BioMed Central

Research article

Open Access

Adaptive designs based on the truncated product method
Markus Neuhäuser[1] and Frank Bretz*[2]

Address: [1]Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Hufelandstr. 55, D-45122 Essen, Germany and [2]Novartis Pharma AG, WSJ-27.1.005, 4002 Basel, Switzerland

Email: Markus Neuhäuser - markus.neuhaeuser@medizin.uni-essen.de; Frank Bretz* - frank.bretz@novartis.com

* Corresponding author

Abstract

Background: Adaptive designs are becoming increasingly important in clinical research. One approach subdivides the study into several (two or more) stages and combines the *p*-values of the different stages using Fisher's combination test.

Methods: Alternatively to Fisher's test, the recently proposed truncated product method (TPM) can be applied to combine the *p*-values. The TPM uses the product of only those *p*-values that do not exceed some fixed cut-off value. Here, these two competing analyses are compared.

Results: When an early termination due to insufficient effects is not appropriate, such as in dose-response analyses, the probability to stop the trial early with the rejection of the null hypothesis is increased when the TPM is applied. Therefore, the expected total sample size is decreased. This decrease in the sample size is not connected with a loss in power. The TPM turns out to be less advantageous, when an early termination of the study due to insufficient effects is possible. This is due to a decrease of the probability to stop the trial early.

Conclusion: It is recommended to apply the TPM rather than Fisher's combination test whenever an early termination due to insufficient effects is not suitable within the adaptive design.

Background

Randomized controlled experiments were introduced by Sir Ronald A. Fisher in the 1920s for agricultural studies and not in order to compare the effects of different treatments in humans. However, according to Palmer [1] the way clinical trials are conducted today is essentially unchanged from Fisher's day. In contrast to agricultural studies most clinical trials require periodic monitoring of the accumulating data, e.g. to minimize the number of experimental patients who will continue with an inferior treatment [[2], p. 360].

Adaptive designs with at least one interim analysis can potentially be used for periodic monitoring. All information from the first stage(s) can be used to plan the following stage(s). A number of adaptive designs have been proposed recently, for an overview see Bauer et al. [3]. Here, we consider the adaptive procedure according to Bauer and Köhne [4] that uses Fisher's product test.

Let k be the number of stages (i.e., there are $k - 1$ interim analyses), and let p_i be the one-sided *p*-value observed with the *i*-th stage's data, $i = 1, ..., k$. According to Fisher's product criterion [[5], pp. 37–39] the null hypothesis H_0 can be rejected at the end of the trial if

$$\prod_{i=1}^{k} p_i \le c_\alpha = \exp\left[-\frac{1}{2} \chi^2_{2k} (1-\alpha) \right],$$

where $\chi^2_{2k} (1-\alpha)$ is the (1 - α)-quantile of the central χ^2-distribution with $2k$ degrees of freedom.

In clinical trials boundaries for early stopping after an interim analysis may be incorporated. Obviously, in the case of $p_1 \le c_\alpha$ early stopping with the rejection of H_0 is possible after stage one. In general, H_0 can be rejected after the j-th stage if $\prod_{i=1}^{j} p_i \le c_\alpha$. In addition, one may terminate the trial due to insufficient effects. A lower limit α_0 can be included so that the trial is terminated without rejecting H_0 if $p_1 \ge \alpha_0$. According to Bauer and Köhne [[4], p. 1031] a value of 0.5 may be a suitable choice for α_0. Bauer and Röhmel [[6], p. 1596] recommended $\alpha_0 = 1$ for establishing a dose-response relationship, that is, no early stopping without rejecting H_0 at all. In this context, an early stopping due to insufficient effects is not feasible since doses in a plateau region could have been used. In that case, different doses may be used in the following stage.

Note that, in case of $\alpha_0 < 1$, larger boundaries for $\prod_{i=1}^{j} p_i$ apply for early stopping with the rejection of H_0. For a two-stage design, one can reject H_0 after stage one if $p_1 \le \alpha_1$ for a value of α_1 that lies between c_α and α [4,6]. This value can be calculated iteratively using the formula [[4], p. 1032]

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_{0}^{c_\alpha / p_1} dp_2 \; dp_1 = \alpha_1 + c_\alpha \left[\ln \alpha_0 - \ln \alpha_1 \right] = \alpha.$$

As an alternative to Fisher's product test, Zaykin et al. [7] recently introduced a truncated product method for combining p-values. To be precise, instead of calculating the product of all p-values, they suggested the use of the product of only those p-values that do not exceed some fixed cut-off value τ, $0 < \tau \le 1$. The truncated product W_τ is defined as

$$W_\tau = \prod_{i=1}^{k} p_i^{I(p_i \le \tau)}$$

where $I(.)$ is the indicator function. Since the p-values of the different stages are independent,
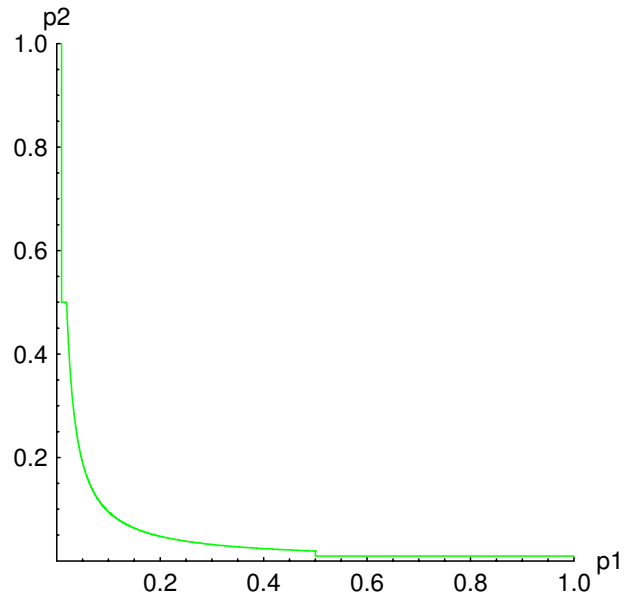
Figure 1
The rejection region of the truncated product method for k = 2 and τ = 0.5.

$$pr(W_\tau \le w) = \sum_{i=0}^{k} \binom{k}{i} (1-\tau)^{k-i} \left(w \sum_{j=0}^{i-1} \frac{(i \ln \tau - \ln w)^j}{j!} I(w \le \tau^i) + \tau^i I(w > \tau^i) \right)$$

[[7], p. 173] holds for $w < 1$ under the overall null hypothesis (i.e., under the assumption that each stage's null hypothesis is true). Figure 1 displays the rejection region for k = 2 and τ = 0.5.

When using the truncated product method, the (1-α)-quantile of the distribution of W_τ, $\tilde{c}_\alpha$, is the critical value for the combination test. Analogous to Fisher's combination test an $\tilde{\alpha}_1$ can be calculated for given α_0 such that the overall type I error rate is α.

Zaykin et al. [7] and Neuhäuser [8] investigated the truncated product method for combining a large number of p-values and demonstrated by simulation that it can provide high power. In this paper we investigate whether the truncated product method is also useful for the adaptive design described above. In contrast to previous applications [7,8] we consider classical experimental questions involving only few p-values. Very recently, a rank truncated product was proposed as a further alternative [9]. That method uses the product of the K most significant p-values where K can be chosen. Since we consider the

Table 1: Power to reject H_0 in a two-stage design with α_0 = 0.5 (combination of t tests, one-sided, α = 0.05)

δ =	0.1	0.2	0.3	0.4	0.5
25 observations per group in stage one, 75 observations per group in stage two					
τ = 1	0.149	0.343	0.595	0.808	0.929
τ = 0.5	0.153	0.352	0.605	0.815	0.931
50 observations per group and stage					
τ = 1	0.162	0.377	0.644	0.854	0.959
τ = 0.5	0.165	0.384	0.652	0.860	0.961
75 observations per group in stage one, 25 observations per group in stage two					
τ = 1	0.166	0.386	0.654	0.860	0.962
τ = 0.5	0.167	0.389	0.657	0.863	0.963

combination of 2 to 4 *p*-values only, the rank truncated product does not seem to be appropriate for our aim.

We first present the comparison of the combinations with and without truncation for designs with two stages. Afterwards, designs with more than two stages are investigated. We then illustrate the method using two examples, and conclusions are given in final section.

Methods
In order to compare the adaptive procedures with and without truncation we consider the situation of two parallel groups with means μ_1 and μ_2. There are 100 observations per stage. These observations are subdivided into two groups and are assumed to be normally distributed with a common, but unknown variance σ^2. Student's *t* test is performed in each of the two stages with a one-sided significance level of α = 5%.

The overall *p*-value, i.e. the *p*-value of the combination test, is defined as follows [10]: In case the study stops after stage 1, the overall *p*-value equals p_1. Otherwise, the overall p-value is $\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_{0}^{1} I(p_1 p_2 \leq c_\alpha) dp_2 dp_1$ for Fisher's combination test and

$\tilde{\alpha}_1 + \int_{\tilde{\alpha}_1}^{\alpha_0} \int_{0}^{1} I(p_1^{I(p_1 \leq \tau)} \cdot p_2^{I(p_2 \leq \tau)} \leq \tilde{c}_\alpha) dp_2 dp1$ for the truncated product test.

The case α_0 = 0.5
First, we consider a study that is terminated early due to insufficient effects if $p_1 \geq \alpha_0$ = 0.5. Without any truncation (i.e., τ = 1) we have c_α = 0.0087 and α_1 = 0.0233 in this case [4]. However, when we set τ = α_0 = 0.5, a smaller value for α_1 but a larger boundary for

$W_{\tau=0.5} = p_1^{I(p_1 \leq 0.5)} \cdot p_2^{I(p_2 \leq 0.5)}$ is obtained. To be precise, the trial can be terminated early with the rejection of H_0 if $p_1 \leq \tilde{\alpha}_1$ = 0.0190, and there is a significance at the end of the trial if $W_{\tau = 0.5} \leq \tilde{c}_\alpha$ = 0.0095.

Although α_1 is decreased the overall power can increase in case of truncation as the boundary $\tilde{c}_\alpha$ for $W_{\tau = 0.5}$ is larger than that for $W_{\tau = 1}$. Table I displays the overall power, that is, the p$_{}$ H_0 after any stage, for different alternatives $\delta = \dfrac{(\mu_1 - \mu_2)}{\sigma}$ (see the appendix for details about the calculation of the power). The power is slightly higher in case of truncation. The difference is very small when the ratio (sample size in stage one)/(sample size in stage two) is large. The reason is that the probability to stop already after the first stage depends on the sample size in stage one.

The area of the rejection region of Fisher's test that can be relocated in case of $\alpha_0 < 1$ [[4], p.1032] has, under H_0, the probability $\Pr(p_1 \geq \alpha_0$ and $p_1 p_2 \leq c_\alpha) = c_\alpha(-\ln\alpha_0)$. In case of truncation with $\tau = \alpha_0$, an area with probability $\Pr(p_1 \geq \alpha_0$ and $p_2 \leq \tilde{c}_\alpha) = \tilde{c}_\alpha (1 - \tau)$ can be relocated. Since $c_\alpha(-\ln\alpha_0) > \tilde{c}_\alpha (1 - \tau)$ for practically relevant situations (see e.g. Table II), we have $\tilde{\alpha}_1 < \alpha_1$. Hence, the probability to terminate the trial after stage one is lower in case of truncation with $\tau = \alpha_0$.

For instance, in the case of 50 observations per group and stage and δ = 0.4 (α = 0.05) the probabilities to reject H_0 after the first stage are $\Pr(p_1 \leq \alpha_1)$ = 0.496 and $\Pr(p_1 \leq \tilde{\alpha}_1)$ = 0.461, respectively. The probability to stop without rejecting H_0 is $\Pr(p_1 \geq \alpha_0)$ = 0.023 irrespective of truncation. With the fixed sample size of 100 per stage the expected total sample size is 200 - 100·Pr(stop after first

Table 2: Boundaries c_α and $\tilde{c}_\alpha$ for two to four stages

Number of stages (k)	c_α	$\tilde{c}_\alpha$ for $\tau = 0.5$
	$\alpha = 0.025$	
2	0.00380	0.00408
3	0.00072	0.00085
4	0.00015	0.00020
	$\alpha = 0.05$	
2	0.00870	0.00948
3	0.00184	0.00222
4	0.00042	0.00057

Figure 2
Power to reject H_0 in a two-stage design with $\alpha_0 = 1$. (50 observations per group and stage, combination of *t* tests, one-sided, $\alpha = 0.05$)
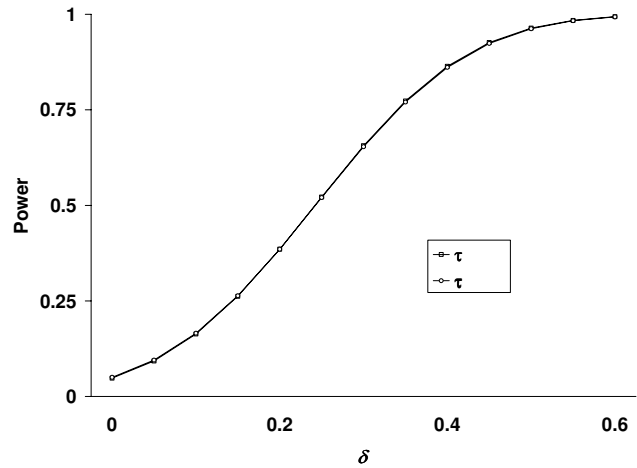
stage). This expected total sample size is 148 for $\tau = 1$, but 152 in case of truncation. Hence, the slight increase in power is connected with a larger expected total sample size.

An a priori fixed sample size for stage two is uncommon within an adaptive design. Instead, a sample size reassessment can be carried out during the interim analysis [11]. Using p_1 and the difference and variability observed in stage one, we simulated the sample size for stage two needed for an overall power of 80%. The results (not shown) indicate that, in this case, the application of the truncated product method can lead to a smaller expected total sample size.

Nevertheless, there is still a smaller probability to stop the trial after the first stage when the truncation is applied. That is a clear disadvantage in clinical development where early decisions are desirable. Therefore, despite the (small) improvement in terms of power, a truncation does not seem to be preferable within a two-stage adaptive design when $\alpha_0 < 1$.

The case $\alpha_0 = 1$
As mentioned in the introduction, $\alpha_0 = 1$ can be a suitable choice, for example when establishing a dose-response relationship. The choice $\alpha_0 = 1$ leads to the same rejection boundary c_α for the interim and the final analysis, respectively. Hence, there is $\alpha_1 = c_\alpha$ and $\tilde{\alpha}_1 = c_\alpha$. Since $c_\alpha < \tilde{c}_\alpha$, the expected total sample size is decreased due to truncation even in case of a fixed sample size for stage two. For instance, in the case of 50 observations per group and stage and $\delta = 0.4$ ($\alpha = 0.05$) the probability to reject H_0 after the first stage is $\Pr(p_1 \le \alpha_1) = 0.342$ for $\tau = 1$, but $\Pr(p_1 \le \tilde{\alpha}_1) = 0.354$ for $\tau = 0.5$. The resultant expected total sample sizes are 166 and 165, respectively. Therefore, a gain in power would be of more importance in case $\alpha_0 = 1$.

However, as demonstrated in Figure 2 there is hardly any difference in power between the choices $\tau = 0.5$ and $\tau = 1$. Nevertheless, the application of the truncated product method is preferable in the case $\alpha_0 = 1$ because there is a lower expected total sample size and a higher probability to reject H_0 already after the first stage.

The value $\tilde{c}_\alpha (= \tilde{\alpha}_1)$ increases with a decreasing truncation point τ. Hence, in order to increase the probability to reject H_0 after stage one, one may argue that a smaller value of τ is preferable. However, this is not the case because the overall power depends on the choice of τ, too. For example, consider 50 observations per group and stage and $\delta = 0.4$ ($\alpha = 0.05$) again. In this case, the overall power is 0.861 for $\tau = 1$, 0.864 for $\tau = 0.5$, but only 0.830 for $\tau = 0.2$.

We now present results for adaptive designs with three and four stages, respectively, and $\alpha_0 = 1$. Again, the behaviour of the strategies is investigated for fixed sample sizes in the separate study stages without including the option for sample size reassessment. The trial can be terminated

with the rejection of H_0 after the j-th stage if $\prod_{i=1}^{j} p_i \le c_\alpha$ in case of $\tau = 1$ or if $\prod_{i=1}^{j} p_i^{I(p_i \le \tau)} \le \tilde{c}_\alpha$ in case of truncation.

Table 3: Simulated power to reject H$_0$ and expected total sample sizes in three- and four-stage designs with α_0 = 1 (50 observations per group and stage, combination of *t* tests, one-sided, α = 0.05)

	δ =	0.1	0.2	0.3	0.4	0.5
			3 stages			
Overall power	τ = 1	0.198	0.498	0.789	0.950	0.993
	τ = 0.5	0.198	0.502	0.799	0.953	0.993
Expected total sample size	τ = 1	293.3	278.7	250.0	213.7	179.6
	τ = 0.5	292.7	276.9	247.1	209.9	176.1
			4 stages			
Overall power	τ = 1	0.230	0.590	0.883	0.984	0.999
	τ = 0.5	0.233	0.596	0.888	0.985	0.999
Expected total sample size	τ = 1	389.0	360.0	308.5	254.1	207.6
	τ = 0.5	387.6	356.2	302.5	246.8	202.3
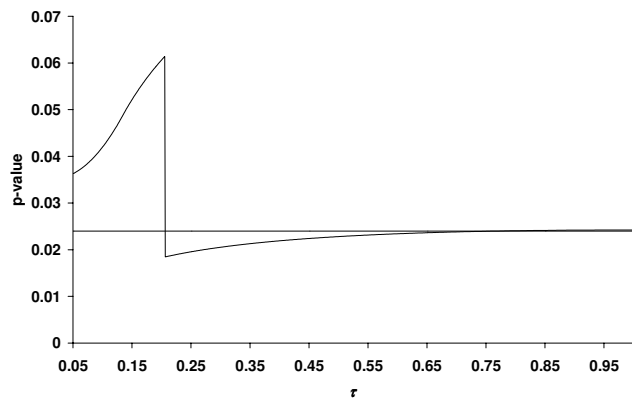
Figure 3
The overall *p*-value of the final analysis based on the combination of p_1 = 0.206 and p_2 = 0.0178 (first example) in dependence of the truncation point τ. The horizontal reference line corresponds to Fisher's product criterion.

For up to four stages, Table II displays the boundaries c_α and $\tilde{c}_\alpha$ for τ = 0.5.

The choices τ = 1 and τ = 0.5 were compared in a Monte Carlo simulation study performed using SAS version 8.2. For each configuration, 10,000 simulation runs were created. Table III shows the overall power and the expected total sample sizes. Always, the truncation is more powerful than the choice τ = 1, however, the difference in power is small. Furthermore, as in the case of k = 2, the expected total sample size is smaller when the truncated product method is applied (α_0 = 1). The decrease of the expected total sample size is more pronounced for larger values of k. Therefore, the truncation can be recom-

mended again. It reduces the expected total sample size without a loss in power.

Discussion

In this section we only consider the case α = 0.025 and α_0 = 1. The first example discussed in this section was presented by Bauer and Röhmel [6]. In a two-stage dose-response study the effect of a new drug on blood pressure was investigated. Assume that the trial would have started with two medium doses. The *p*-value for the one-sided *t* test between these two doses in the interim analysis was p_1 = 0.206. Thus, the study continued with the comparison placebo vs. a higher dose, and the second stage led to p_2 = 0.0178. The product in the final analysis was $p_1 p_2$ = 0.00367, the corresponding overall *p*-value of the non-truncated product test is 0.024. Hence, the combination test is significant even at the 0.025 level.

Figure 3 shows the overall *p*-value of the combination test in case of truncation. Note that TPM *p*-values may be calculated using a C++ code offered by Zaykin et al. [7] which is available at ftp://statgen.ncsu.edu/pub/zaykin/tpm, in addition, the method is implemented in the SAS procedure psmooth. There is no large influence of τ as long as this truncation point is larger than max(p_1,p_2). When τ is slightly smaller than max(p_1,p_2), i.e. for $\tau \rightarrow$ max(p_1,p_2) with $\tau <$ max(p_1,p_2), the *p*-value reaches a local maximum of 0.061. For $\tau <$ min(p_1,p_2) the *p*-value equals 1. Hence, a too small choice of τ is risky. Thus, the analysis of this example may be a further indication that the choice τ = 0.5 is reasonable. In fact, in this example any $\tau >$ 0.206 would have been a powerful alternative to Fisher's criterion.

The second example is a hypothetical clinical study with two stages. We consider a scenario as Bauer and Köhne [[4], p. 1038] in their example. A clinical trial investigates a new therapy for an indication in which no efficient standard therapy is available. For the first stage five individual endpoints have been selected. The first stage's sample size is 30 each in the therapy and the control group. The changes to the baseline measurements of the five endpoints were combined into a single generalized least squares (GLS) criterion according to O'Brien [12], and the first stage's *p*-value was p_1 = 0.1758. Hence, the study continued.

For the second stage the set of five endpoints may be reduced for different reasons such as observed effects and variability, burden to the patients, and costs. The test statistic for the second stage was again the corresponding GLS criterion. In this example, this led to a *p*-value of a similar magnitude as in the first stage: p_2 = 0.1517. Therefore, in the final analysis we have $p_1 p_2$ = 0.0267, and the corresponding overall *p*-value is 0.1233 when Fisher's

Table 4: The overall *p*-value of the final analysis based on the combination of $p_1 = 0.1758$ and $p_2 = 0.1517$ (second example) in dependence of the truncation point τ, $\tau = 1$ corresponds to Fisher's product criterion.

τ	*p*-value for TPM
0.1	1
0.2	0.0801
0.3	0.0964
0.4	0.1064
0.5	0.1130
0.6	0.1174
0.7	0.1203
0.8	0.1221
0.9	0.1230
1.0	0.1233

original product test is applied. The overall *p*-values for different choices of τ are displayed in Table IV. Here, the TPM gives a smaller overall *p*-value than Fisher's method for all considered values of the truncation point with the exception of $\tau = 0.1$. However, that value is smaller than $\min(p_1, p_2)$. In this example $\alpha_0 = 1$ may be appropriate because no efficient standard therapy is available, the sample size of stage 1 is relatively small, and there might be only one endpoint showing a difference between the therapy and the control group.

Conclusion

The application of the truncated product method instead of Fisher's combination test within an adaptive design hardly changes the overall power. Therefore, to decide whether or not a truncation is useful one should focus on the probability to stop early and on the expected total sample size. According to these criteria, a truncation seems to be preferable in case of $\alpha_0 = 1$, but not for $\alpha_0 < 1$.

A variety of other combination functions exists [13], for example, the inverse normal method was proposed for adaptive designs [14]. According to Rice [15] Fisher's test is "inappropriate when asking whether a set of tests, on balance, supports or refutes a common null hypothesis ... because ... Fisher's statistic is more sensitive to smaller, as compared to larger, *P*-values" [[15], p. 303–305]. In contrast, the inverse normal method is not differentially sensitive to data that support or refute a common null hypothesis. Thus, one may argue that the inverse normal method is more appropriate for an adaptive design if each stage tests the same null hypothesis. However, in the context of a dose-response study, discussed here as a motivation for $\alpha_0 = 1$, different doses may be tested in different stages, that is, the hypotheses tested change. The resultant question is whether at least one stage is significant, and a

high sensitivity to small *p*-values is desirable. Consequently, Fisher's test or TPM are appropriate. An additional advantage of these two combination methods is that an early termination with rejection of the null hypothesis is possible with $\alpha_0 = 1$ and a full level α combination test at the end.

There is also some literature related to the efficiency of adaptive designs, and to the choice of combination functions. Wassmer [16], for example, compared Fisher's product criterion with an alternative adaptive design proposed by Proschan and Hunsberger [17] based on a conditional power function. Wassmer [16] concluded that "no substantial differences between the procedures were found in terms of rejection regions, power, and expected sample sizes". One of the first to investigate optimal adaptive designs for the control of conditional power were Brannath and Bauer [18]. They constructed two-stage designs with overall and conditional power, which minimize the expected sample size for different specifications of the alternative. It transpires that there is a variety of different options to combine *P*-values and there is no consensus on the best method to use. In this paper we improve under special conditions Fisher's combination test using the truncated product method.

It is worthwhile to note that the truncation point τ must be specified a priori in the study protocol. Unless determined a priori, the truncated product method can be misused to alleviate an observed significance. A post-hoc choice based on the observed maximum of the individual *p*-values is therefore not permitted. As discussed above, $\tau = 0.5$ may be a suitable choice. A further argument for this choice is that those *p*-values are excluded from the product that indicate a difference in the unanticipated direction. Note that the truncated product does not follow a χ^2-distribution. Thus, a penalty results for the exclusion of large *p*-values. Nevertheless, this exclusion can be advantageous as demonstrated by Zaykin et al. [7] and above for the case of adaptive designs.

For the presentation of the power a one-sided significance level of $\alpha = 5\%$ was chosen in this paper. However, completely analogous results can be found in case of $\alpha = 2.5\%$. Regarding the choice of α for one-sided tests it is referred to Neuhäuser [19].

Appendix

The power of a two-stage test according to Bauer and Köhne [4], that is, a combination with $\tau = 1$, is given e.g. by Wassmer [[20], p. 833].

In case of truncation with $\tau = \alpha_0 > \alpha$ the power is

$$1 - \Pr(p_1 \geq \alpha_0) - \Pr(\tilde{\alpha}_1 < p_1 < \alpha_0 \ and \ \tilde{c}_\alpha / p_1 < p_2 \leq \tau) - \Pr(\tilde{\alpha}_1 < p_1 < \alpha_0 \ and \ p_2 > \tau)$$
$$= 1 - \int_{\alpha_0}^{1} f_\delta(p_1) dp_1 - \int_{\tilde{\alpha}_1}^{\alpha_0} \int_{\tilde{c}_\alpha / p_1}^{\tau} f_\delta(p_1, p_2) dp_2 dp_1 - \int_{\tilde{\alpha}_1}^{\alpha_0} \int_{\tau}^{1} f_\delta(p_1, p_2) dp_2 dp_1,$$

where f_δ denotes the respective density under the alternative δ [20]. In case of truncation with $\tau > \alpha$, but $\alpha_0 = 1$, the power is

$$1 - \Pr(\tilde{c}_\alpha < p_1 \leq \tau \ and \ \tilde{c}_\alpha / p_1 < p_2 \leq \tau) - \Pr(\tilde{c}_\alpha < p_1 \leq \tau \ and \ p_2 > \tau)$$
$$- \Pr(p_1 > \tau \ and \ \tilde{c}_\alpha < p_2 \leq \tau) - \Pr(p_1 > \tau \ and \ p_2 > \tau)$$
$$= 1 - \int_{\tilde{c}_\alpha}^{\tau} \int_{\tilde{c}_\alpha / p_1}^{\tau} f_\delta(p_1, p_2) dp_2 dp_1 - \int_{\tilde{c}_\alpha}^{\tau} \int_{\tau}^{1} f_\delta(p_1, p_2) dp_2 dp_1$$
$$- \int_{\tau}^{1} \int_{\tilde{c}_\alpha}^{\tau} f_\delta(p_1, p_2) dp_2 dp_1 - \int_{\tau}^{1} \int_{\tau}^{1} f_\delta(p_1, p_2) dp_2 dp_1.$$

Wassmer [20] presented a SAS/IML program to calculate the power for the two-stage test without truncation. Modifications of this program were used to calculate the different powers given above.

Abbreviations
TPM – truncated product method

Authors' contributions
MN performed most of the statistical analyses and drafted the manuscript. FB participated in the statistical analyses and helped to draft the manuscript. Both authors read and approved the final manuscript.

References
1. Palmer CR: **Ethics, data-dependent designs, and the strategy of clinical trials: time to start learning-as-we-go?** *Statistical Methods in Medical Research* 2002, **11**:381-402.
2. Gauch HG: **Scientific method in practice.** Cambridge University Press: Cambridge; 2002.
3. Bauer P, Brannath W, Posch M: **Flexible two-stage designs: an overview.** *Methods of Information in Medicine* 2001, **40**:117-121.
4. Bauer P, Köhne K: **Evaluation of experiments with adaptive interim analyses.** *Biometrics* 1994, **50**:1029-1041. Correction in *Biometrics* 52:380
5. Hedges LV, Olkin I: **Statistical methods for meta-analysis.** Academic Press: Orlando; 1985.
6. Bauer P, Röhmel J: **An adaptive method for establishing a dose-response relationship.** *Statistics in Medicine* 1995, **14**:1595-1607.
7. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS: **Truncated product method for combining P-values.** *Genetic Epidemiology* 2002, **22**:170-185.
8. Neuhäuser M: **Tests for genetic differentiation.** *Biometrical Journal* 2003, **45**:974-984.
9. Dudbridge F, Koeleman BPC: **Rank truncated product of P-values, with application to genomewide association scans.** *Genetic Epidemiology* 2003, **25**:360-366.
10. Brannath W, Posch M, Bauer P: **Recursive combination tests.** *Journal of the American Statistical Association* 2002, **97**:236-244.
11. Friede T, Kieser M: **A comparison of methods for adaptive sample size adjustment.** *Statistics in Medicine* 2001, **20**:3861-3873.
12. O'Brien PC: **Procedures for comparing samples with multiple endpoints.** *Biometrics* 1984, **40**:1079-1087.
13. Loughin TM: **A systematic comparison of methods for combining p-values from independent test.** *Computational Statistics and Data Analysis* 2004, **47**:467-485.
14. Lehmacher W, Wassmer G: **Adaptive sample size calculations in group sequential trials.** *Biometrics* 1999, **55**:1286-1290.
15. Rice WR: **A consensus combined P-value test and the family-wide significance of component tests.** *Biometrics* 1990, **46**:303-308.
16. Wassmer G: **A comparison of two methods for adaptive interim analyses in clinical trials.** *Biometrics* 1998, **54**:696-705.
17. Proschan MA, Hunsberger SA: **Designed extension of studies based on conditional power.** *Biometrics* 1995, **51**:1315-1324.
18. Brannath W, Bauer P: **Optimal conditional error functions for the control of conditional power.** *Biometrics* 2004, **60**:715-723.
19. Neuhäuser M: **The choice of α for one-sided tests.** *Drug Information Journal* 2004, **38**:57-60.
20. Wassmer G: **A technical note on the power determination for Fisher's combination test.** *Biometrical Journal* 1997, **39**:831-838.

Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/5/30/prepub