively summarize multiple related studies. There are two major statistical approaches to testing in meta-analysis. The fixed effects approach assumes that all studies are governed by a common treatment effect and tests whether this single effect is zero. The random effects approach allows different treatment effects for different studies and tests whether the mean effect differs from zero. Both have their place, although there is increasing emphasis on the results of the random effects analysis (National Research Council, 1992; Mosteller and Colditz, 1996).

Testing under either approach typically depends on an asymptotic approximation of a test statistic to a standard normal distribution. For fixed effects models, the approximation is asymptotic on the total number of subjects, whereas for random effects models, it is asymptotic on the number of studies, k. Although the total number of subjects is generally large in meta-analyses, they typically entail fewer than 20 studies. For example, DerSimonian and Laird (1986) described seven medical meta-analyses, six of which had $k < 10$. The use of the asymptotic approximation for relatively few studies can substantially inflate the type I error rate. This is analogous to using a standard normal null distribution for a t-test.

This paper investigates procedures that address the problem of type I error rate inflation. We show that using a t-reference distribution with $k-1$ degrees of freedom (t_{k-1}) usually inflates the type I error rate by a modest amount for realistic scenarios. We also show that applying a t-test to the k observed treatment effects, whereby each study is weighted equally, seems to control the type I error rate. To ensure an α-level procedure, we propose a group permutation method where we permute the treatment and control group labels *en masse* within each trial. This provides a permutation reference distribution for the usual random effects estimate of the treatment effect. This is similar to the permutation methods for a cluster randomized trial where pairs of communities are randomized to treatment or control groups (Gail et al., 1996). Confidence intervals based on the permutation distribution are also derived.

We illustrate our methods by performing a meta-analysis of the effect of serum cholesterol reduction. Several meta-analyses of this issue have been conducted with conflicting conclusions (Thompson, 1993). In this paper, we focus on eight studies of patients with no history of heart attacks. Although use of cholesterol-lowering drugs for the secondary prevention of heart attacks is well accepted, the use of such drugs in relatively healthy patients is more controversial. A strongly significant benefit of cholesterol reduction is suggested using a standard random effects analysis. However, this significance is based on the approximation $k \approx \infty$, where $k = 8$. We reanalyze these trials using the proposed methods.

2. Setup

Suppose that there are k clinical trials and that X_i is the estimated treatment effect in the ith trial. Under the classic random effects model, it is assumed that

$$X_i = \Delta_i + \epsilon_i,$$

where Δ_i is the true effect of the treatment in the ith study, assumed to be normally distributed with mean μ and variance τ^2, and ϵ_i is an independent "error" term assumed to be normally distributed with mean zero and variance σ_i^2. For studies with a continuous endpoint, X_i is often given by the treatment minus the control difference in mean response. For studies with a binary outcome, such as survival, X_i is often given by an estimate of the log of the odds ratio. If the studies are large, then ϵ_i should be roughly normal, regardless of the endpoint.

We assume that the main goals of the meta-analysis are to test whether μ is zero and to form a confidence interval. A standard way (cf., DerSimonian and Laird, 1986) to do this is to replace the σ_i^2's by their estimated values, as reported in the individual trials ($\widehat{\sigma}_i^2$), and then estimate τ^2 by

$$\widehat{\tau}^2 = \max\left(0, \frac{\sum v_i(X_i - \widehat{\mu}_0)^2 - (k-1)}{\sum v_i - \sum v_i^2/\sum v_i}\right), \qquad (1)$$

where $v_i = 1/\widehat{\sigma}_i^2$ and $\widehat{\mu}_0 = \Sigma v_i X_i/\Sigma v_i$. Based on $\widehat{\tau}^2$, one estimates μ by

$$\widehat{\mu} = \sum_{i=1}^{k} \widehat{w}_i X_i, \qquad (2)$$

where $\widehat{w}_i = (\widehat{\tau}^2 + \widehat{\sigma}_i^2)^{-1}/\{\Sigma_j(\widehat{\tau}^2 + \widehat{\sigma}_j^2)^{-1}\}$. If all variances were known, then $\widehat{\mu}$ would have a normal distribution with mean μ and known variance $\text{var}[\widehat{\mu}] = \{\Sigma_j 1/(\tau^2 + \sigma_j^2)\}^{-1}$. Because the variances are not known, $\widehat{\mu}$ is treated as approximately normal with mean μ and estimated variance $\widehat{\text{var}}[\widehat{\mu}] = \{\Sigma_j 1/(\widehat{\tau}^2 + \widehat{\sigma}_j^2)\}^{-1}$. One tests whether $\mu = 0$ by forming $R = \widehat{\mu}/(\widehat{\text{var}}(\widehat{\mu}))^{1/2}$ and comparing it to a standard normal distribution. We will call this test R_Z. Ninety-five percent confidence intervals take the form $\widehat{\mu} \pm 1.96(\widehat{\text{var}}(\widehat{\mu}))^{1/2}$.

The use of a standard normal reference distribution is justified asymptotically as k approaches infinity. For small values of k, the normal approximation can be poor. Suppose $\sigma_i^2 = \sigma^2$ so that X_i is normally distributed with mean μ and variance $\sigma^2 + \tau^2$. If we take $\widehat{\tau}^2 = \{\Sigma v_i(X_i - \widehat{\mu}_0)^2 - (k-1)\}/\{\Sigma v_i - \Sigma v_i^2/\Sigma v_i\}$, rather than the maximum of this and zero, then R reduces to $\overline{X}/(S_X^2/k)^{1/2}$, where S_X^2 is the sample variance. Thus, R has a t_{k-1} null distribution, and we can calculate the type I error rate, $P(|R| > 1.96)$. For an $\alpha = 0.05$ test, the rejection rates are .14, .09, .07, and .06 for $k = 4, 8, 16$, and 32, respectively, which are substantially greater than the .05 rate for typical k's.

The preceding analysis suggests that as an *ad hoc* measure, one might use a t_{k-1} null distribution for R. See also Raghunathan and Ii (1993) for a heuristic argument. We will call this test R_T. As another *ad hoc* measure, one could apply a simple t-test to the observed X_i's, e.g., $T = \overline{X}/(S_X^2/k)^{1/2}$. If the σ_i^2's are all equal, then this has an exact t_{k-1} null distribution. With unequal sample sizes, Efron (1969) argued that using a t-distribution will be conservative in most cases, provided that each X_i is symmetric about zero on the null hypothesis. Such orthant symmetry is implied by the standard null assumptions, which have X_i normal with mean zero and variance $\tau^2 + \sigma_i^2$. The numerator in T, $\overline{X}$, is obtained as (2) with $\tau^2 \to \infty$. The usual fixed effects estimate of μ is given by (2) with $\tau^2 = 0$. Thus, this simple t-test can be viewed as a counterpart to the usual fixed effects test.

A valid reference distribution for $\widehat{\mu}$ can be derived by a rerandomization argument, provided that X_i is symmetric. This is similar to using a permutation distribution in community randomized trials where pairs of communities are randomized to treatment and control groups (see Gail et al. [1996] for a thorough investigation). In meta-analysis, the treatment and control groups within a trial form a pair. Under the hypothesis H_0: $\mu = 0$, the sign of X_i is equally likely to be positive or negative for symmetric X_i. As this method corresponds to changing treatment and control labels *en masse* in each trial, we call this a group permutation method. There are 2^k possible permutations of the signs of X_i: let the pth permutation be given by $\boldsymbol{Z}^p$, where Z_i^p is $+1$ or -1. The pth permuted data set is $\boldsymbol{X}^p = (Z_1^p|X_1|, \ldots, Z_k^p|X_k|)'$. Let $\widehat{w}_i^p = \{(\widehat{\tau}^p)^2 + \widehat{\sigma}_i^2\}^{-1}/[\Sigma_j\{(\widehat{\tau}^p)^2 + \widehat{\sigma}_j^2\}^{-1}]$, where $(\widehat{\tau}^p)^2$ is (1) based on $\boldsymbol{X}^p$. Note that $\widehat{w}_i^p$ is solely a function of the $|X_i|$'s and $\boldsymbol{Z}^p$. Conditional on the $|X_i|$'s and under orthant symmetry, the signs $\boldsymbol{Z}$ of the X_i's are i.i.d. Bernoulli(1/2). A valid null distribution for $\widehat{\mu}$ can be obtained by fixing the $|X_i|$'s and considering all (equally likely) values of

$$\widehat{\mu}^p = \sum_{i=1}^{k} \widehat{w}_i^p Z_i^p |X_i|. \qquad (3)$$

Note that the weights $\widehat{w}_i^p$ are reestimated with each permutation. It is tempting to simplify matters by using the original weights $\widehat{w}_i$ in (3), but this makes the implicit assumption that given the $|X_i|$'s and the $\widehat{w}_i$'s, the signs of the X_i's are i.i.d. Bernoulli(1/2). This is not the case because the $\widehat{w}_i$'s depend on $\widehat{\tau}^2$, which in turn depends on $\Sigma v_i X_i/\Sigma v_i$.

Programming the permutation procedure is straightforward. For small values of k, the exact permutation distribution can be calculated as follows:

(1) Set a $2^k \times k$ $\boldsymbol{Z}$ matrix of ± 1's with the pth row $\boldsymbol{Z}^p$ a unique permutation. Set $p = 1$.
(2) Make rerandomized data $\boldsymbol{X}^p = (Z_1^p|X_1|, \ldots, Z_k^p|X_k|)'$.
(3) Estimate $\widehat{\mu}^p$ from formulas (1) and (2) using $\boldsymbol{X}^p$.
(4) Store $\widehat{\mu}^p$ and increment p by 1. If $p > 2^k$, stop. Otherwise, go to step 2.

If 2^k is too large, the permutation distribution can be approximated by making a large $B \times k$ matrix $\boldsymbol{Z}^*$, where each row is k independent realizations of a ± 1 Bernoulli(1/2) random variable. We then replace $\boldsymbol{Z}$ with $\boldsymbol{Z}^*$ and 2^k with B in the previously mentioned algorithm.

The two-sided p value based on the group permutation method is the proportion of $|\widehat{\mu}^p|$'s $\geq |\widehat{\mu}|$. For example, if $\widehat{\mu}$ is the second-largest value among the 2^k unique $\widehat{\mu}^p$'s, the two-sided p value is $4/2^k$. Thus, we need $k \geq 6$ to even assign a two-sided p value < 0.05. For discrete ϵ_i, e.g., X_i is the difference in proportions, there may be redundant or zero $|X_i|$'s, and $k = 6$ may be insufficient to obtain a p value ≤ 0.05. An S-plus program is available from the first author.

3. Permutation Confidence Intervals

To calculate an exact permutation confidence interval, we proceed by considering the family of tests defined by the family of null hypotheses H_0^c: $\mu = c$, and then by determining the lower and upper limits. The upper limit c_U is such that the one-tailed p value for the test of $\mu = c_U$ versus $\mu < c_U$ is $\alpha/2$. For simplicity, we limit our discussion to the determination of c_U; the determination of c_L is identical.

To test H_0^c, we transform the data to $X_i - c$ and calculate $\widehat{\mu}$ based on these data; e.g., $\widehat{\mu}(c) = \widehat{\mu}(0) - c$, where $\widehat{\mu}(0)$ is the estimate of μ based on the original data. We then reject the null hypothesis H_0^c: $\mu = c$ in favor of the alternative $\mu < c$ if $\widehat{\mu}(c)$ lies at or to the left of the $100\alpha/2$ percentile of the permutation distribution of $\widehat{\mu}(c)$. It remains to determine the smallest c satisfying this condition in an efficient manner. The solution is c_U.

Consider the pth permutation $Z_1^p, \ldots, Z_k^p$. The estimate of $\mu(c)$ based on this permutation is given by

$$\widehat{\mu}^p(c) = \sum_{i=1}^{k} \widehat{w}_i^p Z_i^p (X_i - c) = \widehat{\mu}^p(0) - c\beta^p,$$

where $\beta^p = \Sigma_i \widehat{w}_i^p Z_i^p$. To determine c_U, it is sufficient to determine $p^-(c)$, the permutation p value associated with the test of H_0^c versus the alternative $\mu < c$, and then find the smallest c such that $p^-(c) \le \alpha/2$. At any c we have

$$\begin{aligned} p^-(c) &= 1/2^k \#\{p : \widehat{\mu}^p(c) \le \widehat{\mu}(c)\} \\ &= 1/2^k \#\{p : \widehat{\mu}^p(0) - c\beta^p \le \widehat{\mu}(0) - c\} \\ &= 1/2^k \#\{p : \widehat{\mu}(0) - \widehat{\mu}^p(0) \ge c(1 - \beta^p)\}, \end{aligned}$$

where $\#\{p : \mathcal{S}(p)\}$ is the number of p's satisfying $\mathcal{S}(p)$. But the last term is an integer-valued decreasing function of c, so it suffices to look at the jump points: $\{c : \widehat{\mu}(0) - \widehat{\mu}^p(0) = c(1 - \beta^p)\ p = 1, \ldots, 2^k\}$. Let the $2^k - 1$ ordered solutions be $c_{(1)}, \ldots, c_{(2^k-1)}$ (we include no solutions of $\{c : 0 = c0\}$ which occurs when $\boldsymbol{Z}^p = 1$). We set c_U as the smallest of the $c_{(p)}$s such that $p^-(c) \le \alpha/2$.

4. Simulation

A simulation was performed to evaluate the different procedures in terms of the type I error rate and power. We generated data according to the model $X_i = \Delta_i + \epsilon_i$, where the Δ_i's were drawn from a normal distribution with mean zero and variance τ^2, and the ϵ_i's were normally distributed with mean zero and variance σ_i^2. We let $k = 2, 4, 8, 16$, and 32 and take $\tau^2 = r\overline{\sigma^2}$, where $\overline{\sigma^2}$ is the average of the σ_i^2's, and we let $r = 0, .5, 1$, and 10.

To specify the σ_i^2's, we began by approximating the distribution of the log-transformed sample sizes ($\log_{10}(N)$) for three similar meta-analyses with continuous outcome by a gamma distribution using the method of moments (Allender et al., 1996; Cutler, Follmann, and Allender, 1997; Whelton et al., 1997). The 10th, 50th, and 90th percentiles for N based on this fitted distribution are 14, 40, and 144, respectively. We wanted the σ_i^2's to roughly mimic the heterogeneity seen in practice. Thus, for each simulated meta-analysis, we drew 1000 $\boldsymbol{N}$'s, each of dimension k, where each element is i.i.d. from the preceding distribution. For each $\boldsymbol{N}$, we calculated the sample variance and then ordered the $\boldsymbol{N}$'s based on these 1000 sample variances. For the simulated meta-analysis, we would choose the $\boldsymbol{N}$ corresponding to the qth percentile of the sample variances of the $\boldsymbol{N}$'s. Finally, we specified $\sigma_i^2 \propto 1/N_i$. We let $q = .10, .50$, and .90 to reflect meta-analyses with little, typical, and substantial amounts of heterogeneity among the N's. To help interpret this, note that in 1 time out of 10, a meta-analyst could expect more variability among the N's than for our $q = 0.90$ scenario. For each unique combination of k, r, and q, 10,000 meta-analyses were simulated.

For each replication of the simulation, we calculate the test statistic R as $\widehat{\mu}/(\widehat{\text{var}}[\widehat{\mu}])^{1/2}$, taking the σ_i^2's as known at their true values and estimating τ^2 using (1). We compare R to the standard normal and t_{k-1} null distributions, denoted by R_Z and R_T, respectively. We also calculate $T = \overline{X}/(S_X^2/k)^{1/2}$ and perform the permutation test, which we call $\widehat{\mu}_P$. For $k \le 8$, the permutation distribution is obtained by enumerating all 2^k values. For $k > 8$, the permutation distribution is approximated by 1024 randomly selected values.

The middle columns of Table 1 present the simulated type I error rates. For brevity, only the results for $k = 4, 8$, and 16 are given. Use of a standard normal null distribution results in substantial type I error inflation unless $\tau^2 = 0$. Use of the

Table 1

Rejection rates for four test procedures. Each line is based on 10,000 *simulated meta-analyses.*

	q = Percentile of		Null rejection rates				Power		
k	$S^2(N_i)$'s	$\tau^2/\overline{\sigma^2}$	R_Z	R_T	$\widehat{\mu}_P$	T	R_T	$\widehat{\mu}_P$	T
4	.10	0	.03	.00	.00	.04	.70	.00	.72
		.5	.08	.00	.00	.05	.77	.00	.75
		1	.10	.01	.00	.05	.78	.00	.76
		10	.15	.05	.00	.05	.79	.00	.78
	.50	0	.04	.00	.00	.04	.68	.00	.62
		.5	.10	.02	.00	.05	.77	.00	.73
		1	.12	.03	.00	.04	.79	.00	.75
		10	.16	.05	.00	.05	.79	.00	.79
	.90	0	.03	.00	.00	.03	.64	.00	.47
		.5	.15	.04	.00	.05	.79	.00	.69
		1	.16	.05	.00	.04	.79	.00	.74
		10	.17	**.06**	.00	.05	**.79**	.00	.78
8	.10	0	.04	.02	.05	.05	.72	.71	.67
		.5	.08	.04	.05	.05	.79	.75	.75
		1	.09	.05	.05	.05	.80	.76	.77
		10	.09	.05	.04	.05	.79	.78	.80
	:50	0	.04	.01	.05	.04	.71	.65	.57
		.5	.08	.04	.04	.04	.80	.74	.72
		1	.10	**.06**	.05	.05	**.80**	.76	.76
		10	.10	**.06**	.05	.05	**.80**	.77	.80
	.90	0	.04	.01	.05	.04	.68	.51	.42
		.5	.11	**.07**	.05	.05	**.80**	.72	.69
		1	.12	**.07**	.05	.05	**.80**	.75	.74
		10	.11	**.07**	.05	.05	**.81**	.78	.80
16	.10	0	.04	.03	.05	.05	.75	.75	.61
		.5	.07	.05	.05	.05	.80	.78	.73
		1	.07	**.06**	.05	.05	**.80**	.79	.76
		10	.07	.05	.04	.05	.80	.79	.80
	.50	0	.04	.02	.05	.05	.73	.72	.52
		.5	.08	**.06**	.05	.05	**.80**	.78	.71
		1	.08	**.06**	.05	.05	**.80**	.78	.74
		10	.08	**.06**	.05	.05	**.80**	.80	.80
	.90	0	.04	.02	.05	.05	.70	.62	.39
		.5	.09	**.07**	.05	.05	**.81**	.77	.70
		1	.09	**.06**	.05	.05	**.80**	.78	.74
		10	.09	**.06**	.05	.05	**.80**	.80	.80

t_{k-1} reference distribution provides much better protection, although type I error rates of 0.06 to 0.07 often occur with $q \neq 0.1$. For $k = 32$, rejection rates for R_Z and R_T are similar and slightly better than the R_T rejection rates for $k = 16$. For $k = 2$, R_Z performs much worse than for $k = 4$, whereas R_T has rejection rates $< .04$ and generally $< .0001$. With $q = .99$, or extreme variability among the σ_i^2's, R_T had type I error rates of .06 to .10 for scenarios with inflation with $q = .90$ (not shown). We also performed simulations with 75% of the σ_i^2's $=1$ and 25% of them $= 100$, varying k and τ^2 as in Table 1. For this extreme setting, the type I inflation of R_Z and R_T was greater than that for Table 1, e.g., for R_T with $\tau^2 = \overline{\sigma}^2$, the rates are .13, .15, and .09 for $k = 4, 8$, and 16, respectively. The rejection rates for $\widehat{\mu}_P$ and T did not exceed .05 for any of the settings.

Given a symmetric distribution for the Δ_i's, the permutation method theoretically controls the type I error rate if ϵ_i, and hence X_i, is symmetric. Asymmetry of the distribution of X_i requires small and unequal sample sizes between treatment and control groups within trials and an asymmetric distribution for ϵ_i, e.g., a binomial with $\theta \neq .5$. To explore whether this is a practical concern, we also conducted a simulation with $X = \overline{Y}_T - \overline{Y}_C$ and $\overline{Y}_T$ $(\overline{Y}_C)$ the mean of a binomial with 50 (10) Bernoulli outcomes with probability θ. For each simulated meta-analysis, θ was drawn from a beta distribution with $\mu = E[\theta] = .1$ and a fixed $\tau^2 = \text{var}[\theta] = 0, .0001, .001$, or .01. A fivefold difference in sample sizes is extreme, but we found no inflation of the type I error rate for any scenarios with $k = 4, 8, 16$, or 32. Part of the reason is that for discrete distributions, the number of unique nonzero $|X_i|$'s may well be less than k. This reduces the number of possible p values and makes it harder to have a p value less than .05.

We next evaluate the power for the $R_T, \widehat{\mu}_P$, and T tests: R_Z is not evaluated because it has unacceptable type I error inflation. The use of R_T can inflate the type I error rate; however, some may feel that the amount of inflation is acceptable. We mimic the scenarios for null simulations, but choose $E[\Delta_i] = \mu = (\text{var}[\widehat{\mu}])^{1/2}\{t(k-1, .025) + t(k-1, .20)\}$ so that at least when the σ_i^2's are all the same, the simple t-test should have power around .80. The results are given in the last three columns of Table 1.

For $q = .1$, where the R_T test generally did not inflate the type I error rtae for any τ^2, the R_T test generally does best: for $k = 4$, the R_T and T tests are similar and best, whereas for $k = 8$, R_T generally does better than the other two tests. For $k = 16$, the R_T and $\widehat{\mu}_P$ tests are similar and best. For $q = .5$ or .9, the R_T test almost never has less power than the other tests; however, this advantage is partially due to its type I error inflation on the null distribution. If type I error control is important, then $\widehat{\mu}_P$ and T are the only options. Provided $k = 8$ or 16, the $\widehat{\mu}_P$ test is similar or better than the T test. The relative inferiority of the t-test increases with smaller τ^2. For $k = 4$, T is preferred, as $\widehat{\mu}_P$ literally has no power because a permutation p value ≤ 0.05 is not possible.

The power for R_T is unfairly inflated when it is not an α-level procedure. To make a fairer comparison, we calculated an approximate deflation term for the power of R_T, reasoning as follows. If an $\alpha = 0.07$ t-test based on $k = 8$ studies has a power of .81, one can calculate that the power for an $\alpha = 0.05$ t-test would be .75. Thus, as a crude approximation, one could replace the .81 for R_T in the last row for $k = 8$ with .75. Overall, we found that for each percent of type I error inflation, power should be reduced by about .05, .03, and .02 for $k = 4, 8$, and 16, respectively, under this approximation. If these adjustments are made, the power of R_T is reduced so that the permutation test is competetive or superior for $k = 8$ and 16, provided that $q \neq .1$ and $\tau^2 \neq 0$.

5. Example

Thompson (1993) presented a thorough meta-analysis of randomized trials that tested the effect of reducing serum cholesterol on ischemic heart disease. One analysis of interest was the subset of eight trials that enrolled patients without histories of heart attacks. Table 2 presents the raw data, along with the estimated treatment effect X_i and its within-study variance σ_i^2, estimated using Peto's method (Yusuf et al., 1985). Using (1), $\widehat{\tau}^2$ is estimated to be .002. Note that X_i is an estimate of the log-odds ratio.

The random effects test statistic $R = -.191/.059 = -3.24$ gives a p value of 0.001 when compared with a standard normal distribution. However, this approach is not defensible here due to its type I error inflation. We next apply the simple t-test and group permutation methods of this paper to these data. Here, T is $\overline{X}/(S_X^2/8)^{1/2} = -2.38$, which gives a p value of 0.05 and a 95% confidence interval of $\exp(-.353 \pm 2.36(.175/8)^{1/2}) = (.50, 1.00)$. The group permutation procedure gives an exact p value of 0.04 and an exact permutation confidence interval of (.72, .97), which is not symmetric about the point estimate of $\exp(\widehat{\mu}) = \exp(-.191) = .83$.

Another possibility is to use a t_7-reference distribution for $R = -3.24$. Even though there is substantial heterogeneity among the σ_i^2's, the type I error rate should be inflated by, at most, a modest amount. This procedure provides a p value for R of .014 and a 95% confidence interval of $\exp(-.191 \pm 2.36(.059)^{1/2})$=(.72, .95). All three proposed procedures have p values between 0.01 and 0.05, suggesting a significant benefit of cholesterol-lowering drugs for the primary prevention of heart attacks.

As this is just one data set, we explore how the four procedures perform under repeated sampling with these σ_i^2's. We generate data, as for Table 1, but with the σ_i^2's given by Table 2. The results are given in Table 3. The use of R_Z roughly doubles the type I error rate, whereas the use of R_T inflates the type I error rate to .06 or .07, provided that $\tau^2 \neq 0$. As

Table 2
Data from Thompson (1993) *for a meta-analysis of the effect of serum cholesterol in patients with no history of heart attacks*

	IHD events/sample size			
Study	Treatment	Control	X_i	σ_i^2
WHO	173/5331	210/5296	−.207	.011
LRC	157/1906	193/1900	−.230	.013
Minnesota	131/4541	121/4516	.076	.016
Helsinki	56/2051	84/2030	−.425	.030
Los Angeles	52/424	65/422	−.263	.040
Upjohn	36/1149	42/1129	−.177	.053
EXCEL	62/6582	20/1663	−.265	.076
McCaughan	2/88	2/30	−1.330	1.353

Table 3

Simulated rejection rates for a meta-analysis, like that in Table 2. *Each line is based on* 10,000 *simulated meta-analyses.*[a]

		Rejection rates			
Hypothesis	$\tau^2/\overline{\sigma}^2$	R_Z	R_T	$\widehat{\mu}_P$	T
Null	0	.04	.01	.05	.02
	.5	.11	**.06**	.05	.03
	1	.10	**.06**	.05	.04
	10	.11	**.07**	.05	.05
Alternative	0	*	.71	.64	.31
	.5	*	**.80**	.75	.60
	1	*	**.80**	.75	.67
	10	*	**.80**	.77	.79

[a] * denotes that the test substantially inflates the type I error rate.

τ^2 is unknown and estimated to be nonzero, the use of R_T subjects the meta-analyst to potential criticism. The use of the permutation distribution for $\widehat{\mu}$ and the simple t-test do not inflate the type I error rate. Under the alternative, the unadjusted power of R_T is never worse than the permutation test, which is generally better than the simple t-test. Incorporating the simple power deflation from Section 4 to R_T gives deflated powers of 71, .77, .77, and .74 for the 4 $\tau^2/\overline{\sigma}^2$'s, which are quite similar to the $\widehat{\mu}_P$ for $\tau^2 \neq 0$.

6. Discussion

The usual approach to random effects meta-analysis suffers from serious type I error rate inflation under common scenarios. The inflation is worst for meta-analyses with relatively few studies, but with 16 studies, the type I error rate can be doubled, and even with 32 studies, the type I error rate can be inflated by 40–60%. This paper has proposed three different methods for analyzing random effects meta-analyses without this drawback. If control of the type I error rate is important, we recommend using either group permutation methods or a simple t-test. The group permutation method should be used, provided that there are enough unique nonzero $|X_i|$'s. For continuous X_i, we found that $k \geq 7$ worked well (simulations are not reported). Otherwise, the simple t-test can be used. The usual random effects test compared with a t_{k-1} distribution was also suggested. This provides a much improved procedure compared with using a standard normal null distribution. For the simulations in this paper, there was generally at most a modest 20–40% type I error inflation, though the inflation can be greater. If this amount of uncertainty about the type I error rate is acceptable, then R_T will be a simple and attractive option.

The group permutation method can also be applied more generally. For example, $\overline{X}$ or $\widehat{\mu}_0$ could be used as test statistics rather than $\widehat{\mu}$. They should be more powerful than $\widehat{\mu}$, provided that τ^2 is quite large or close to zero, respectively. More generally, meta-analyses of observational studies, which summarize each study with a slope estimate, could also apply group permutation ideas. If there is truly no relationship between Y and X, then the slope is as likely to be positive as negative and a permutation-type distribution could be used here as well. Finally, a two-stage regression model, e.g., $X_i = \Delta_i + Z_i\beta + \epsilon_i$, might be postulated, where Z_i is a study-level covariate such as the baseline difference in blood pressure. Group permutation methods could be applied here to simultaneously test whether $E[\Delta_i]$ and β equal zero. Such a test should be more powerful than $\widehat{\mu}_P$ if the postulated model holds. See Gail et al. (1996) for a discussion of permutation tests for a similar setting.

Acknowledgements

The authors wish to thank Drs Craig Borkowf, Margaret Wu, and Nancy Geller for their helpful comments that improved this manuscript.

Résumé

L'approche standard pour l'inférence dans les meta-analyses à effets aléatoires repose sur une approximation de la distribution de la statistique du test sous l'hypothèse nulle, par une loi normale. Cette approximation est asymptotique en k, le nombre d'études, et peut être substantiellement erronée dans les meta-analyses médicales, qui souvent ont peu d'études. Ce papier propose des méthodes de permutation et *ad hoc* pour tester le modèle à effets aléatoires. Avec la méthode de permutation par groupe, nous permutons aléatoirement l'identité des groupes traitement et contrôle dans chaque essai. L'idée est similaire à l'utilisation d'une distribution de permutation pour une essai d'intervention dans des communautés où les communautés sont randomisées par paires. La méthode des permutations contrôle en théorie le risque de première espèce pour les scénarios typiques de meta-analyses. Nous suggérons aussi deux procédures ad hoc. Notre première suggestion est d'utiliser une distribution t de référence avec $k-1$ degrés de liberté plutôt que la loi normale pour les tests statistiques usuels des effets aléatoires. Nous étudions aussi l'usage d'une simple test de tZ pour les effets traitement rapportés.

References

Allender, P. S., Cutler, J., Follmann, D. A., Cappuccio, F., Pryer, J., and Elliott, P. (1996). Dietary calcium and blood pressure: An overview of randomized clinical trials. *Annals of Internal Medicine* **124,** 825–831.

Cutler, J., Follmann, D. A., and Allender, P. (1997). Randomized trials of sodium reduction: An overview. *American Journal of Clinical Nutrition* **65,** 643S–651S.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7,** 177–188.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association* **64,** 1278–1302.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15,** 1069–1092.

Mosteller, F. and Colditz, G. A. (1996). Understanding research synthesis (meta-analyses). *Annual Review of Public Health* **17,** 1–23.

National Research Council (1992). *Combining Information: Statistical Issues and Opportunities for Research.* Washington: National Academy Press.

Raghunathan, T. E. and Ii, Y. (1993). Analysis of binary data from a multicentre trial. *Biometrika* **80,** 127–139.

Thompson, S. G. (1993). Controversies in meta-analysis: The case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research* **2,** 173–192.

Whelton, P. K., He, J., Cutler, J. A., Brancati, F. L., Appel, L. J., Follmann, D. A., and Klag, M. J. (1997). The effects of oral potassium on blood pressure: A quantitative overview of randomized, controlled clinical trials. *Journal of the American Medical Association* **277,** 1624–1632.

Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Progress in cardiovascular disease. **27,** 335–371.

Received March 1998. *Revised August* 1998.
Accepted September 1998.