Author(s): Frank Miller, Tim Friede, Meinhard Kieser

Article Title: Blinded assessment of treatment effects utilizing information about the randomization block length
Year of publication: 2009
Link to published version: http://dx.doi.org/ 10.1002/sim.3576
Publisher statement: The definitive version is available at www3.interscience.wiley.com

Blinded assessment of treatment effects utilising information about the randomisation block length

Frank Miller[1], Tim Friede[2] and Meinhard Kieser[3,*]

Astra Zeneca, Södertälje, Sweden[1]

Warwick Medical School, The University of Warwick, UK[2]

Institute of Medical Biometry and Informatics, University of Heidelberg, Germany[3]

[*]Correspondence to: Meinhard Kieser, Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg.

E-mail: meinhard.kieser@imbi.uni-heidelberg.de

Short title: Blinded assessment of treatment effects

SUMMARY

It is essential for the integrity of double-blind clinical trials that during the study course the individual treatment allocations of the patients as well as the treatment effect remain unknown to any involved person. Recently, methods have been proposed for which it was claimed that they would allow reliable estimation of the treatment effect based on blinded data by using information about the block length of the randomisation procedure. If this would hold true, it would be difficult to preserve blindness without taking further measures. The suggested procedures apply to continuous data and their characteristics were illustrated by applying them to a single simulated data set per scenario considered. We investigate the properties of these methods more thoroughly by repeated simulations per scenario. Furthermore, a method for blinded treatment effect estimation in case of binary data is proposed, and blinded tests for treatment group differences are developed both for continuous and binary data. We report results of comprehensive simulation studies that investigate the features of these procedures. It is shown that for sample sizes and treatment effects which are typical in clinical trials, no reliable inference can be made on the treatment group difference which is due to the bias and imprecision of the blinded estimates.

KEY WORDS: clinical trial; blindness; treatment effect; block randomisation; trial integrity.

1. INTRODUCTION

One of the most important issues in the design of clinical trials is the implementation of procedures that eliminate or minimise bias. Blinding and randomisation are generally considered to be the most important techniques in this context [1]. While randomisation eliminates bias in treatment assignment, blinding minimises the risk of trial personnel or patients being aware of the assigned intervention and thus from being influenced by this. Knowledge of treatment group assignment may, beside others, affect the response of the participants, influence the use of ancillary interventions of the investigators, or lead to a differential judgment of the outcome by the assessors [2]. But even if individual treatment assignments are unknown, knowledge of the size of the treatment effect during an ongoing trial may influence the attitude of anyone involved in the trial and may thus lead to the occurrence of bias by mechanisms as such described above. For this reason, regulatory guidelines strongly demand that the results of interim analyses are not disseminated to the personnel and patients involved in a trial [1, 3]. It is therefore standard that any comparison between treatment groups that is performed mid-course of a trial is assigned to an independent committee which maintains the results strictly confidentially [4]. The only information that is disseminated beyond the small circle of committee members is whether the trial was stopped or continued. As a further measure, information on who is assigned to which treatment group is not included in the trial database while the study is ongoing but is kept separately with special access rights.

Usually, general information on the implemented randomisation procedure is given in the protocol, such as the method used for random sequence generation or prognostic

variables that are taken into account in stratified randomisation. Quite frequently, block randomisation is applied to minimise the risk of obtaining an undesirable sample size imbalance between the treatment groups. In order to prevent from the possibility of predicting the assignment of future patients from guessed or known assignments of patients already allocated, it is recommended not to specify the block length in the protocol [5]. However, trial statisticians are generally aware of the block lengths and - as they have access to the trial database - they also know the order in which the patients were recruited. Therefore, they are aware of which patients are in which block. It is evident that for extreme outcome scenarios this knowledge can be used to derive information about the treatment effect even if the individual treatment group allocation remains blinded. For example, in the case of a binary endpoint "event yes / no" where in every block either the event occurs for all patients or does not occur for any patient, it is clear that there is no difference in treatment effect between the groups. The question arises, whether reliable conclusions about the treatment effect can be drawn from the blinded data also in more realistic situations. The message given in a recent paper by van der Meulen [6] raises hopes or fears - depending on the perspective - that this would be possible. In the abstract of [6] it was claimed that when randomisation is done using permuted blocks "statistical inference of the treatment effects can be conducted" before unblinding "yielding consistent and rather precise estimates" [6, *p.* 479]. In the light of the above-mentioned fact that the integrity of the trial is questioned if information about the extent of the treatment effect becomes known and if actions are taken based on this information during the course of a trial, this would create serious problems.

Other methods that allow a blinded estimation of the treatment effect have been proposed in the literature before. In the framework of blinded sample size re-estimation, Gould and Shih [7] suggested a method to estimate the within-group variance of normally distributed data. This procedure is based on an EM algorithm and does not make use of information about the randomisation. It provides not only an estimate of the within-group variance but also of the treatment effect. It was shown that this approach exhibits some major deficits and is not appropriate for sample size re-estimation based on the blinded variance estimate [8, 9]. Waksmann [10] corrected the Gould-Shih procedure such that now the resulting estimates converge to the ML estimates. He investigated the characteristics of the resolved method in a thorough simulation study and found that for sample sizes and treatment effects that usually occur in clinical trials, the estimates of the treatment group difference are not accurate enough to be of practical value. The same conclusion was drawn by Xing and Ganju [11] for the procedure they derived by making use of the knowledge of the randomisation block length. Their primary focus was the construction of a blinded variance estimator for the purpose of blinded sample size re-estimation in case of continuous data. They briefly mention that a blinded estimator of the treatment effect can be obtained as a by-product, based on the difference between the pooled one-sample variance estimator for all data and their blinded within-group variance estimator (which is a multiple of the variance of the block means). Based on heuristic arguments and simulation results, Xing and Ganju concluded that “there is no risk of unblinding the trial” because “the variation in the estimate of Δ^2 is large enough to be practically useless”. In contrast to the other references mentioned above, van der Meulen [6] focuses on estimation of the treatment effect using estimators based on all available effect information in the blinded data.

Interestingly, he derived blinded moment and ML estimators of the treatment group difference for continuous data that use the knowledge of the block length applied in randomisation. His work on the estimators is therefore a good basis to investigate the amount of effect-information contained in blinded data. Besides, van der Meulen's [6] investigations suggest that the blinded estimates of the variance he derived with the ML method are more precise than the blinded treatment effect estimates. Recently, Ganju and Xing [12] discussed specifically the blinded treatment effect estimation of van der Meulen and concluded that "the blinded method is refined enough for blinded variance estimation but blunt enough for inferring efficacy".

The aim of our paper is to assess the risk of unblinding the treatment effect from blinded inference. This goal is examined (i) by investigating the procedure proposed by van der Meulen for continuous data more thoroughly by simulation studies; (ii) by developing alternative approaches for a blinded assessment of treatment effects by utilising knowledge of the randomisation block length in case of continuous and binary outcomes; and (iii) by investigating comprehensively the characteristics of the new methods. Thereby, previous work on blinded estimation of treatment group differences is extended in various ways. Section 2.1 takes a closer look to the characteristics of the procedure proposed by van der Meulen for blinded estimation of treatment effects in case of continuous data by presenting results of repeated simulations. In Section 2.2, a blinded test for difference of means is derived and its properties are investigated. An example illustrates the findings in Section 3. Related methods for binary data did not exist up to now. Section 4 presents blinded estimators of and a test for treatment group difference in case of binary outcomes and shows their characteristics. We summarise the findings in Section 5 and discuss them in the context of the literature.

2. BLINDED ASSESSMENT OF TREATMENT EFFECT FOR CONTINUOUS DATA

2.1 Blinded estimation of treatment effect

Here and in the remainder of the paper the treatment effect, i.e., the difference of the expected values in the two intervention groups, is denoted by Δ. Hence, in case of continuous data $\Delta = \mu_1 - \mu_2$ with group expectations μ_1, μ_2. We start with a block length of $l = 2$ and adopt the notation used in Reference [6]. Let Y_{i1}, Y_{i2} denote the 1st and 2nd observation in block i, $i = 1, ..., k$, where $\text{var}(Y_{ij}) = \sigma^2$, $i = 1, ..., k$, $j = 1, 2$. Further, $Z_{1i} = Y_{i1} - Y_{i2}$ is the within-block difference in block i and $Z_{2i} = Y_{i1} + Y_{i2}$ the within-block sum in block i, $i = 1, ..., k$; $Z_{2i}/2$ is therefore the mean in block i. The between-block variation (i.e., the variance of block means) can thus be expressed as $S_{BB}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(Z_{2i}/2 - \bar{Z}_{2\bullet}/2\right)^2$, where $\bar{Z}_{2\bullet}$ denotes the mean of Z_{2i}, $i = 1, ..., k$. Obviously, the expectation of S_{BB}^2 is given by $E(S_{BB}^2) = \sigma^2/2$. The within-block variation is given by $S_{WB}^2 = \frac{1}{k}\sum_{i=1}^{k} Z_{1i}^2/2$ and has expectation $E(S_{WB}^2) = \sigma^2 + \Delta^2/2$. Consequently, an unbiased estimator of Δ^2 is given by $2S_{WB}^2 - 4S_{BB}^2$, and the absolute value $|\Delta|$ of Δ can be estimated by $\sqrt{\max\left(2S_{WB}^2 - 4S_{BB}^2, 0\right)}$. This is the moment estimator $\widehat{|\Delta|}$ of $|\Delta|$ presented in [6]. No assumptions are to be made regarding the distribution of the observations when deriving this estimator. Alternatively, the

maximum likelihood (ML) estimator $\left|\hat{\Delta}\right|$ of $|\Delta|$ can be derived under the assumption of normally distributed observations Y_{ij}. The log likelihood is then given by

$$l(\mu_1, \mu_2, \sigma^2) = \sum_{i=1}^{k} \log\left(\frac{1}{2} \varphi(y_{i1}; \mu_1, \sigma^2)\, \varphi(y_{i2}; \mu_2, \sigma^2) + \frac{1}{2} \varphi(y_{i1}; \mu_2, \sigma^2)\, \varphi(y_{i2}; \mu_1, \sigma^2) \right),$$

where $\varphi(y; \mu, \sigma^2)$ is the density of a normal distribution with mean μ and variance σ^2 [6]. The ML estimates can be obtained by maximising the log likelihood using the Newton-Raphson algorithm, as for example implemented in the SAS/IML function nlpnra. The moment estimators for $|\Delta|$, $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$ and σ^2 can be used to initialise the algorithm. Analogously, the moment estimator and the ML estimator of $|\Delta|$ can be derived for a block length of $l = 4$ (see [6, *p.* 483*ff*] and Appendix A for a correction of the moment estimator given in [6]).

Figures 1 and 2 show the simulated bias $\left|\hat{\Delta}\right| - |\Delta|$ of the moment estimator and the ML estimator $\left|\hat{\Delta}\right|$ for $|\Delta|$ for block lengths of $l = 2$ and $l = 4$, respectively. Considered were (standardised) treatment effects $\Delta = 0, 0.25, 0.5$ and 1.0 ($\sigma = 1$) and total sample sizes 20, 40, …, 100, 200, …, 800, 1,000 which were balanced between the two groups. It should be noted that total sample sizes of 506, 128 and 34 are required to achieve a power of 80 per cent at one-sided level $\alpha = 0.025$ with the unblinded *t*-test for standardised differences of $\Delta/\sigma = 0.25, 0.5$ and 1.0, respectively. For each situation, 1,000 simulations were performed.

– Please insert Figures 1 and 2 about here. –

Both the moment estimator and the ML estimator overestimate the absolute value of the treatment effect for $\Delta = 0$ and $\Delta = 0.25$ for all sample sizes, while underestimation

occurs for $\Delta = 1$ (with the only exception of $l = 4$ and total sample sizes smaller than 40). As could be expected, the extent of bias is for the same sample sizes generally higher for block size $l = 4$ than for block size 2. For $\Delta = 0.5$, a positive bias is observed for small sample sizes which changes to a negative bias when the sample size is increased. The shift from overestimation to underestimation occurs for higher sample sizes for block size $l = 4$ as compared to $l = 2$. In all considered situations, the mean estimates resulting from the moment estimator are slightly smaller than those for the ML estimator, but the difference is negligible for practical purposes. For $\Delta > 0$, the bias approaches zero only for sample sizes which would have been used if the relevant effect assumed at the planning stage was much smaller than the true treatment effect Δ. As a reviewer noted, this result is also reflected in Table 1 of van der Meulen's paper [6]: For the data set he used, the treatment effect is estimated precisely from blinded data only when the trial has extremely high power (> 99%); see also [12].

The simulations also show that the standard deviations of the two blinded estimators are very similar as well. The variability is higher for block size $l = 4$ as compared to $l = 2$. In comparison to unblinded estimation of the treatment effect, the standard deviation of the blinded estimates is considerably higher by a factor of up to 3 for the range of investigated parameter constellations (results not shown).

What do these results mean in terms of information about the treatment effect that can be inferred from blinded data of an ongoing trial? For example, if a clinical trial is powered for $1-\beta = 0.80$ at a clinically relevant treatment effect of $\Delta^* = 0.25$, this requires a total sample size of $n^* = 506$ (balanced groups, *t*-test at one-sided level $\alpha = 0.025$). If block randomisation with a block size 2 is applied and the treatment effect is estimated by application of the blinded moment or ML estimator based on the

data of 500 patients utilising the information $l = 2$, then Figure 1 shows that the expected estimated treatment group difference amounts to about 0.27 if the true treatment effect is in fact $\Delta = 0.25$. However, if there is actually no treatment group difference at all ($\Delta = 0$), the expected blinded estimate is about 0.21. It is therefore evident that in this situation one is not able to distinguish whether a blinded estimate of, say, 0.25 arises from overestimation of a non-existing treatment group difference or from a treatment effect $\Delta > 0$.

– Please insert Figure 3 about here. –

Figure 3 considers this aspect in more detail. The box-whisker plots show for block lengths $l = 2$ and $l = 4$ the range of the central 90 per cent of values arising from 1,000 simulated blinded treatment effect estimates (moment estimator grey, ML estimator white). It can be seen that there is a heavy overlap of the distributions of the blinded estimators that originate from different values of Δ. Furthermore, with the exception of the constellations of a total sample size of 500 and $l = 2, \Delta = 0.75, 1$ and $l = 4, \Delta = 1$, the box-whisker plot includes the value zero. Hence, for adequately powered studies a blinded estimate of zero gives no indication at all about the actual underlying treatment effect. The other way round, for $\Delta = 0$ and $l = 2$ ($l = 4$) the upper whisker reaches the value 1.25, 1.0, and 0.65 (1.5, 1.25, and 0.85) and total sample sizes of 40, 100, and 500, respectively. This means that only huge treatment effect estimates would give cause to the suspicion of a non-zero treatment effect. Figure 3 shows that such estimates occur with considerable probability only for at least as extreme true treatment effects. As a consequence, for treatment group differences and sample sizes which are common in clinical trials, a differentiation between various extents of Δ based on the blinded moment or ML estimator is not possible. This uncertainty about the actual amount of Δ

is a consequence of the substantial positive bias of the blinded estimators for the situation $\Delta = 0$ combined with a high variability of the estimates.

In order to study the sensitivity of our findings to deviations from the normal distribution we conducted simulations with non-normal distributions including distributions with heavy tails (*t*-distributions with small number of degrees of freedom) and skewed distributions (e.g. log normal distribution). We found that the results for the moment estimator are robust against such deviations. However, the ML estimator was sensitive to deviations from normality as expected with at times severe downwards bias in estimation of the treatment effect Δ. The general observation for non-normal distributions was that the performance of the blinded estimators is not better as seen for normal data. A differentiation between various extents of Δ based on the blinded moment or ML estimator is even more difficult.

2.2 Blinded tests for treatment effect

From the above mentioned formulae for the within-block variation and the between-block variation, a blinded F-test can be derived. As the F-test and the moment estimator are closely related, it cannot be expected that this test will be more revealing than the findings presented in Section 2.1, but is an alternative way to look at this problem. For block length $l = 2$ the test statistics is given by $F_{blind,\, l=2} = \dfrac{S_{WB}^2}{S_{BB}^2} = \dfrac{2 \cdot \dfrac{1}{k}\sum_{i=1}^{k} Z_{1,i}^2}{\dfrac{1}{k-1}\sum_{i=1}^{k}\left(Z_{2,i} - \bar{Z}_{2\bullet}\right)^2}$ and follows the non-central F-distribution $F_{k,k-1}\left(\dfrac{k\Delta^2}{2\sigma^2}\right)$ with $(k, k-1)$ degrees of freedom and non-centrality parameter $\dfrac{k\Delta^2}{2\sigma^2}$. Analogously, a test statistics $F_{blind,\ell=4}$ for a

blinded F-test can be derived for block length four, which is given in Appendix A. $F_{blind,\,l=4}$ can again be interpreted as the ratio of the within-block variation to the between-block variation and follows the non-central F-distribution $F_{3k,k-1}\left(\frac{k\Delta^2}{\sigma^2}\right)$. Due to the known distribution of the test statistics, the power properties of the blinded F-tests can be evaluated analytically.

– Please insert Figure 4 about here. –

Figure 4 shows the relationship between the treatment effect Δ^* used for sample size calculation and the ratio r of the true treatment effect Δ to the effect Δ^* that is required to achieve a power of $1-\beta = 0.5$ and 0.8, respectively, for the blinded F-test with significance level $\alpha = 0.05$. The sample size is chosen such that the unblinded two-sided *t*-test at $\alpha = 0.05$ has a power of $1-\beta^* = 0.8$ at the treatment effect $\Delta^*\ (\sigma{=}1)$. This means that r is determined such that $1-\text{Probf}_{(l-1)k,\,k-1,\,nc}(f_{(l-1)k,\,k-1,\,1-\alpha}) = 1-\beta$, where k is the nearest integer to $(4/l)(z_{1-\alpha/2}+z_{1-\beta^*})^2(\sigma/\Delta^*)^2$ and $nc = r^2\cdot(z_{1-\alpha/2}+z_{1-\beta^*})^2$. $\text{Probf}_{df_1,\,df_2,\,\vartheta}$ denotes the distribution function of the non-central F distribution with (df_1, df_2) degrees of freedom and non-centrality parameter ϑ, $f_{df_1,\,df_2,\,\gamma}$ is the γ quantile of the central F distribution with the same degrees of freedom, and z_γ is the γ quantile of the standard normal distribution.

For $l = 2$ $(l = 4)$ and a clinically relevant treatment effect $\Delta^* = 0.25$ assumed for sample size calculation, a 2.7 (3.7) times larger true treatment effect Δ is necessary to achieve a 50 per cent power for the blinded F-test. A factor of about 3.4 (4.6) between the actual treatment effect and the effect used for sample size calculation is required to

have an adequate power of 80 per cent. From another perspective, if $\Delta = \Delta^* = 0.25$ the factor between the chosen and the required sample size amounts to about 46 (136) for a desired power of 50 per cent of the blinded F-test and to about 104 (309) for 80 per cent. For a clinically relevant treatment effect of $\Delta^* = 0.5$ the situation becomes a little bit less extreme. However, still about double (2.8 times) this treatment effect is necessary to achieve a power of 50 per cent for the blinded F-test, and the true effect needs to be about 2.6 (3.7) times as large as this effect to get a power of 80 per cent for the blinded test. In other words, for $\Delta = \Delta^* = 0.5$ the total sample size required to achieve 80 per cent power for the blinded F-test amounts to about $n = 3,560$ ($n = 10,340$) while the chosen sample size is $n^* = 128$.

In practice, one would usually try to learn from the blinded data while the study is still ongoing. Hence, the blinded test would be applied before the data of all n^* patients are available. As a consequence, the required factors by which the true effect has to exceed the one used for sample size calculation to achieve a power of 50 or 80 per cent for the blinded test are even higher as those reported above.

These considerations show that the blinded F-test has reasonable power only when the true treatment effect is several times larger than the clinically relevant effect assumed in the sample size calculation of the study, i.e. if the study is an overpowered study from a retrospective point of view. Though there is a possibility for this to happen, this is a very rare event in clinical practice.

3. EXAMPLE

In order to illustrate the results, we consider placebo-controlled clinical studies in the acute treatment of major depression. The change in total score of the Hamilton Rating Scale for Depression (HAM-D, 17-item version) [13] between baseline and end of therapy is often used as the primary endpoint in such studies, and a difference of $\Delta^* = 3$ points can be considered as clinically relevant [14]. If the standard deviation σ is assumed to lie within the range of 5 to 8 as usually observed for this outcome (see, for example, Reference [15]). This corresponds to total sample sizes of 90 to 230 required to achieve a power of 80 per cent for the treatment group difference $\Delta^* = 3$ (two-sided *t*-test, $\alpha = 0.05$). These sample sizes match well the number of patients typically included in studies investigating the efficacy of the treatment of acute major depression (see, for example, the systematic reviews [15, 16]).

The true treatment group difference needed to achieve a power of $1 - \beta = 0.80$ for the blinded F-test would amount to 7.4 ($\sigma = 5$) or 8.8 ($\sigma = 8$) HAM-D points, respectively, for block length two and 10.5 ($\sigma = 5$) or 11.9 ($\sigma = 8$), respectively, for block length four. For all existing antidepressants, such huge overall treatment effects are out of reach. The other way round, if the true treatment effect would amount to $\Delta = 3$ HAM-D points, the sample size required to attain a power of $1 - \beta = 0.80$ for the blinded F-test would be 1,800 ($\sigma = 5$) or 10,700 ($\sigma = 8$), respectively, for block length two, and 5,200 ($\sigma = 5$) or 31,500 ($\sigma = 8$), respectively, for block length four. However, there is no reason to perform such large-scale placebo-controlled studies investigating the efficacy of short-term treatment of major depression.

In a double-blind, placebo-controlled trial Kalb *et al.* [17] investigate the efficacy of St. John's wort in patients with mild to moderate depression. The primary variable was the change in HAM-D total score between baseline and end of the six weeks acute treatment phase. Due to the uncertainty about the variability of the primary endpoint and the expected treatment effect, the study was planned with an interim analysis that should be performed when roughly half of the projected sample size of 130 was achieved. A block length of two was used for randomisation. For those 72 patients that were included in the interim analysis, 35 blocks were complete such that a blinded assessment could be based on 70 patients. Applying the blinded ML- and moment estimator to this data set leads to estimated treatment effects of 3.2 and 2.4 points, respectively. Employing the one-sample estimator of the standard deviation of 6.3 results in an estimated standardised treatment effect of 0.4 to 0.5. In view of the results shown in Figure 3, this estimate is compatible with true treatment effects from 0.0 to 1.0. Application of the blinded F-test leads to a test statistic of $F = 1.08$, which corresponds to a *p*-value of $p = 0.42$. Hence, blinded estimation and testing give no hint for a significant difference between the two treatment groups. However, the unblinded estimate of the difference between the group means amounts for the same data set to 5.1 and the two-sample *t*-test results in a two-sided *p*-value of $p = 0.0004$. As a consequence, the study could be stopped after the planned interim analysis with a significant superiority of St. John's wort to placebo (pre-defined critical boundary for early rejection $\alpha_1 = 0.0207$), whereas the blinded methods would not have given any indication to complete the trial early. Therefore, this real study data give another counter-example to the claim that blinded inference "takes away the need of conducting interim analyses" [6].

4. BLINDED ASSESSMENT OF TREATMENT EFFECT FOR BINARY DATA

4.1 Blinded estimation of treatment effect

In case of binary data, we consider the ML estimator and the moment estimator for blinded estimation of the treatment effect. We will see that this approach leads to a simple expression for the estimator. In order to avoid repetition we restrict the presentation here to block length $l = 2$ as in this situation most information can be drawn from the blinded data thus marking the best case scenario for the performance of blinded estimation methods.

The information included in blinded binary data is the number of events that occurred in each block and the order they occurred. In the case of block size $l = 2$, the numbers a_0, a_{10}, a_{01}, a_2 of blocks with 0 events, with one event that occurs for the first patient, with one event that occurs for the second patient and with two events, respectively, can be observed. If a_{10} or a_{01} is "large", one would become suspicious that there may be a difference between the treatment groups. In contrast, for every block with no event or two events one knows for sure that there is no difference between the number of events between the treatment groups.

The blinded ML estimator for the difference in event rates between the two treatments can be derived analytically. If p_i denotes the true event rate in treatment group $i = 1, 2$, then the probabilities of observing a block with 0, 2, or 1 event are $(1-p_1)(1-p_2)$, $p_1 p_2$, and $1-(1-p_1)(1-p_2)-p_1 p_2$, respectively. The probabilities to

observe a block with one event that occurs in the first patient or a block with one event that occurs in the second patient are equal. Hence, the likelihood function is given by $\left[(1-p_1)(1-p_2)\right]^{a_0}\left[p_1p_2\right]^{a_2}\left[(1-(1-p_1)(1-p_2)-p_1p_2)/2\right]^{a_{10}+a_{01}}$. We can restrict on using the total number of blocks with one event $a_1 = a_{10} + a_{01}$ instead of (a_{10}, a_{01}), since the likelihood function shows that (a_0, a_1, a_2) is a sufficient statistic. If the expression above is re-parameterised with $\pi = (p_1 + p_2)/2$, $q = (p_1 - p_2)/2$, and $d = q^2$, it can be seen after some calculation that the log likelihood is maximised for $\hat{\pi} = (a_1 + 2a_2)/(2k)$ (which is the observed overall event rate) and for $\hat{d} = (a_1^2 - 4a_0a_2)/(2k)^2$ if $(a_1^2 - 4a_0a_2)/(2k)^2 \geq 0$, and $\hat{d} = 0$ otherwise. Hence, the blinded ML estimator for $|\Delta| = |p_1 - p_2| = 2\sqrt{d}$ in case of block length $l = 2$ is given by $\widehat{|\Delta|} = \frac{1}{k}\sqrt{\max(a_1^2 - 4a_0a_2, 0)}$.

Another possibility is to apply the moment estimator presented in Section 2.1 directly to binary data. In case of binary data and block length $l = 2$ we find $S_{WB}^2 = a_1/(2k)$ and $S_{BB}^2 = (a_0\hat{\pi}^2 + a_1(\hat{\pi} - 0.5)^2 + a_2(1-\hat{\pi})^2)/(k-1)$. With these expressions, the moment estimator $\sqrt{\max\left(2S_{WB}^2 - 4S_{BB}^2, 0\right)}$ for $|\Delta| = |p_1 - p_2|$ can be calculated.

– Please insert Figure 5 about here. –

Figure 5 shows the box-whisker plots covering the central 90 per cent of the empirical distribution of the blinded estimates for $l = 2$ (moment estimator grey, ML estimator white) arising from 1,000 simulations for each situation. Overall rates $\pi = (p_1 + p_2)/2$ of 0.2, 0.5, treatment effects $\Delta = p_1 - p_2$ of 0, 0.05, …, 0.3 and balanced total sample sizes of 100, 500 and 1000 were considered. Qualitatively, the results are very similar to

those obtained for continuous data. Again the box-whisker plots show an extensive overlap for different values of π and Δ. Moreover, they cover the value zero for almost all considered situations with the exception of some combinations of a total sample size of 500 or 1,000 and treatment effects of 0.25 or 0.3. However, in these situation the total sample size required to achieve a power of 80 per cent with the unblinded chi-square test is extremely much lower and ranges from 54 ($\pi = 0.2, \Delta = 0.3$) to 86 ($\pi = 0.5, \Delta = 0.3$). Hence, for reasonably powered studies a blinded estimate of zero for the treatment group difference is both in accordance with the assumption $\Delta = 0$ and $\Delta > 0$. As for continuous data, the upper whiskers reach very large treatment effects for $\Delta = 0$, for example values of 0.50, 0.32, and 0.27 for an overall rate of $\pi = 0.5$ and total sample sizes of 100, 500, and 1,000, respectively. Therefore, only if even more extreme values are obtained from blinded estimation of the group difference it is worth to speculate whether the true treatment effect might be different from zero. However, such large estimated effects occur in turn only for very large true differences with noticeable probability. In summary, it can be concluded that for treatment effects and sample sizes which are usually met in clinical trials different values of Δ cannot be distinguished by blinded moment or ML estimation.

4.2 Blinded tests for treatment effect

In Section 2.2 we introduced the F-test $F_{blind,\, l=2} = \frac{S_{WB}^2}{S_{BB}^2}$ for continuous data that we could apply for binary data as well. In case of binary data and block length $l = 2$ we can employ the expressions for S_{WB}^2 and S_{BB}^2 given in the previous section. Unlike with normal data, however, the reference distribution is not an F-distribution. One way of

obtaining a reference distribution is by considering all permutations of '1' (event) and '0' (no event) keeping the total number of events fixed. It can be shown that the between-block variation S_{BB}^2 is invariant to permutations. Therefore, the test statistics reduces to a_1 after elimination of invariants, with larger values of a_1 being stronger evidence against the null hypothesis of no effect. A *p*-value for the test is given by

$$\sum_{b_2=\max\{0,e-k\}}^{floor(e/2)} \binom{k}{b_2} \binom{k-b_2}{b_1} \; 2^{b_1} \;\; \mathrm{I}\{b_1 \geq a_1\},$$

where $floor(x)$ is the largest integer that is less than or equal to *x*, $e = 2a_2 + a_1$, $b_1 = b_1(b_2) = e - 2b_2$, and $\mathrm{I}\{\cdot\}$ denotes the indicator function.

– Please insert Figure 6 about here. –

Similar to Figure 4 in Section 2.2, Figure 6 shows the relationship between the risk difference Δ^* used for sample size calculation and the ratio *r* of the actual risk difference Δ to the effect Δ^* that is required to achieve a power of $1-\beta = 0.5$ and 0.8, respectively, for the blinded permutation test given above with an overall event rate of $\pi = 0.2$ and 0.5. Using the approximation formula given in [18] the sample size is calculated such that the unblinded two-sided Fisher's exact test has a power of $1-\beta^* = 0.8$ for level $\alpha = 0.05$, overall event rate $\pi = 0.2, 0.5$, and difference in event rates Δ^*. The ratios *r* were determined by means of line searches over Δ with step size 0.005. Hereby the power values of the blinded test were simulated using 50,000 replications.

As can be seen from Figure 6 the risk differences need to be 2 to 4 times larger than anticipated in order to obtain a power in the range of 0.5 to 0.8 for the considered

combinations of rate differences and overall event rates. This is in keeping with similar effects observed in Figure 4.

Alternative test statistics could be used for the permutation test. For instance, the MLE of the treatment difference given above or the likelihood ratio test (LRT) statistic would make good choices. Both statistics lead to tests equivalent to the one used above.

5. DISCUSSION

We saw that it is possible to construct both for continuous and for binary data estimators and tests for the treatment effect that are based on blinded data by utilising information about the randomisation procedure. Although we demonstrated this only for block sizes of 2 and 4, this applies for other block sizes. Not surprisingly, the information that can be gained from blinded data is highest for the smallest possible block size of 2. But even in this situation, a reasonable precision for the treatment effect estimate can hardly be obtained in realistic scenarios. Furthermore, the blinded tests have sufficient power only if the study is severely overpowered: For a block size of 2, the actual treatment effect has to exceed the clinically relevant effect by a factor of at least 2. A situation where this may happen is in case of multiple co-primary endpoints with differing treatment effects. When the study sample size is chosen such that adequate power is assured for the outcome with the smallest minimally clinically relevant effect size, an excessive power may result for the endpoint with the largest treatment effect. If a single endpoint is used, the actual treatment effect will usually be not much larger than the clinically relevant treatment effect assumed in the sample size calculation of the study. In

indications where treatments are already approved and on the market, the clinically relevant effect is generally near the observed effect of the standard. Hence, in case of indications with existing treatments a reasonable power of the blinded test is only achieved if a new treatment shows a “jump” in efficacy, which may happen only in exceptional cases in drug development. Consequently, the only realistic situations where reliable information about the treatment effect can be gained from blinded data in case of an overpowered study are in trials which are performed with common treatments but in new populations, or in indications where no efficacious treatment exists up to now. Especially in these situations, adequate measures should be taken to exclude the use of blinded data for inference about the treatment effect. First of all, the use of alternative randomisation techniques to block randomisation (see e.g. [19]) could be considered. If block randomisation is used the block length should be chosen not too small. Fixed small block lengths are often used to avoid imbalances in patient numbers between the intervention groups, for example, if randomisation is stratified by centres with small sample sizes per centre. The results of our investigations add to the known disadvantages of this practice [20] and should be carefully weighted against its merits before application. Secondly, details about the randomisation procedure should not be described in the protocol but specified in a separate document that is withheld from all persons involved in the study. Furthermore, the process of generating the allocation sequence should be separated from the personnel that has access to the trial database. By this, it can definitely be avoided that knowledge about the randomisation scheme is used to derive information about the treatment effect from the blinded study data by application of the methods described above.

In case of a reasonable possibility of an overpowered study, the introduction of an interim analysis that allows for early stopping in case of overwhelming efficacy might be an option in addition to the measures mentioned above. This is a much more efficient way of conducting the trial than trying to recover treatment effect estimates from blinded data.

In summary, for all proposed approaches for blinded assessment of the treatment effect it holds true what Waksman states on the corrected EM algorithm based method, namely “it is interesting to note that the procedure can accurately and precisely estimate the difference when either the sample size or the true difference is sufficiently large” [10]. However, and this is the reassuring message: For combinations of sample sizes and treatment effects which are typical in clinical trials, no reliable inference can be made on the treatment group difference due to the bias and imprecision of the blinded estimates.

APPENDIX A: BLINDED MOMENT ESTIMATOR AND BLINDED F-TEST FOR CONTINUOUS DATA AND BLOCK LENGTH $l = 4$

Let B denote the covariance matrix of $Z_i = (Z_{1,i}, Z_{2,i}, Z_{3,i})^T$ with $Z_{1,i} = Y_{1,i} - Y_{2,i}$, $Z_{2,i} = Y_{2,i} - Y_{3,i}$, $Z_{3,i} = Y_{3,i} - Y_{4,i}$, i.e,

$$B = \sigma^2 \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \text{ and } B^{-1} = \frac{1}{4\sigma^2} \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} = \frac{1}{4\sigma^2} A.$$

Then $Z_i \sim N_3(\pm\Delta\eta_i, B)$ and $Z_i^T B^{-1} Z_i = \left\| B^{-1/2} Z_i \right\|^2 \sim \chi^2_{3;\,\Delta^2 \eta_i^T B^{-1} \eta_i}$, where $\eta_1 = (1, 0, -1)^T$, $\eta_2 = (1, -1, 1)^T$, $\eta_3 = (0, 1, 0)^T$, and $\chi^2_{3;nc}$ denotes the chi-square distribution with 3 degrees of freedom and non-centrality parameter *nc*. Since $\eta_i^T B^{-1} \eta_i = 1/\sigma^2$ for all *i*, $Z_i^T B^{-1} Z_i \sim \chi^2_{3;(\Delta/\sigma)^2}$ and $Z_i^T A Z_i \sim 4\sigma^2 \cdot \chi^2_{3;(\Delta/\sigma)^2}$.

Therefore, $E\left(\frac{1}{k} \sum_{i=1}^{k} Z_i^T A Z_i \right) = 12\sigma^2 + 4\Delta^2$ and since $E\left(\frac{1}{k-1} \sum_{i=1}^{k} (Z_{4,i} - \bar{Z}_{4,\bullet})^2 \right) = 4\sigma^2$, we get with the method of moments the estimator

$\hat{\Delta}^2 = \max\left(\frac{1}{4k} \sum_{i=1}^{k} Z_i^T A Z_i - \frac{3}{4} \frac{1}{k-1} \sum_{i=1}^{k} (Z_{4,i} - \bar{Z}_{4,\bullet})^2, 0 \right)$ which differs from the expression for $\hat{\delta}^2$ given in Reference [6] on page 484.

Analogously to block length $\ell = 2$, the following test statistics for a blinded F-test can be derived for block length $\ell = 4$: $F_{blind,\,l=4} = \dfrac{\dfrac{1}{3k}\sum_{i=1}^{k} Z_i^T A Z_i}{\dfrac{1}{k-1}\sum_{i=1}^{k}\left(Z_{4,i} - \bar{Z}_{4\bullet}\right)^2}$, where as in Reference [6] $Z_{4,i} = Y_{1,i} + Y_{2,i} + Y_{3,i} + Y_{4,i}$.

REFERENCES

1. ICH. ICH Harmonised Tripartite Guideline E9: Statistical Principles for Clinical Trials. *Statistics in Medicine* 1999; **18**:1905-1942. DOI: 10.1002/(SICI)1097-0258(19990815)18:15<1903::AID-SIM188>3.0.CO;2-F.
2. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *The Lancet* 2002; **359**:696-700. DOI:10.1016/S0140-6736(02)07816-9.
3. CHMP. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. Doc. Ref. CHMP/EWP/2459/02. London: 2007. http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf (accessed Nov. 3, 2008).
4. Fleming TR, Sharples K, McCall J, Moore A, Rodgers A, Stewart R. Maintaining confidentiality of interim data to enhance trial integrity and credibility. *Clinical Trials* 2008; **5**:157-167. DOI: 10.1177/1740774508089459.
5. CPMP. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. *Statistics in Medicine* 1995; **14**:1659-1682. DOI: 10.1002/sim.4780141507.
6. Van der Meulen EA. Are we really that blind? *Journal of Biopharmaceutical Statistics* 2005; **15**:479-489. DOI: 10.1081/BIP-200056540.
7. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics – Theory and Methods* 1992; **21**:2833-2853. DOI: 10.1080/03610929208830947.

8. Friede T, Kieser M. On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine* 2002; **21**:165-176. DOI: 10.1002/sim.977
9. Friede T, Kieser M. Authors' Reply to Letter to the Editor by Gould AL, Shih WJ. *Statistics in Medicine* 2005; **24**:154-156. DOI: 10.1002/sim.1894.
10. Waksman JA. Assessment of the Gould-Shih procedure for sample size re-estimation. *Pharmaceutical Statistics* 2007; **6**:53-65. DOI: 10.1002/pst.244.
11. Xing B, Ganju J. A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine* 2005; **24**:1807-1814. DOI: 10.1002/sim.2070.
12. Ganju J, Xing B. Re-estimating the sample size of an on-going blinded trial based on the method of randomization block sums. *Statistics in Medicine* 2009; **28**:24-38. DOI: 10.1002/sim.3442.
13. Hamilton, M. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry* 1960; **23**:56-62.
14. Montgomery SA. Clinically relevant effect sizes in depression. *European Neuropsychopharmacology* 1994; **4**:283-284.
15. Linde K, Mulrow CD, Berner M, Egger M. *St John's Wort for Depression (Review)*. The Cochrane Collaboration: Oxford, 2005. DOI: 10.1002/14651858.CD000448.pub3.
16. Barbui C, Furukawa TA, Cipriani A. Effectiveness of paroxetine in the treatment of acute major depression in adults: a systematic re-examination of published and unpublished data from randomized trials. *Canadian Medical Association Journal* 2008; **178**:296-305. DOI:10.1503/cmaj.070693.

17. Kalb R, Trautmann-Sponsel RD, Kieser M. Efficacy and tolerability of Hypericum extract WS 5572 versus placebo in mildly to moderately depressed patients. A randomized double-blind multicenter clinical trial. *Pharmacopsychiatry* 2001; **34**:96-103. DOI: 10.1055/s-2001-14280.
18. Casagrande JT, Pike MC. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* 1978; **34**:483-486.
19. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice.* Wiley: New York, 2002.
20. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *The Lancet* 2002; **359**:515-519. DOI:10.1016/S0140-6736(02)07683-3.

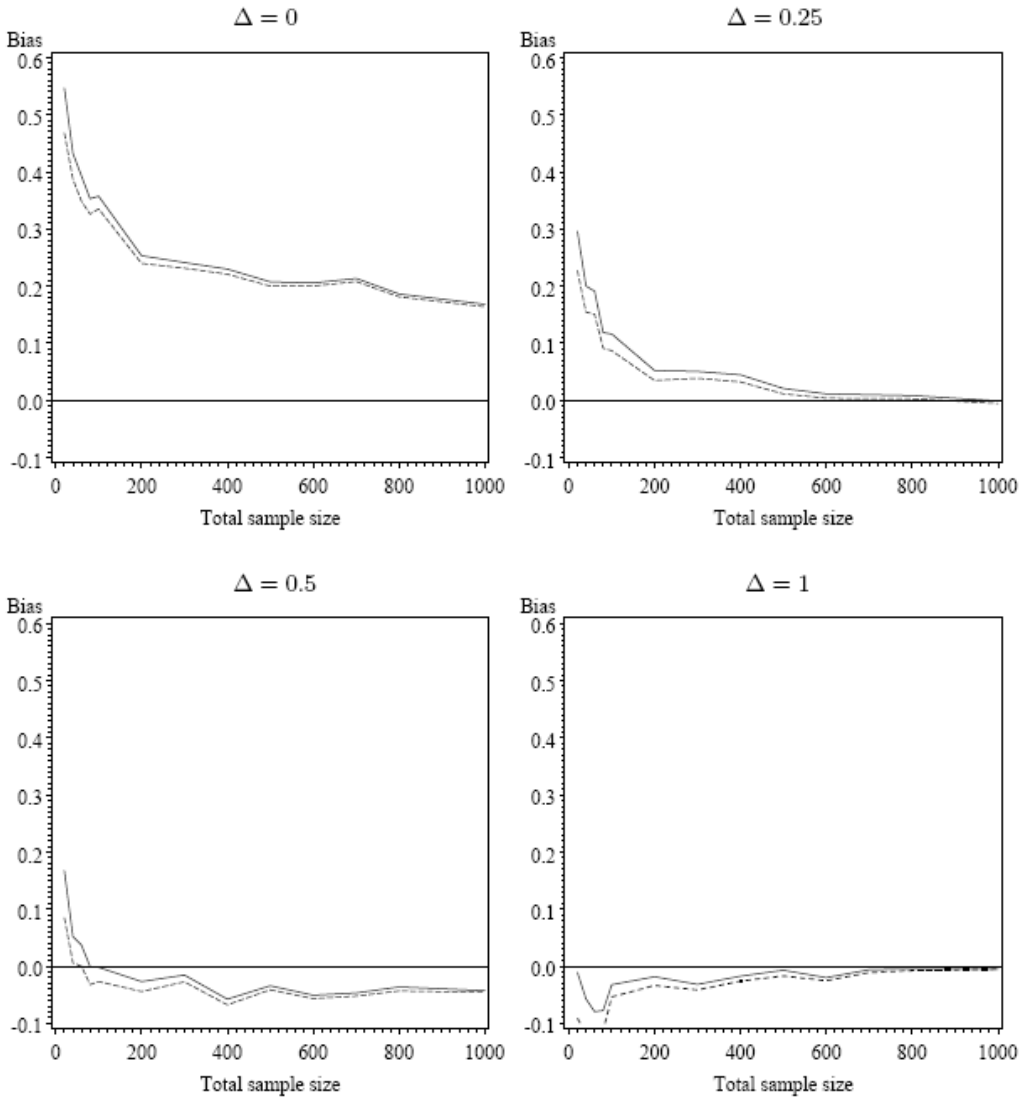

Figure 1. Mean simulated bias of the blinded treatment effect estimates for block length $l = 2$ in case of normally distributed data depending on the true treatment effect Δ ($\sigma = 1$) and the total sample size n used for estimation (solid line: blinded ML estimator, dotted line: blinded moment estimator).

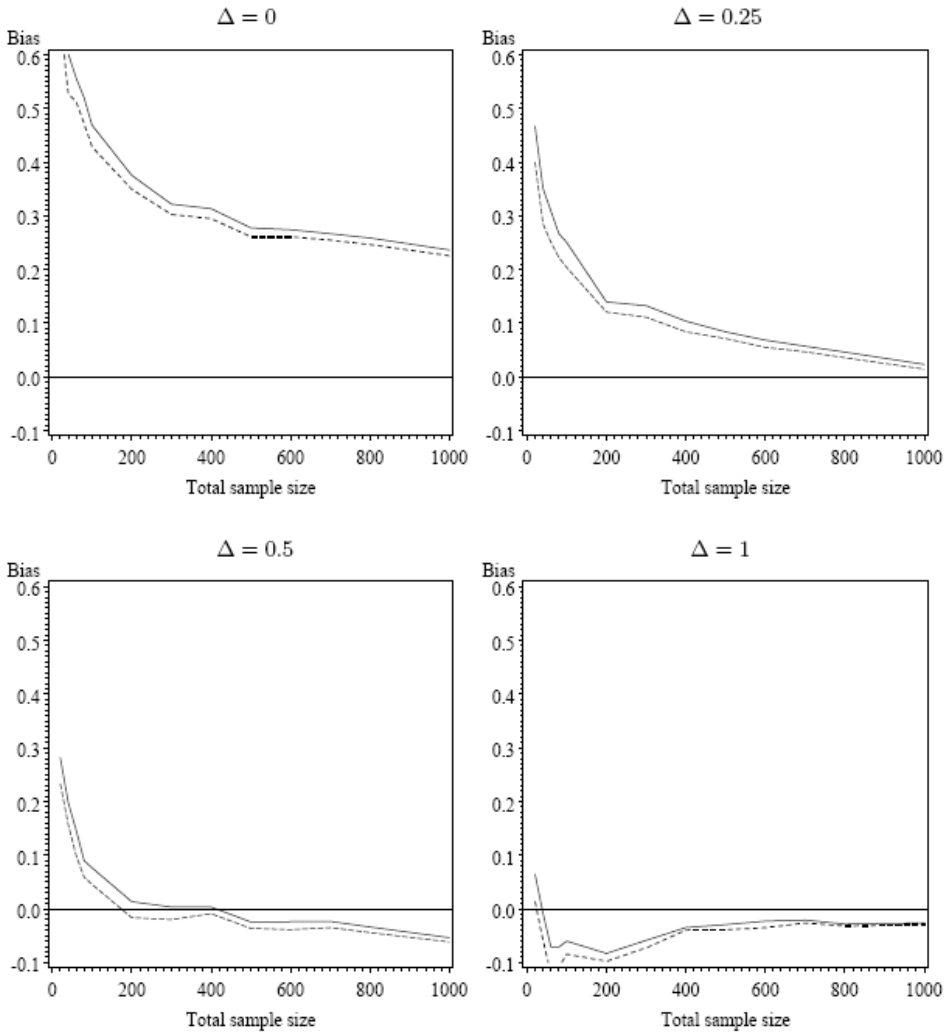

Figure 2. Mean simulated bias of the blinded treatment effect estimates for block length $l = 4$ in case of normally distributed data depending on the true treatment effect Δ ($\sigma = 1$) and the total sample size n used for estimation (solid line: blinded ML estimator, dotted line: blinded moment estimator).

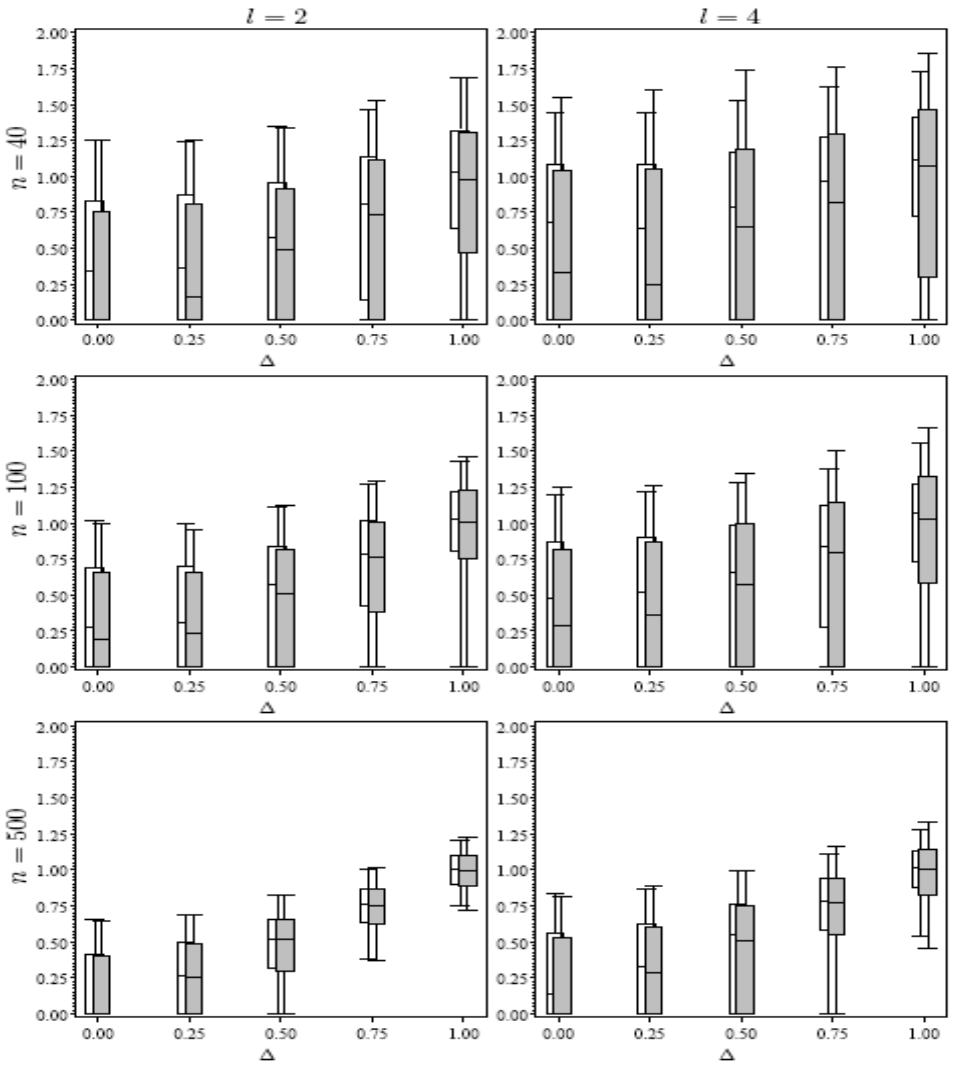

Figure 3. Box-whisker plot of the simulated distributions of the blinded estimates $|\hat{\Delta}|$ of the absolute value of the treatment effect $|\Delta|$ in case of normally distributed data depending on the true treatment effect Δ ($\sigma = 1$) and the total sample size n used for estimation (white: blinded ML estimator, grey: blinded moment estimator; in the box-whisker plot, bottom and top of the boxes are located at the first and third quartile, respectively, the central horizontal line marks the median, and 5 per cent and 95 per cent quantile are indicated by the ends of the whiskers).

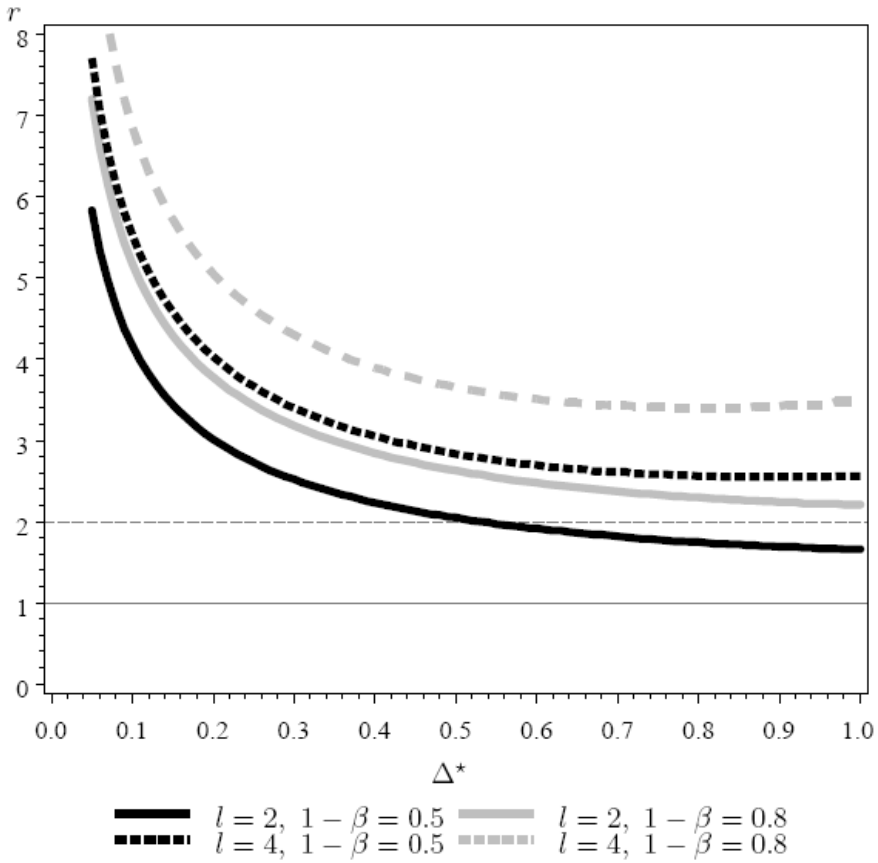

Figure 4. Ratio r of the true treatment effect Δ to the assumed effect Δ^* that is required to achieve a power of $1-\beta = 0.50$ and 0.80, respectively, for the blinded F-test. The sample size is chosen such that the unblinded two-sided *t*-test at $\alpha = 0.05$ has a power of 80 per cent at the assumed treatment effect Δ^* (σ=1).

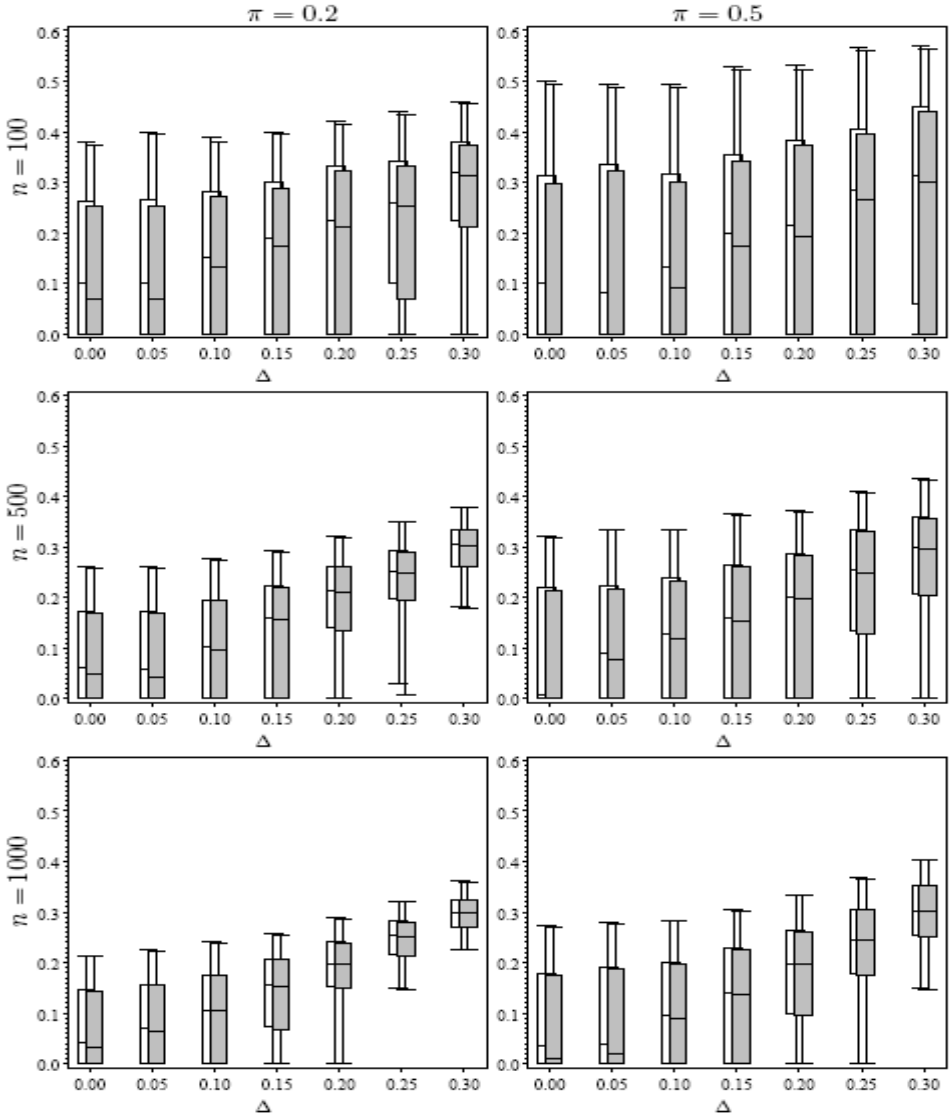

Figure 5. Box-whisker plot of the simulated distributions of the blinded estimates $|\hat{\Delta}|$ of the absolute value of the treatment effect $|\Delta|$ in case of binary data depending on the true treatment effect Δ, the overall event rate π and the total sample size n used for estimation (white: blinded ML estimator, grey: blinded moment estimator; in the box-whisker plot, bottom and top of the boxes are located at the first and third quartile, respectively, the central horizontal line marks the median, and 5 per cent and 95 per cent quantile are indicated by the ends of the whiskers).

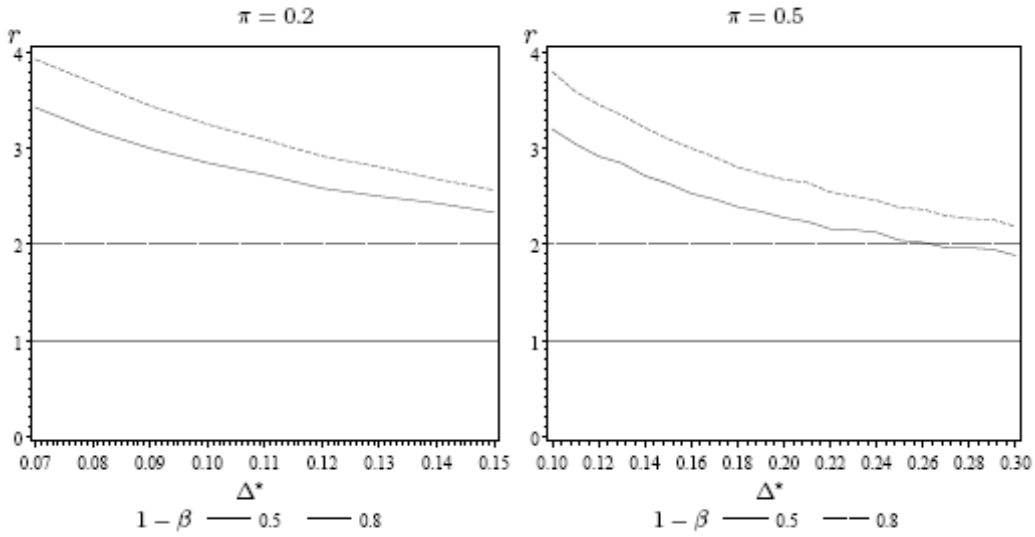

Figure 6. Ratio r of the true treatment effect Δ to the assumed effect Δ^* that is required to achieve a power of $1-\beta = 0.50$ and 0.80, respectively, for the blinded permutation test with overall event rates of $\pi = 0.2$ and 0.5. The sample size is chosen such that the unblinded two-sided Fisher's exact test at $\alpha = 0.05$ has a power of 80 per cent at the assumed treatment effect Δ^*.