Combined Test Procedures in the Meta-Analysis of Controlled Clinical Trials

Guido KNAPP and Joachim HARTUNG

Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

Abstract. Concerning the actual significance level, we investigate the commonly used combined test procedures of meta-analysis, which contain the choice of the model, in which the analysis is carried out, and the commonly used tests for treatment effect in the fixed and random effects model, and some new combined test procedures, which use an alternative test statistic in the random effects model or the t-distribution as test distribution of the commonly used test statistics. A simulation study indicates that the new combined test procedures are better with respect to a prescribed significance level.

1. Introduction

In this paper we consider the test for treatment effect in the meta-analysis of controlled clinical trials, i. e. we want to judge if an overall treatment effect is present given stochastically independent study-specific estimates of the treatment effect. The test for treatment effect is carried out either in the fixed effects model of meta-analysis assuming homogeneous treatment effects in all clinical trials or in the random effects model of meta-analysis if heterogeneity of the study-specific treatment effects is present. Before using the test for treatment effect one usually makes a decision which of the two models one takes and this decision may affect the performance of the test for treatment effect in the chosen model.

Often, one can find the proposal that the choice of the model should be based on the test of homogeneity, cf. for instance [1]. But the test of homogeneity in most cases has too low power to detect differences between the study-specific treatment effects and the false use of the fixed effects model, if heterogeneity is present, can lead to a substantial increase of the type I error rate of the commonly used test for treatment effect as pointed out in [2].

Another criterion for the choice between the two models is given in [3] where it is recommended to use the fixed effects model if the method of moments estimator of the between-study variance proposed in [4] yields a negative estimate, and if the estimated value of the between-study variance is positive, the random effects model should be used.

As tests for treatment effect we consider the commonly used ones in the fixed effects and in the random effects model, respectively, and the alternative test proposed in [5] for the random effects model. It is known that the commonly used tests for treatment effect may lead to a large number of unjustified significant evidences even if one carries out the analysis in the correct model, cf. for instance [6] and [7]. So, we will compare the commonly used decision rules and some new decision rules, which contain either the alternative test from [5] or the commonly used test statistics of the test for treatment effect but instead of the standard normal distribution an appropriate t-distribution is used as test distribution, cf. [8].

2. The test statistics

Let us consider k controlled clinical trials and let us denote by $\theta_1, \dots, \theta_k$ the one-dimensional parameters of interest, where each parameter stands for the treatment effect in a study. In each trial an estimate of the parameter θ_i, say $\hat{\theta}_i$, is available with an estimate of the variance of $\hat{\theta}_i$, say $\hat{\sigma}^2\left(\hat{\theta}_i\right)$. It is usually assumed that the study-specific estimators $\hat{\theta}_i$, are at least approximately normally distributed and (nearly) unbiased. If all study-specific treatment effects are equal, i. e. it holds $\theta_1 = \dots = \theta_k = \theta$, then the feasible estimator of the overall treatment effect θ in the fixed effects model is given by

$$\hat{\theta}_{FE} = \sum_{i=1}^{k} \frac{\hat{v}_i}{\hat{v}} \hat{\theta}_i , \quad \hat{v} = \sum_{i=1}^{k} \hat{v}_i , \quad \hat{v}_i = \left[\hat{\sigma}^2 \left(\hat{\theta}_i \right) \right]^{-1} , \quad i = 1, \ldots, k .$$

The test statistic for the test for treatment effect, i. e. the hypothesis is $H_0\!:\theta = 0$, is given by

$$T_1 = \frac{\hat{\theta}_{FE}}{\sqrt{1/\hat{v}}} = \frac{\sum_{i=1}^{k} \hat{v}_i \hat{\theta}_i}{\sqrt{\hat{v}}}$$

and the hypothesis H_0 is rejected at level α if the observed absolute value of T_1 exceeds the (1-$\alpha/2$)-quantile of the standard normal distribution denoted by $u_{1-\alpha/2}$.

The assumption of homogeneous treatment effects can be formally tested using the test statistic

$$Q = \sum_{i=1}^{k} \hat{v}_i \left(\hat{\theta}_i - \hat{\theta}_{FE} \right)^2 ,$$

which is at least approximately χ^2-distributed with (k-1) degrees of freedom under the hypothesis of homogeneity, cf. [9], and the hypothesis of homogeneity is rejected at level α if the observed value of Q exceeds the (1-α)-quantile of the χ^2-distribution with (k-1) degrees of freedom denoted by $\chi^2_{k-1;1-\alpha}$.

In the random effects model the study-specific treatment effects $\theta_i, i = 1, \ldots, k$, may vary, and it is assumed that θ_i, $i = 1, \ldots, k$, are stochastically independent normally distributed random variables with mean θ and variance τ^2. Then, the marginal distribution of the study-specific estimators $\hat{\theta}_i$ is the normal distribution with mean θ and variance $\tau^2 + \sigma^2\left(\hat{\theta}_i\right)$, cf. [3]. The parameter τ^2 is often called between-study variance.

The feasible estimator of the overall treatment effect θ in the random effects model is given by

$$\hat{\theta}_{RE} = \sum_{i=1}^{k} \frac{\hat{w}_i}{\hat{w}} , \quad \hat{w} = \sum_{i=1}^{k} \hat{w}_i , \quad \hat{w}_i = \left[\hat{\tau}^2 + \hat{\sigma}^2 \left(\hat{\theta}_i \right) \right]^{-1} , \; i = 1, \ldots, k ,$$

where $\hat{\tau}^2$ is a nonnegative estimator of the between-study variance τ^2, for instance the truncated versions of the DerSimonian-Laird estimator $\hat{\tau}^2_{DSL}$ or of the restricted maximum likelihood estimator $\hat{\tau}^2_{REML}$, cf. [1].

The test statistic for $H_0{:}\,\theta = 0$ is then given by

$$T_2 = \frac{\hat{\theta}_{RE}}{\sqrt{1/\hat{w}}} = \frac{\sum_{i=1}^{k} \hat{w}_i \hat{\theta}_i}{\sqrt{\hat{w}}}$$

and the hypothesis H_0 is rejected at level α if the observed absolute value of T_2 exceeds $u_{1-\alpha/2}$.

In the random effects model Hartung, cf. [5], derived a new estimator for the variance of the best linear unbiased estimator of the overall treatment effect which is stochastically independent of the best estimator of θ and is distributed as a multiple of a χ^2-distribution with (k-1) degrees of freedom. Thus, an alternative test statistic of the test for treatment effect in the random effects model can be constructed. The feasible alternative test statistic is given by

$$T_3 = \frac{\hat{\theta}_{RE}}{\sqrt{\hat{\sigma}^2\left(\hat{\theta}_{RE}\right)}} \qquad \text{with} \qquad \hat{\sigma}^2\left(\hat{\theta}_{RE}\right) = \frac{1}{k-1}\sum_{i=1}^{k}\frac{\hat{w}_i}{\hat{w}}\left(\hat{\theta}_i - \hat{\theta}_{RE}\right)^2,$$

and the hypothesis $H_0{:}\,\theta = 0$ is rejected at level α if the observed absolute value of T_3 exceeds the (1-*α/2*)-quantile of the t-distribution with (k-1) degrees of freedom denoted by $t_{k-1;1-\alpha/2}$.

3. Combined test procedures

In table 1 we put together the combined test procedures for the hypothesis $H_0{:}\,\theta = 0$ which we will investigate.

The test ψ_1 reflects the usual procedure that the model choice depends on the test of homogeneity and for the test for treatment effect the commonly used test statistic is used with the standard normal distribution as test distribution. The decision rule of the test ψ_2 is quite similar to the previous one but instead of the standard normal distribution the t-distribution with (k-1) degrees of freedom is used as test distribution, cf. [8].

The test ψ_3 coincides with the proposal in [3] that the model choice should be dependent on the sign of the DerSimonian-Laird estimator $\hat{\tau}^2_{DSL}$ and then the commonly used test is applied in the corresponding model. The decision rule of the test ψ_4 resembles the proposal in [3] with the difference that the t-distribution is used as test distribution.

The test ψ_5 is the alternative test proposed in [5] for the random effects model and we will investigate this test if it is always used irrespective of the true underlying model. In contrast to the test ψ_5, the test ψ_6 requires the additional consideration of the commonly used test of the fixed effects model if the DerSimonian-Laird estimator of the between-study variance does not yield a positive estimate. The test ψ_6 is motivated for correcting a possible anti-conservative attitude of the test ψ_5 if the between-study variance is small.

Table 1: Decision rules of the test for treatment effect

Test	Decision rule: Reject $H_0{:}\theta = 0$ if
ψ_1	$\left(Q \le \chi^2_{k-1;1-\alpha}\ and\ \lvert T_1\rvert > u_{1-\alpha/2}\right)$ or $\left(Q > \chi^2_{k-1;1-\alpha}\ and\ \lvert T_2\rvert > u_{1-\alpha/2}\right)$
ψ_2	$\left(Q \le \chi^2_{k-1;1-\alpha}\ and\ \lvert T_1\rvert > t_{k-1;1-\alpha/2}\right)$ or $\left(Q > \chi^2_{k-1;1-\alpha}\ and\ \lvert T_2\rvert > t_{k-1;1-\alpha/2}\right)$
ψ_3	$\left(\hat{\tau}^2_{DSL} \le 0\ and\ \lvert T_1\rvert > u_{1-\alpha/2}\right)$ or $\left(\hat{\tau}^2_{DSL} > 0\ and\ \lvert T_2\rvert > u_{1-\alpha/2}\right)$
ψ_4	$\left(\hat{\tau}^2_{DSL} \le 0\ and\ \lvert T_1\rvert > t_{k-1;1-\alpha/2}\right)$ or $\left(\hat{\tau}^2_{DSL} > 0\ and\ \lvert T_2\rvert > t_{k-1;1-\alpha/2}\right)$
ψ_5	$\lvert T_3\rvert > t_{k-1;1-\alpha/2}$
ψ_6	$\left(\hat{\tau}^2_{DSL} > 0\ and\ \lvert T_3\rvert > t_{k-1;1-\alpha/2}\right)$ or $\left(\hat{\tau}^2_{DSL} \le 0\ and\ \lvert T_1\rvert > u_{1-\alpha/2}\ and\ \lvert T_3\rvert > t_{k-1;1-\alpha/2}\right)$

4. Simulation study

In a simulation study we investigate the six decision rules from table 1 concerning their actual significance level given a nominal significance level of $\alpha = 0.05$. We consider the meta-analysis of 10 controlled clinical trials where in each trial a new treatment is compared to a standard treatment. As parameter of interest we choose the risk difference, i. e. the difference of the probabilities of success of the two treatments. The probabilities of success are randomly chosen from the interval [0.5 ; 0.8] under the hypothesis that the overall risk difference is equal to zero.

We present the results for three different patterns of sample sizes where in each pattern the number of patients in each study and treatment group is equal and as different sample sizes we consider n_{ij} = 10, 20, and 40 for i=1, ..., 10 clinical trials and j =1, 2 treatment groups in each trial. As values of the between-study variance τ^2 we take 0, 0.01, 0.1, and 1. The between-study variance τ^2 is always estimated using the truncated version of the DerSimonian-Laird estimator.

The results of the simulation study are given in table 2 where each reported estimated type I error rate is based on 10,000 runs of the corresponding model.

Table 2: Estimated type I error rates (in %) for the hypothesis $H_0\!:\theta = 0$ in the meta-analysis of 10 controlled clinical trials using the decision rules from table 1 and the parameter of interest is the risk difference

k = 10		Tests for $H_0\!:\theta = 0$ at level $\alpha = 0.05$					
n_{ij}	τ^2	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
10	0	6.75	3.44	5.21	2.47	5.95	4.41
	0.01	9.26	5.36	6.72	3.68	6.33	5.43
	0.1	9.84	6.36	8.40	5.10	5.29	5.41
	1	8.58	5.29	8.57	5.28	5.10	5.10
20	0	5.95	3.12	4.63	2.36	5.53	3.81
	0.01	9.03	5.21	6.54	3.35	5.31	4.75
	0.1	8.27	5.30	8.03	4.94	5.03	5.06
	1	8.72	5.41	8.72	5.41	5.43	5.43
40	0	5.13	2.55	4.15	1.93	5.30	3.38
	0.01	9.80	5.90	7.17	3.93	4.97	4.84
	0.1	8.41	5.14	8.38	5.13	5.02	5.02
	1	8.35	5.27	8.35	5.27	5.21	5.21

From table 2 we see that the commonly used test procedures ψ_1 and ψ_3, which contain the model choice between the fixed and the random effects model as well as the standard normal distribution as test distribution of the test for treatment effect, have estimated type I error rates up to nearly 10% if heterogeneity is present, and even if the sample sizes within the studies increase, both tests remain rather anti-conservative. For large values of τ^2 both tests have nearly identical estimated type I error rates but for smaller positive values the test

ψ_1 always yields the larger estimated type I error rates. For $\tau^2 = 0$ we observe that the estimated type I error rates of the tests ψ_1 and ψ_3 decline for growing sample sizes within the studies. So, for small sample sizes within the studies the test ψ_3 has acceptable results concerning the prescribed significance level and becomes a bit conservative for increasing sample sizes. The test ψ_1, however, attains the nominal significance level for larger sample sizes, whereas for smaller sample sizes the test is anti-conservative.

Due to its construction the combined tests ψ_2 and ψ_4, which use the t-distribution as test distribution of the test for treatment effect, have always estimated type I error rates which are below the estimated type I error rates of the tests ψ_1 and ψ_3. Moreover, the test ψ_2 always has larger estimated type I error rates than the test ψ_4. If no or less heterogeneity is present, the tests ψ_2 and ψ_4 can become rather conservative, whereas for growing between-study variance both tests yield satisfactory and nearly identical results concerning the prescribed significance level.

The test ψ_5 almost always has estimated type I error rates near the prescribed level irrespective of the true underlying model. Only for small sample sizes and less heterogeneity the test is a bit anti-conservative. In this situation the test ψ_6 leads to an improvement concerning the actual significance level. For growing between-study variance both tests, ψ_5 and ψ_6, are nearly identical.

Acknowledgement

The support of the Deutsche Forschungsgemeinschaft (SFB 475, “Reduction of complexity in multivariate data structures”) is gratefully acknowledged.

References

[1] S.-L.T. Normand, Meta-Analysis: Formulating, Evaluating, Combining, and Reporting, *Statistics in Medicine* **18** (1999) 321-359.

[2] S. Ziegler and N. Victor, Gefahren der Standardmethoden für Meta-Analysen bei Vorliegen von Heterogenität, *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **30** (1999) 131-140.

[3] A. Whitehead and J. Whitehead, A General Parametric Approach to the Meta-Analysis of Randomized Clinical Trials, *Statistics in Medicine* **10** (1991) 1665-1677.

[4] R. DerSimonian and N.M. Laird, Meta-Analysis in Clinical Trials, *Controlled Clinical Trials* **7** (1986) 177-188.

[5] J. Hartung, An Alternative Method for Meta-Analysis, *Biometrical Journal* **41** (1999) 901-916.

[6] Y. Li, L. Shi and H.D. Roth, The Bias of the Commonly-Used Estimate of Variance in Meta-Analysis, *Communications in Statistics - Theory and Methods* **23** (1994) 1063-1085.

[7] A. Böckenhoff and J. Hartung, Some Corrections of the Significance Level in Meta-Analysis, *Biometrical Journal* **40** (1998) 937-947.

[8] D.A. Follmann and M.A. Proschan, Valid Inference in Random Effects Meta-Analysis, *Biometrics* **55** (1999) 732-737.

[9] W.G. Cochran, The Combination of Estimates From Different Experiments, *Biometrics* **10** (1954) 101-129.