The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints

M. Mielke

Department of Mathematical Stochastics, University Göttingen
e-mail: mmielke@math.uni-goettingen.de

A. Munk

Department of Mathematical Stochastics, University Göttingen
e-mail: munk@math.uni-goettingen.de

A. Schacht

Lilly Deutschland GmbH, Bad Homburg
e-mail: schacht_alexander@lilly.com

Abstract: The objective of this paper is to develop statistical methodology for non-inferiority hypotheses to censored, exponentially distributed time to event endpoints. Motivated by a recent clinical trial in depression, we consider a gold standard design where a test group is compared to an active reference and to a placebo group. The test problem is formulated in terms of a retention of effect hypothesis. Thus, the proposed Wald-type test procedure assures that the effect of the test group is better than a pre-specified proportion Δ of the treatment effect of the reference group compared to the placebo group. A sample size allocation rule to achieve optimal power is presented, which only depends on the pre-specified Δ and the probabilities for the occurrence of censoring. In addition, a pretest is presented for either the reference or the test group to ensure assay sensitivity in the complete test procedure. The actual type I error and the sample size formula of the proposed tests is explored asymptotically and by means of a simulation study showing good small sample characteristics. To illustrate the procedure a randomized, double blind clinical trial in depression is evaluated. An R-package for implementation of the proposed tests and for sample size determination accompanies this paper on the author's web page.

Keywords and phrases: censored data, exponential data, gold standard design, non-inferiority, optimal allocation, retention of effect.

1. INTRODUCTION

Recently, non-inferiority trials have gained in importance supported by the declaration of Helsinki [1] and a vigorous discussion on designing and evaluating such trials is ongoing. For a selective list of fundamental references see Jones et al. [2], Röhmel [3], D'Agostino et al. [4] and Senn [5]. In active controlled non-inferiority trials without a placebo arm the assay sensitivity, i.e. the ability of a study to distinguish between effective, less effective, and ineffective therapies (ICH 2000 [6]), is based on results from historical trials. In this manner the constancy condition is presumed, i.e. the active control effects in the active control trial patient population and the historical trial patient population are assumed to be equal. This assumption is not directly verifiable and its violation could result in statistical uncontrolled errors. Actually, for the treatment of depression there exists evidence that the placebo response is substantial (Dworkin [7]) and that it is increasing over time (Walsh [8]). The mentioned problems of active controlled trials are discussed by Rothmann et al. [9] and by Temple & Ellenberg [10] with regard to the declaration of Helsinki. They recommend the inclusion of a concurrent placebo group due to the problems of assay sensitivity if the

[*] Preprint submitted to *Statistics in Medicine* (5 September 2007), Accepted 28 April 2008.

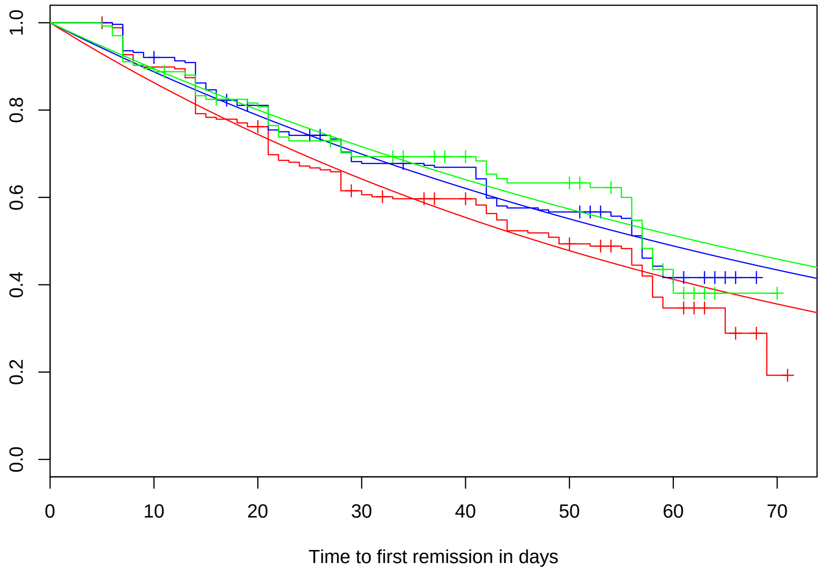

Fig 1. *Three-arm study for treatment of depression: Kaplan-Meier curves with marks for censoring times and fitted exponential survival curves for the endpoint "time to first response", test treatment (red), reference treatment (blue), and placebo (green).*

patients are not harmed by deferral of therapy and are fully informed about alternatives. Such a study design with reference, test treatment, and placebo group, which is called a gold standard design (Koch and Röhmel [11]), can be used to demonstrate superiority of either the reference or the test treatment to placebo as well as non-inferiority of the test treatment compared to the reference.

Pigeot et al. [12] as well as Koch & Röhmel [11] consider the gold standard design with normal endpoints and homogeneity of the variances. In both papers the non-inferiority test problem is based on the differences in means. In the case of binary endpoints we refer to Tang & Tang [13] and Kieser & Friede [14].

As emphasized above, the gold standard design is recommended in clinical studies which investigate the treatment of depression. In the therapy of depression, achieving remission is the clinically desired goal (Nierenberg & Wright [15]) whereas remission is defined as maintaining the Hamilton Rating Scale of Depression (HAM-D) total score at ≤ 7. Kieser & Friede [14] provide the statistical methodology to examine remission as binary endpoint or to be more precisely, the question whether the patient does achieve remission after treatment of acute symptoms or not. However, Yadid et al. [16] point out that in addition to remission the fast onset of action and the prevention of relapse are important and thus are the major goals of the present research. The primary endpoint time to first remission incorporates this issue. The occurrence of remission can be investigated over the complete time interval of the study. In the present, we analyze a randomized, double blind clinical trial on major depression where a new antidepressant is compared to a standard antidepressant, known for having a fast onset of action, and to placebo. The data representing the time to first remission are displayed in Figure 1. The time points are discretized due to weekly visit intervals, which are sufficient to monitor the occurrence of remission over the entire study period. In this paper, we assume that the time points to first remission are i.i.d. right censored, exponentially distributed in each group. The PP-plots in Figure 2 indicate a quite good fit of this model. We mention, however, that our method could also be generalized to different models such as Weibull or Gamma distributed endpoints.

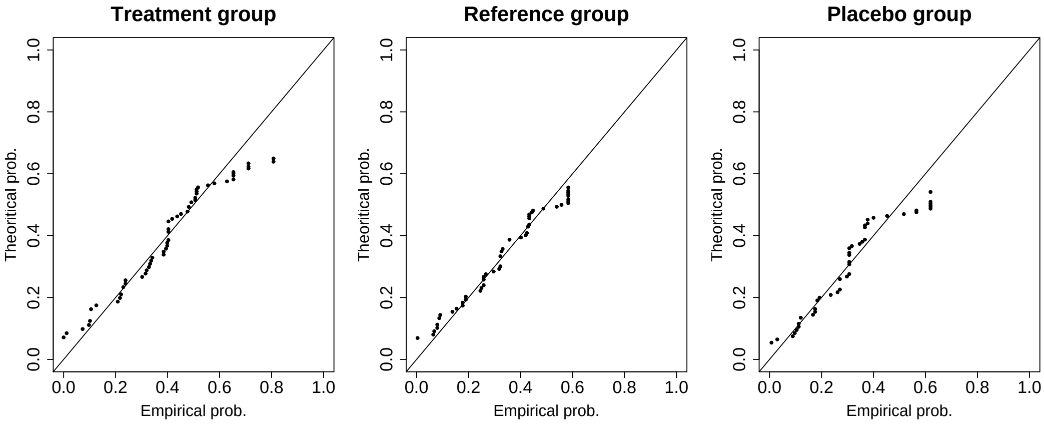

Fig 2. *PP-Plots for fitted exponential model vs. Kaplan-Meier.*

As statistical methods for showing non-inferiority have been mainly developed for binary or normal endpoints, the problem of assessing non-inferiority, where the primary variable is a survival time, has been far less investigated. Freitag [17] presents a review of existing approaches for the two-sample case. In contrast, the objective of this paper is to develop the statistical methodology for planning and assessing non-inferiority in a gold standard design, which the CPMP [18] recommends for clinical investigation of medicinal products for the treatment of depression. The endpoints are assumed to be censored, exponentially distributed. The presented analysis to show non-inferiority is based on a retention of effect hypothesis and on the resulting Wald-type test, which relies on the asymptotic normality of the maximum likelihood estimator. Throughout the paper, the resulting test is denoted as *RET (Retention of Effect Test)*. The hypothesis and the RET are presented in Section 2. In Section 3, a sample size formula to determine the required sample sizes to obtain a desired power is derived. This can be used to find the optimal allocation of the samples in terms of maximizing the power and minimizing the total sample size for given type II error, respectively. As a special case, we investigate the case of homogeneous censoring probabilities in all groups, where it turns out that the optimal allocation and the estimation of the asymptotic variance become particularly simple and only depend on Δ, the non-inferiority margin. In Section 4, we incorporate the retention of effect test in a complete two step test procedure which ensures assay sensitivity via a pretest for superiority of either the reference or the test treatment to placebo. Moreover, the matter of sample size adjustment for the complete test procedure is discussed. A major finding is that for the commonly used alternative of equal effects for the new and the reference treatment no correction of sample size is necessary to obtain a power of $1-\beta$ for the complete test procedure when the sample size is determined via the formula of the RET presented in Section 3. In particular, this result is valid independent of the considered active control effect. In Section 5, the results of a simulation study to investigate the finite sample behavior of the test decisions are presented. In Section 6, we revisit the clinical trial in major depression, introduced above, and apply our procedures. The new treatment turns out to be non-inferior to the reference. Finally, we comment on our software for analysis and planning of the introduced test procedures in Section 7 and conclude with a discussion in Section 8. All proof are deferred to a technical report [19] in order to keep the paper short and concise.

2. MODEL, HYPOTHESIS AND WALD-TYPE TEST

2.1. Model and Hypothesis

Let T_{ki} for $i = 1, \dots, n_k$ be independent, exponentially distributed survival times with expectation $\mathrm{E}[T_{ki}] = \lambda_k$, $k = R, T, P$, where R, T, and P abbreviates reference, treatment, and placebo group, respectively, in a three-arm clinical trial. Further, let the corresponding censoring times U_{ki} be independent distributed according to G_k where U_{ki} is independent of T_{ki} for $i = 1, \dots, n_k$ and $k = R, T, P$. The observations consist of pairs (X_{ki}, δ_{ki}), where $X_{ki} = \min\{T_{ki}, U_{ki}\}$ are the observed survival times and $\delta_{ki} = \mathbf{1}_{\{T_{ki} \leq U_{ki}\}}$, $i = 1, \dots, n_k$, $k = R, T, P$, are the corresponding censoring indicators. Hence, $\delta_{ki} = 1$ stands for an uncensored observation. Moreover, none of the groups should asymptotically vanish, i.e. for $k = R, T, P$ and $n = n_R + n_T + n_p$

$$\frac{n_k}{n} \longrightarrow w_k$$

holds for $n_R, n_T, n_p \to \infty$ and some $w_k \in (0, 1)$. Further, we assume that the probabilities for an uncensored observation should be positive, i.e.

$$p_k := P(\delta_{ki} = 1) > 0$$

for $k = R, T, P$. We emphasize that no assumptions on the censoring distribution G_k are made and that we only have to incorporate the probabilities p_k in the following modeling process.

Furthermore, we assume that small values for the observations X_{ki} are associated with higher efficacy of the treatment, e.g. we observe the time which elapses until healing, or in general until a positive impact occurs. Therefore, small values of λ_T are desirable. The hazard ratio, which is in the case of exponentially distributed endpoints just the ratio of the λ's, is the usual way of comparing time to event endpoints. Therefore, we consider the retention of a control effect on the log relative risk scale, which yields the following test problem

$$\begin{array}{rl} & H_0^N : \ \log \lambda_T - \log \lambda_P \geq \Delta \left(\log \lambda_R - \log \lambda_P \right) \\ \text{vs.} & K_0^N : \ \log \lambda_T - \log \lambda_P < \Delta \left(\log \lambda_R - \log \lambda_P \right), \end{array} \tag{1}$$

with $\Delta \in [0, \infty)$. The alternative K_0^N means that the test treatment T achieves more than $\Delta \times 100\%$ of the active control effect, where both are compared to placebo and the effect is measured via the log relative risk (cf. Rothmann et al. [9]). Note that the hypothesis can be equivalently formulated as

$$H_0^N : \ \frac{\lambda_T}{\lambda_P} \geq \left(\frac{\lambda_R}{\lambda_P} \right)^{\Delta}.$$

Testing for efficacy of the test treatment over placebo corresponds to a choice of $\Delta = 0$. Setting $\Delta = 1$ or even $\Delta > 1$ implies testing for superiority of the test over the reference treatment and substantial superiority, respectively. The main focus in this paper is on the case $\Delta \in (0, 1)$, which corresponds to showing non-inferiority of the test treatment to the reference and superiority of the test treatment to placebo. To this end, a non-inferiority margin Δ has to be determined in advance which represents a clinically irrelevant relative deviation between treatment and active control. This choice of Δ highly depends on the application and a general recommendation is difficult. We do not want to pursue this topic in detail, however, we refer to Lange & Freitag [20], who give a systematic review of 332 published clinical non-inferiority studies, especially addressing the methods applied to choose the equivalence margin. They point out that it is mostly recommended to set the margin Δ at least to 0.5, yielding an effect of the test treatment that is more closely located to the standard than to placebo. We have evaluated our data example in Section 6 according to this recommendation. Finally, note that our results are formulated such that they cover also the assessment of superiority, i.e. where $\Delta \geq 1$.

2.2. Wald-type test

In order to test problem (1), it is convenient to rewrite the hypothesis as

$$H_0^N : \log \lambda_T - \Delta \log \lambda_R + (\Delta - 1) \log \lambda_P \geq 0,$$

which allows to consider the contrast $\eta = \log \lambda_T - \Delta \log \lambda_R + (\Delta - 1) \log \lambda_P$. The maximum likelihood (ML) estimator for η is given by

$$\hat{\eta} = \log \hat{\lambda}_T - \Delta \log \hat{\lambda}_R + (\Delta - 1) \log \hat{\lambda}_P \, , \tag{2}$$

with the MLE for λ_k

$$\hat{\lambda}_k = \frac{X_k}{\delta_k} \, , \tag{3}$$

for $k = R, T, P$ with $X_k = \sum_{i=1}^{n_k} X_{ki}$ and $\delta_k = \sum_{i=1}^{n_k} \delta_{ki}$. Note that the distribution of (X_k, δ_k) can be represented as a full exponential family and hence the ML-estimator $\sqrt{n}\,(\hat{\eta} - \eta)$ is asymptotically normal with variance (confer Theorem A.1, Mielke et al. [19])

$$\sigma^2 = \frac{1}{w_T \, p_T} + \frac{\Delta^2}{w_R \, p_R} + \frac{(\Delta - 1)^2}{w_P \, p_P}. \tag{4}$$

With $\hat{p}_k = \delta_k / n_k$ and $w_k = n_k / n$ we obtain a surprisingly simple estimator for σ^2 as

$$\hat{\sigma}^2 = n \left(\frac{1}{\delta_T} + \frac{\Delta^2}{\delta_R} + \frac{(\Delta - 1)^2}{\delta_P} \right). \tag{5}$$

Note that only the total number of uncensored observations δ_k, the total sample size n, and the non-inferiority margin Δ are required for computation of $\hat{\sigma}^2$. In particular, this estimator is independent of the observed survival times.

Hence, we obtain as the test-statistic

$$T = \sqrt{n}\, \frac{\hat{\eta}}{\hat{\sigma}} = \frac{\log \hat{\lambda}_T - \Delta \log \hat{\lambda}_R + (\Delta - 1) \log \hat{\lambda}_P}{\sqrt{\frac{1}{\delta_T} + \frac{\Delta^2}{\delta_R} + \frac{(\Delta-1)^2}{\delta_P}}} \, , \tag{6}$$

which is asymptotically standard normally distributed for $\eta = 0$, i.e. if $\log \lambda_T - \log \lambda_P = \Delta\,(\log \lambda_R - \log \lambda_P)$ in (1). Thus, for a given level of significance α the hypothesis H_0^N will be rejected and non-inferiority can be claimed if

$$T < z_\alpha \, , \tag{7}$$

where z_α denotes the α-quantile of the standard normal distribution. In the following, we will call the test in (7) *RET (Retention of Effect Test).*

Remark. Alternatively, the retention of effect hypothesis from (1) could be defined in the ratio of differences in means, i.e. through $H_0^{N^*} : (\lambda_P - \lambda_T) \leq \Delta^*(\lambda_P - \lambda_R)$, (Hung et al. [21]). Our methods can be easily extended to a Wald-type test for $H_0^{N^*}$. However, a straight forward calculation shows that in this case the asymptotic variance will depend on the parameters λ_k, $k = R, T, P$. In contrast for the RET in (7) for the hypothesis (1), the asymptotic variance σ^2 in (4) is independent of the parameters λ_k, $k = R, T, P$. That has the advantage that the variance can be estimated unrestricted, see (5), in contrast to most situations where retention of effect hypothesis are tested by a Wald-type test (Kieser & Friede [14]).

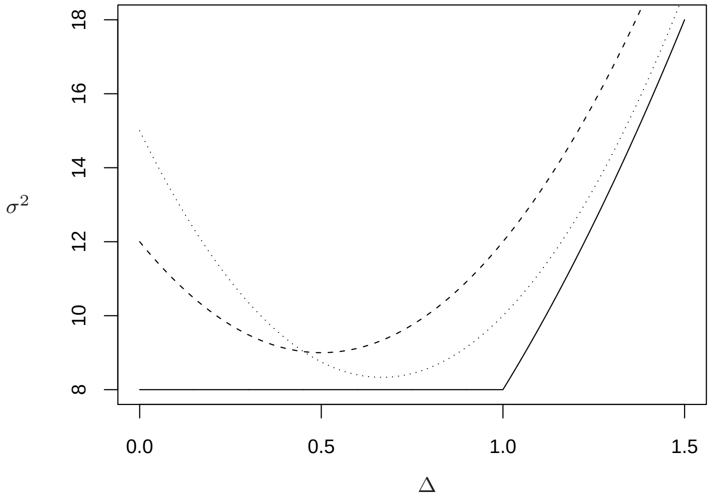

Fig 3. *Comparison of asymptotic variances for optimal allocation (solid line), balanced allocation (dashed line), and 2:2:1 allocation (dotted line),* $p = 0.5$.

3. OPTIMAL ALLOCATION AND SAMPLE SIZE DETERMINATION

The ML-estimator $\hat{\eta}$ in (2) is asymptotically normally distributed for general $\eta = \log \lambda_T - \Delta \log \lambda_R + (\Delta - 1) \log \lambda_P$ (see Theorem A.1, Mielke et al. [19]). Thus, the asymptotic power of the Wald-type test can be approximately calculated for specified λ_k, p_k, and n_k, $k = R, T, P$, given Δ, and significance level α (see Section 2.2) as

$$1 - \beta = P\left(T \leq z_\alpha\right) = P\left(T - \frac{\sqrt{n}\,\eta}{\sigma} \leq z_\alpha - \frac{\sqrt{n}\,\eta}{\sigma}\right) \approx \Phi\left(z_\alpha - \frac{\sqrt{n}\,\eta}{\sigma}\right), \tag{8}$$

where Φ denotes the cumulative distribution function of the standard normal distribution, $w_k = n_k/n$ in the asymptotic variance σ^2 from (4).

In planning a clinical trial, it is a crucial step to determine the required sample size to achieve a given power $1-\beta$ for a specified parameter constellation η in the alternative K_0. By (8) the requirement of achieving at least a given power of $1 - \beta$ is asymptotically equivalent to

$$z_{1-\beta} \leq z_\alpha - \frac{\sqrt{n}\,\eta}{\sigma}.$$

This is equivalent to

$$n \;\geq\; \frac{\sigma^2}{\eta^2}\, \left(z_\alpha - z_{1-\beta}\right)^2 \tag{9}$$

for $\eta \in K_0$, i.e. $\eta < 0$. Note that each term on the right hand side other than σ^2 is fixed in planning a clinical trial. The variance σ^2 depends through w_k, $k = R, T, P$, on the allocation, which is under control of the investigator [12]. Therefore, it could be chosen optimal in terms of minimizing σ^2 and therewith the required sample size in order to achieve a given power $1 - \beta$.

Substituting $w_T = 1 - w_R - w_P$ yields

$$\sigma^2 = \frac{1}{(1 - w_R - w_P)\, p_T} + \frac{\Delta^2}{w_R\, p_R} + \frac{(1-\Delta)^2}{w_P\, p_P}. \tag{10}$$

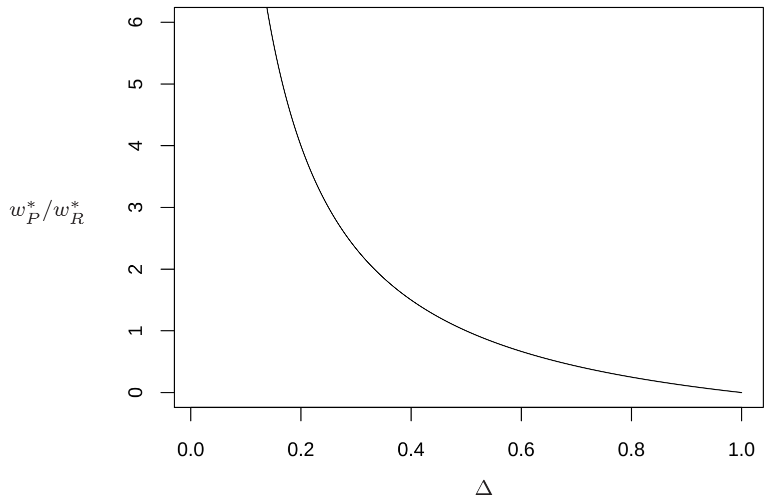

FIG 4. *Optimal allocation for given Δ under the assumption of homogeneous censoring probabilities.*

Minimizing σ^2 as a function of w_R and w_P w.r.t. the constraint $w_R + w_P \leq 1$ gives (see for details Theorem A.2, Mielke et al. [19])

$$w_R^* = \frac{\Delta\, p_R^{-1}}{p_T^{-1} + \Delta\, p_R^{-1} + |1-\Delta|\, p_P^{-1}} \quad \text{and} \quad w_P^* = \frac{|1-\Delta|\, p_P^{-1}}{p_T^{-1} + \Delta\, p_R^{-1} + \,|1-\Delta|\, p_P^{-1}}\,.$$

Thus, the *optimal allocation* of the samples is given by

$$n_T^* : n_R^* : n_P^* \;=\; 1 \;:\; \Delta\sqrt{p_T/p_R} \;:\; |1-\Delta|\sqrt{p_T/p_P}, \tag{11}$$

which yields a minimal total required sample size of

$$n^* = \left(\frac{1}{\sqrt{p_T}} + \frac{\Delta}{\sqrt{p_R}} + \frac{|1-\Delta|}{\sqrt{p_P}}\right)^2 \left(\frac{z_\alpha - z_{1-\beta}}{\eta}\right)^2. \tag{12}$$

From (12) we see that the total required sample size is a monotone decreasing function in each p_k, $k = T, R, P$, and it is minimal in the case of uncensored observations, i.e. $p_T = p_R = p_P = 1$. Further, the monotonicity provides a *worst case scenario* for sample size planning by means of presuming homogeneous censoring probabilities in the three groups, $k = T, R, P$, and setting the common censoring probability to the smallest value, i.e. $p = \min\{p_T, p_R, p_P\}$. In planning a clinical trial, one would expect $p_T, p_R > p_P$ and hence $p_P = \min\{p_T, p_R, p_P\}$ because the reference and the test treatment are expected to be efficient, i.e. $\lambda_T, \lambda_R < \lambda_P$, which implies under identically censoring variables U_k in the groups that reference and test treatment are less affected by censoring than placebo. Hence, a conservative recommendation for planning the trial is to assume that all censoring probabilities equal p_P. This simplifies the optimal allocation rule significantly, as we will see below. In particular, it accentuates that the optimal allocation coincides with the case of normal endpoints by Pigeot et al. [12] and Schwartz & Denne [22].

The assumption of homogeneous censoring probabilities and $\Delta \in [0, 1)$ simplifies the optimal allocation (11) to

$$n_T^* : n_R^* : n_P^* \;=\; 1 \;:\; \Delta \;:\; (1-\Delta).$$

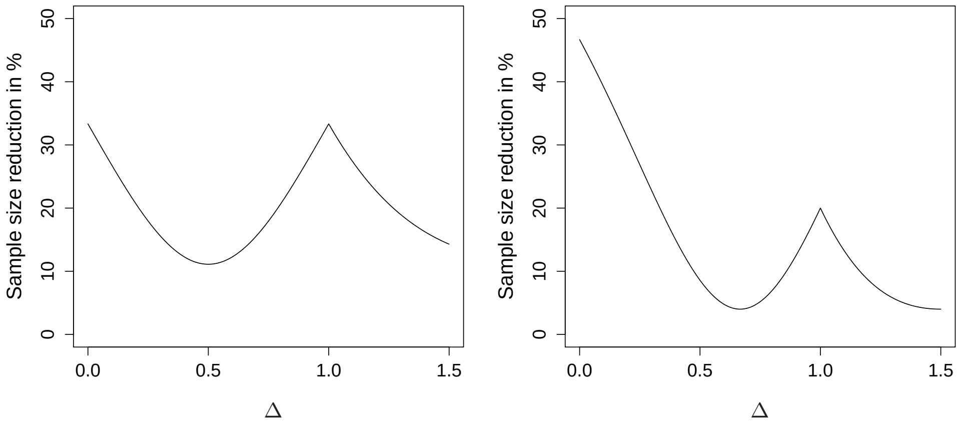

FIG 5. *Reduction in total sample size when optimal allocation is used instead of balanced allocation (left figure), and instead of 2:2:1 allocation (right figure).*

This yields with a minimal asymptotic variance of $\sigma^2_{opt} = 4/p$ the required sample sizes in groups T, R, and P as

$$(n^*_T, n^*_R, n^*_P) = n^* \left(w^*_T, w^*_R, w^*_P\right) = \frac{2}{p}\left(\frac{z_\alpha - z_{1-\beta}}{\eta}\right)^2 (1\ ,\ \Delta\ ,\ 1-\Delta)\,. \tag{13}$$

Hence, in the case of homogeneous censoring probabilities and $\Delta \in [0, 1)$ the first half of the total samples should always be assigned to the test group, the other half allocated in a ratio $n_R : n_P = 1 : (1 - \Delta)/\Delta$ (see Figure 4) to the reference, and the placebo group independent of the censoring probability p. Hence, it is also valid for the non-censored case. The comparison in Figure 3 of the optimal asymptotic variance with the asymptotic variance, when a balanced and $2 : 2 : 1$ allocation, respectively, is used, points out the capability of reduction of the asymptotic variance by reallocating. The condition (9) yields that the reduction in the asymptotic variance from (10) is equivalent to the reduction in required total sample size. This is illustrated in Figure 5, where the reduction for using the optimal allocation instead of a balanced and $2 : 2 : 1$ allocation, respectively, is presented. For the balanced design a reduction of at least 10% is always possible and even more than 30% for Δ close to zero or one. The allocation $2 : 2 : 1$ is more appropriate for $\Delta \in [0.5, 1)$, but a reduction up to 20% is still possible by reallocating to the optimal allocation. Figure 6 presents the total required sample size for $p = 0.5$, and different values of $\Delta = 0.5, 0.7, 0.8, 0.85$ in dependence of the active control effect λ_P/λ_R for the alternative $\lambda_T = \lambda_R$, significance level $\alpha = 0.05$, and a desired power of $1 - \beta = 0.8$.

In the case of homogeneous censoring probabilities and assessing superiority, $\Delta \geq 1$, the optimal allocation from (11) becomes $n^*_T : n^*_R : n^*_P = 1 : \Delta : (\Delta - 1)$. Hence, in contrast to the case of assessing non-inferiority, $\Delta < 1$, the first half of the total samples should always be assigned to the reference and not to the test treatment group, the other half allocated in a ratio $n_T : n_P = 1 : (\Delta - 1)$ to the test treatment and the placebo group.

4. COMPLETE TEST PROCEDURE

The test problem considered so far is to show non-inferiority of the test treatment to the reference. The inclusion of a placebo group makes it possible to directly demonstrate the effectiveness of a therapy and therewith ensures assay sensitivity of the test procedure. Pigeot et al. [12] carry out a pretest for superiority of the reference treatment to placebo, which provides internal evidence of assay sensitivity. Though, Koch [23] points out that this

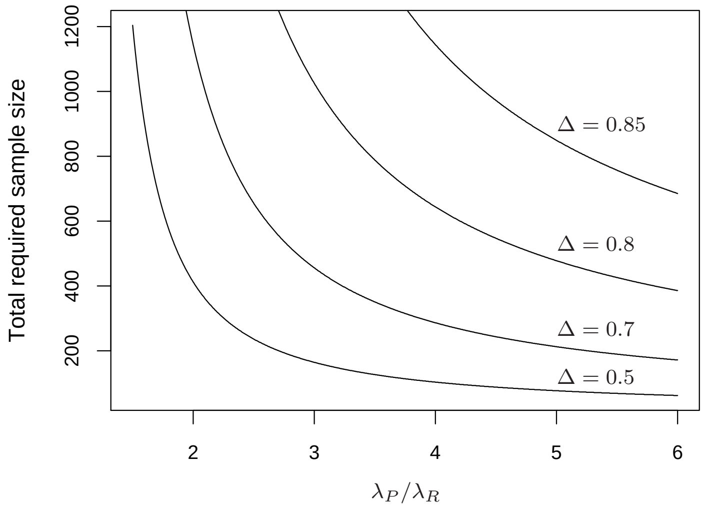

FIG 6. *Total required sample size for* $p = 0.5$ *and* $\Delta = 0.5, 0.7, 0.8, 0.85$ *in dependence of the active control effect* λ_P/λ_R *for the alternative* $\lambda_T = \lambda_R$*, significance level* $\alpha = 0.05$*, and a desired power of* $1 - \beta = 0.8$*.*

procedure would blame a test treatment that has shown to be superior to placebo and non-inferior to the reference for the fact that reference could not beat placebo. Therefore, Koch & Röhmel [11] perform a pretest for superiority of the test treatment to placebo instead. It is not the objective of this paper to take up this discussion. But we state that in any case a two-step test procedure must be conducted to establish non-inferiority and effectiveness of the test treatment, where in a first step a pretest for superiority of either the reference or the test treatment to placebo is performed, and in a second step the non-inferiority is investigated. The pretest for superiority of a treatment to placebo coincides with rejecting the null hypothesis $H_{0,j}^S : \lambda_j \geq \lambda_P$ either for the reference treatment $(j = R)$ or for the test treatment $(j = T)$. Thus, the overall hypothesis is given by

$$H_0 : \quad H_0^N \cup H_{0,j}^S \;=\; \{\log \lambda_T - \log \lambda_P \geq \Delta\,(\log \lambda_R - \log \lambda_P)\} \cup \{\lambda_j \geq \lambda_P\}\,,$$

where H_0 is rejected if the sub-tests for H_0^N and $H_{0,j}^S$ can be rejected. In order to avoid a misunderstanding, note that H_0 either includes $H_{0,R}^S$ or $H_{0,T}^S$ and not both at once. Due to the principles of intersection-union-tests, this test decision for H_0 does not exceed a level α if H_0^N and $H_{0,j}^S$ are tested the level α, respectively. Therefore, the power for rejecting H_0 is reduced compared to simple testing H_0^N. However, in section 4.2 it will turn out that this reduction is negligible for the commonly used alternative $\lambda_T = \lambda_R < \lambda_P$. Similar results were obtained by Pigeot et al. [12] and Kieser & Friede [14] for normal and binomial endpoints, respectively.

4.1. Two-sample Wald-type test for superiority

The ML-estimators $\hat{\vartheta}_j$ for $\vartheta_j := \log \lambda_j - \log \lambda_P$ are obtained as in Section 2.2 by plugging in the ML-estimators given in (3) for λ_k, $k = R, T, P$. Moreover, these estimators can be shown to be asymptotically normal in the same way as $\hat{\eta}$,

$$\sqrt{n_j + n_P}\;(\hat{\vartheta}_j - \vartheta_j) \;\xrightarrow{\mathfrak{D}}\; \mathcal{N}(0, \sigma_j^2)$$

	Δ							
	$H_{0,R}^S$				$H_{0,T}^S$			
λ_T:λ_R:λ_P	0.5	0.7	0.8	0.9	0.5	0.7	0.8	0.9
0.8:1:1.1	**15.24**	**15.04**	**13.76**	**11.26**	**77.11**	**68.87**	**59.59**	**42.32**
0.8:1:1.2	**26.60**	**27.85**	**25.56**	**19.82**	**83.50**	**79.40**	**72.48**	**55.41**
0.8:1:1.5	**51.16**	**59.88**	**59.08**	**49.09**	**91.18**	**92.43**	**90.68**	**81.13**
0.8:1:2	**68.94**	**82.80**	**85.32**	**80.38**	**94.74**	97.42	97.68	95.32
0.8:1:3	**80.39**	**94.04**	96.74	96.68	96.63	99.17	99.59	99.48
0.8:1:5	**86.61**	97.88	99.39	99.72	97.59	99.69	99.93	99.97
0.9:1:1.1	**29.13**	**30.91**	**28.52**	**22.07**	**84.56**	**81.21**	**74.87**	**58.20**
0.9:1:1.2	**49.45**	**57.56**	**56.49**	**46.47**	**90.78**	**91.79**	**89.74**	**79.48**
0.9:1:1.5	**74.83**	**89.13**	**92.13**	**90.09**	95.74	98.45	98.89	98.06
0.9:1:2	**85.35**	97.26	99.06	99.45	97.39	99.61	99.89	99.93
0.9:1:3	**90.42**	99.24	99.91	99.99	98.16	99.88	99.99	100.00
0.9:1:5	**92.85**	99.71	99.99	100.00	98.54	99.95	100.00	100.00
1.1:1:1.5	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\lambda_T = \lambda_R$	96.94	99.99	100.00	100.00	99.21	100.00	100.00	100.00

TABLE 1

*Approximated power of the pretests for $H_{0,R}^S$ and $H_{0,T}^S$ in percent for $\alpha = 0.05$, and a desired power of $1 - \beta = 0.8$ for the RET under optimal allocation. Values **less than 95%** are bold.*

with

$$\sigma_j^2 = (w_j + w_P)\left(\frac{1}{w_j\, p_j} + \frac{1}{w_P\, p_P}\right), \quad j = R, T.$$

Therefore, one rejects $H_{0,j}^S$ at level α, i.e. one concludes superiority of the reference or the test treatment to placebo, respectively, if $\sqrt{n_j + n_P}\, \hat{\vartheta}_j / \hat{\sigma}_j \leq z_\alpha$ holds with

$$\hat{\sigma}_j^2 = (n_j + n_P)(1/\delta_j + 1/\delta_P),$$

and z_α the α-quantile of the standard normal distribution.

4.2. Sample size adjustment for the complete test procedure

In Section 3, we derived for the three-sample non-inferiority RET, the optimal sample size allocation in terms of minimizing the total required sample sizes, and corresponding formulas for sample size determination. Based on these results, we will now derive the approximated power $1 - \tilde{\beta}_j$ of the pretests $H_{0,T}^S$ and $H_{0,R}^S$, respectively, when the sample sizes are determined with (13) to obtain a power of $1 - \beta$ for the RET. It will turn out that for the commonly used alternative $\lambda_T = \lambda_R$ no correction of the sample size is necessary to obtain a power of $1 - \beta$ for the complete test procedure because the power of the pretests is always larger than $1 - \beta$ for $\Delta \geq 1/3$, which covers the range of practical interest for non-inferiority tests, cf. Lange & Freitag [20].

We restrict our considerations to homogeneous censoring probabilities, i.e. $p = p_T = p_R = p_P$, and $\Delta \in [0, 1)$. The power $1 - \tilde{\beta}_j$ of the test decisions for superiority of $j \in \{R, T\}$ to placebo introduced in the previous Section 4.1 can be approximated by

$$1 - \tilde{\beta}_j \;\approx\; \Phi\left(z_\alpha - \sqrt{n_j + n_P}\,\frac{\vartheta_j}{\sigma_j}\right) \tag{14}$$

with Φ the cumulative distribution function of the standard normal distribution. Substituting the approximately required optimal sample sizes (n_T^*, n_R^*, n_P^*) from (13) for the RET to obtain a power of $1-\beta$, which presumes the allocation $n_T^* : n_R^* : n_P^* = 1 : \Delta : (1-\Delta)$, in (14) yields by straightforward calculations

$$1 - \tilde{\beta}_j \;\approx\; \Phi\left(z_\alpha + \frac{(\Delta-1)\,\vartheta_j}{|\eta|}\, b_j(\Delta)\right) \tag{15}$$

with

$$b_T(\Delta) = (z_{1-\beta} - z_\alpha)\sqrt{\frac{2}{(1-\Delta)(2-\Delta)}} \tag{16}$$

and

$$b_R(\Delta) = (z_{1-\beta} - z_\alpha)\sqrt{\frac{2\,\Delta}{1-\Delta}}\ . \tag{17}$$

As a first result, one observes that the power of the pretests is independent of the censoring probability p. For the commonly used alternative $\eta < 0$ with $\lambda_T = \lambda_R > \lambda_P$

$$|\eta| = \Delta \log\lambda_R - \log\lambda_T - (\Delta-1)\log\lambda_P = (\Delta-1)\vartheta_j$$

holds for $j = T, R$, which simplifies (15) to

$$1 - \tilde{\beta}_j \;\approx\; \Phi\left(z_\alpha + b_j(\Delta)\right)\ .$$

The case $\lambda_T = \lambda_R$. Therefore, under the alternative $\lambda_T = \lambda_R$ the power of the pretests is in addition independent of the effect size λ_P/λ_R and only depends on Δ besides α and β. This allows to estimate the power of the pretests from the power $1-\beta$ of the RET as at least $1-\beta$, again, for a range of Δ, s.t. $1/3 \le \Delta < 1$. To this end, observe that $b_T(\Delta)$ and $b_R(\Delta)$ in (16) and (17),respectively, are always $\ge (z_{1-\beta} - z_\alpha)$.

In fact, numerical investigations show that the power of the pretest is often even better. This is illustrated in table 1, which shows for different effect constellations $\lambda_T : \lambda_R : \lambda_P$ the approximated power of the pretests for $H_{0,R}^S$ and $H_{0,T}^S$ given in (15). The values are calculated for a significance level of $\alpha = 0.05$ and a desired power for the RET of 0.8. Under $\lambda_T = \lambda_R$ even for a small Δ of 0.5 the power of both pretests is with 96.94% and 99.21%, respectively, nearly 1 and increases for increasing Δ. Hence, the power of the complete test is almost equal to the power $1-\beta$ of the RET, and no adjustment is required to obtain an overall power of $1-\beta$.

The case of different effect sizes. The power of the pretests is increasing in the test-reference effect λ_T/λ_R for a fixed reference-placebo effect λ_P/λ_R. It can decrease drastically when the complete trial is planned via the RET and under an alternative $\lambda_T < \lambda_R < \lambda_P$, e.g. for $H_{0,R}^S$, $\lambda_T : \lambda_R : \lambda_P = 0.8 : 1 : 1.1$, and $\Delta = 0.5$ one ends up with a power of only 15.24%. Therefore, one has to be aware of a possibly significant reduction in power when planning under an alternative $\lambda_T < \lambda_R < \lambda_P$. Due to the parameter constellations $\lambda_T < \lambda_R$, this loss in power is a more serious problem for $H_{0,R}^S$ than for $H_{0,T}^S$, confer Table 1. Moreover, the test for superiority to placebo is more powerful for the test treatment than for the reference in general and especially also for $\lambda_T = \lambda_R$ due to the sample allocation $n_T^* : n_R^* : n_P^* = 1 : \Delta : (1-\Delta)$ used. Hence, in the considered complete test procedure from a statistical point of view the pretest for $H_{0,T}^S$ is preferred to those for $H_{0,R}^S$. But we emphasize again that both pretests have a power of almost 1 and no adjustment of sample size is required under the commonly used alternative $\lambda_T = \lambda_R$.

5. SIMULATIONS

In the following, we present the main results of extensive simulations studies for the actual type I error of the RET, confer Section 2.2, for the RET sample size formula (13) and for the

		λ_P/λ_R						
n	$n_T : n_R : n_P$	10	8	5	3	2	1.5	1.2
30	1:1:1	**5.69**	**5.89**	**5.71**	**5.71**	**6.02**	**5.94**	**6.17**
	2:2:1	5.35	5.10	5.31	5.26	4.91	5.02	5.20
	2:1:1	4.68	4.44	4.55	4.78	4.64	4.78	4.63
60	1:1:1	**5.90**	**5.67**	**5.62**	5.29	5.21	**5.59**	5.35
	2:2:1	5.23	5.25	4.91	4.56	5.22	5.19	5.26
	2:1:1	4.92	4.23	4.79	4.87	4.84	4.32	5.02
120	1:1:1	**5.80**	5.29	5.47	5.35	**5.60**	5.42	5.47
	2:2:1	4.83	5.00	5.28	4.80	4.88	5.35	4.66
	2:1:1	4.76	4.46	4.30	4.94	4.91	4.56	4.73
240	1:1:1	4.97	5.31	5.43	5.22	4.96	5.09	5.30
	2:2:1	5.05	5.20	5.03	5.06	4.87	5.11	4.87
	2:1:1	5.02	4.77	4.54	5.05	4.79	5.06	4.61
480	1:1:1	5.01	**5.68**	5.11	5.03	5.01	5.10	5.15
	2:2:1	5.13	5.41	5.06	4.86	5.19	5.32	5.07
	2:1:1	4.90	4.75	5.15	5.13	4.55	4.94	5.00
960	1:1:1	5.35	5.13	4.98	5.42	5.02	5.22	5.09
	2:2:1	4.89	4.77	4.74	5.14	4.87	4.83	4.97
	2:1:1	4.61	4.89	4.74	4.63	4.98	5.09	4.64
1440	1:1:1	4.94	5.23	5.08	5.37	4.98	5.25	5.31
	2:2:1	4.84	4.95	5.25	5.08	4.83	4.83	4.75
	2:1:1	5.03	5.10	5.23	5.27	4.60	4.89	4.63

TABLE 2
Simulated actual type I error in % for a nominal significance level of $\alpha = 0.05$, $p_T = p_R = p_P = 0.8$, $\Delta = 0.5$, and 10 000 replications. Values ***larger than 5.5%*** *are bold.*

derived power of the pretests presented in Section 4.1. Moreover, the power of the complete test procedures introduced in the previous section is simulated when the trial is conducted via the optimally allocated sample size for the RET. It turns out that the large sample framework presented in this paper yields even for small total sample sizes quite satisfactory results and that a finite sample adjustment is not necessary, in general.

The following investigations are based on a nominal significance level of $\alpha = 0.05$. However, similar results are obtained for a nominal significance level of $\alpha = 0.025$ (not displayed).

5.1. Type I error

To investigate the finite sample behavior of the RET we simulated the actual type I error for a nominal significance level $\alpha = 0.05$, the optimal allocation ratio, allocation ratios of 1:1:1 and 2:2:1, active control effects of $\lambda_P/\lambda_R = 10, 8, 5, 3, 2, 1.5, 1.2$, retention of effects $\Delta = 0.5, 0.7$, and total sample sizes $n = 30, 60, 120, 240, 480, 960, 1440$. The probabilities for an uncensored observation in the three groups T, R, and P are assumed to be homogeneous with $p_T = p_R = p_P = 0.8$. All parameter constellations were simulated with 10 000 replications. The results for $\Delta = 0.5$ are presented in Table 2. The results for $\Delta = 0.7$ are similar and omitted due to the marginal gain of insight.

For small total sample sizes ($n < 120$) the Wald-type test tends to be somewhat anticonservative for the balanced design and somewhat conservative for the optimal design (2:1:1 for $\Delta = 0.5$), whereas the unbalanced 2:2:1 attains the nominal significance level $\alpha = 0.05$. However, the magnitudes of these discrepancies are negligible and for total sample sizes round about 120 and more the nominal significance level of $\alpha = 0.05$ is attained almost exactly for all three designs. It is worth to note that these observations for the actual type I error can be made independently of the underlying active control effect and the choice of Δ.

		Pretest		**RET**	**Complete test procedure**	
Δ	λ_T:λ_R:λ_P	$H_{0,R}^S$	$H_{0,T}^S$	H_0^N	$H_0^N \cup H_{0,R}^S$	$H_0^N \cup H_{0,T}^S$
0.5	0.8:1:1.1	14.75	77.53	79.92	**11.39**	**71.60**
	0.8:1:1.2	26.99	84.03	80.48	**21.62**	75.95
	0.8:1:1.5	52.43	92.55	80.55	**41.75**	79.53
	0.8:1:2	69.65	95.45	80.13	**55.26**	79.82
	0.8:1:3	79.67	96.80	79.06	**62.37**	78.94
	0.9:1:1.1	28.87	84.99	80.45	**23.02**	76.37
	0.9:1:1.2	49.66	91.05	79.92	**39.38**	78.54
	0.9:1:1.5	74.55	96.01	79.64	**58.86**	79.33
	0.9:1:2	85.73	97.65	79.91	**68.11**	79.81
	0.9:1:3	89.56	98.27	78.85	**70.25**	78.81
	1:1:1.1	96.95	99.24	79.41	76.93	79.39
	1:1:1.2	97.01	99.32	80.26	77.86	80.24
	1:1:1.5	96.99	99.17	79.86	77.60	79.84
	1:1:2	96.56	99.31	80.19	77.43	80.19
	1:1:3	96.24	99.17	79.78	76.62	79.77
0.7	0.8:1:1.1	14.40	68.80	79.49	**11.59**	**63.15**
	0.8:1:1.2	27.04	80.48	80.44	**21.31**	**71.84**
	0.8:1:1.5	60.33	93.24	80.03	**47.73**	78.42
	0.8:1:2	83.64	97.97	79.50	**66.19**	79.26
	0.8:1:3	94.57	99.44	79.76	75.16	79.73
	0.9:1:1.1	30.29	81.82	80.34	**24.41**	**72.66**
	0.9:1:1.2	57.10	91.80	79.67	**45.39**	77.29
	0.9:1:1.5	89.57	98.67	79.64	**71.16**	79.48
	0.9:1:2	97.70	99.78	79.78	77.91	79.75
	0.9:1:3	99.45	99.94	80.19	79.74	80.18
	1:1:1.1	99.98	100.00	80.05	80.03	80.05
	1:1:1.2	100.00	100.00	80.99	80.99	80.99
	1:1:1.5	100.00	100.00	79.85	79.85	79.85
	1:1:2	99.98	100.00	79.06	79.04	79.06
	1:1:3	99.99	100.00	79.96	79.95	79.96

Table 3

Simulated power in %. Desired power of 0.8 for the RET under optimal allocation, significance level $\alpha = 0.05$, $p_T = p_R = p_P = 0.8$, and 10 000 replications. For the complete test procedures values ***less than 75%*** *are bold.*

We summarize that the asymptotic RET yields a test for finite samples, which keeps rather accurately the nominal significance level $\alpha = 0.05$ even for small total sample sizes about 120 and all simulated parameter constellations.

5.2. Power

Table 3 presents the simulated power of the pretest for $H_{0,R}^S$ and $H_{0,T}^S$, respectively, of the RET and the power of complete test procedure, whereas either a test of superiority of the reference treatment to placebo or of the test treatment to placebo is performed in the first step and in the second step the RET. The RET is performed for different parameter constellations $\lambda_T : \lambda_R : \lambda_P$ in the alternative, and for $\Delta = 0.5, 0.7$. Beside the commonly used alternative $\lambda_T = \lambda_R$, we considered alternatives with $\lambda_T < \lambda_R$, since, as pointed out in the previous section, these are the critical parameter constellations in planning the complete test procedure. We considered homogeneous censoring probabilities $p_T = p_R = p_P = 0.8$ as before and a significance level of $\alpha = 0.05$. In each case the used sample sizes were determined according to the optimal allocation (13) to obtain a power of 80% in the RET. The results are based on 10 000 replications.

Hypothesis	p-value in %
H_0^N, $\Delta = 0.5$	1.83
H_0^N, $\Delta = 0.8$	2.51
H_0^N, $\Delta = 1$	4.42
$H_{0,R}^S$	33.34
$H_{0,T}^S$	3.88

Table 4
P-values for the RET and the pretests.

At first, one observes that for all parameter constellations the power of the RET attains exactly 80%. The simulated power for the pretest coincides with the theoretically computed one in Table 1. Moreover, the simulations approve the assumption that no sample size adjustment for the complete test procedure is necessary under the commonly used alternative $\lambda_T = \lambda_R$. In contrast, under alternatives $\lambda_T < \lambda_R$ the complete test procedure with a pretest for superiority of the reference to placebo sustains a drastically loss in power, whereas the complete test procedure with a pretest for superiority of the test treatment to placebo keeps a power of at least 70% in all cases.

6. ANALYSIS OF A THREE ARM CLINICAL TRIAL IN DEPRESSION

In this section, we revisit the example in treatment of major depression of the introduction. In this randomized, double blind study a new antidepressant (T) is compared to a standard antidepressant (R), known for having a fast onset of action, and to placebo (P). The comparison is based on the analysis of the time to first remission whereas remission is defined as maintaining the Hamilton Rating Scale of Depression (HAM-D) total score at ≤ 7 as aforementioned. The data set consists of $n_T = 262$, $n_R = 267$, and $n_P = 135$ pairs of observations, the time to first remission in days, and the censoring indicator with a fraction of 0.51, 0.46, and 0.41 uncensored observations, respectively. For the ML-estimators we obtain $\hat{\lambda}_T = 67.75$, $\hat{\lambda}_R = 83.84$, and $\hat{\lambda}_P = 89.87$. Thus, one would guess that the new antidepressant has the fastest onset of action followed by the reference and by placebo. The PP-plots in Figure 2 in the introduction indicate a quite good fit of the exponential model. Note that due to the heavy censoring at the right tail the quality of fit is decreased, of course.

The resulting p-values for the RET and the pretests are presented in Table 4. If we presume the commonly used significance level of 5%, the hypothesis of the RET could be even rejected for $\Delta = 1$ and hence not only non-inferiority but also superiority of the new treatment to the standard treatment could be claimed. The pretest with the new treatment ($H_{0,T}^S$) would reject in favor of superiority of new treatment to placebo. In contrast, the pretest with the reference treatment ($H_{0,R}^S$) would fail, i.e. it does not reject. This fact supports the view of Koch & Röhmel [11] to perform the pretest for $H_{0,T}^S$ instead of $H_{0,R}^S$.

The present sample size allocation is approximately $n_T : n_R : n_P \approx 2 : 2 : 1$. Hence, if we consider $\Delta = 0.5$, a sample size reduction of roughly 10% would have been possible by reallocating to the optimal allocation $2 : 1 : 1$ (see Figure 5).

7. SOFTWARE

The R source code of the functions for planning and analyzing the test procedures presented in this paper are available in a package at the author's web page

`http://www.stochastik.math.uni-goettingen.de/RET.zip`

This package includes the functions *ret* and *pretest* to compute the p-values for the RET and the pretest, respectively. Moreover, a function *ret_samplesize* is provided to determine the required sample sizes to obtain a desired power for the RET. The package also includes a brief instruction manual for applying these functions.

8. DISCUSSION

In this paper, we have presented a full analysis and planning of three-arm non-inferiority trials including an active control and a placebo group (gold standard design), where the observations are randomly right censored as we have assumed it for the depression study. As a byproduct we obtain also a two-arm study, which in this context also never has been considered so far. To this end we assumed exponentially distributed endpoints (time to first remission). Our analysis is based on a Wald-type test (RET) for a retention of control effect $\eta = \log \lambda_T/\lambda_P - \Delta \log \lambda_R/\lambda_P$, which shows quite good small sample characteristics. Due to the choice of scale the asymptotic variance σ^2 of the ML-estimator $\hat{\eta}$ only depends on the non-inferiority margin Δ, the probabilities of an uncensored observation p_k, $k = T, R, P$, the sample allocation, and not on the parameters λ_k, $k = T, R, P$, again. This makes estimation of σ^2 particularly simple. The probabilities p_k, $k = T, R, P$, completely characterize the censoring scheme and no further specifications of the censoring variables U_k are necessary.

There are two major findings. First, the optimal allocation of samples for the RET becomes particularly simple when the trial is planned under a worst case scenario, i.e. under homogeneous censoring probabilities in all groups with $p = \min\{p_T, p_R, p_P\}$, and does not depend on the parameters $\lambda_T, \lambda_R, \lambda_P$, and on the censoring distribution, again, viz. $n_T^* : n_R^* : n_P^* = 1 : \Delta : (1 - \Delta)$, exactly as for the Wald-type test in a trial with normal endpoints (cf. Pigeot et al. [12]). Second, independent of the censoring distribution, the power $1-\beta$ of the two-sample pretest for superiority of the test and the reference treatment, respectively, to placebo is automatically guaranteed, when the three-sample trial is planned under the commonly used alternative $\lambda_T = \lambda_R < \lambda_P$ with a sample size to keep the power $1-\beta$ for RET, provided the non-inferiority margin $1/3 \leq \Delta < 1$. In particular, the power of the pretest is independent of the considered active control effect λ_R/λ_P. Moreover, numerical investigations show that the power of the pretest is often even better and no sample size adjustment is necessary to obtain a power of $1 - \beta$ for the complete test procedure. In addition, it turns out that from a statistical point of view the pretest for superiority of the test treatment to placebo is preferred to those of the reference to placebo due to a larger power. This may be also adorable from a clinical perspective because it allows to reveal the test treatments efficacy in direct comparison with placebo instead via the indirect assessment by an additional standard.

The presented approach is based on the asymptotic normality of $\hat{\eta}$, and to our knowledge this is the first paper which shows this for censored observations. This suggests that our method can be extended to other parametric models such as Weibull or Gamma distributed endpoints, and as well as to other censoring schemes such as interval censored observations, as they may occur, e.g. in cancer studies. Certainly, the case of uncensored, exponentially distributed endpoints is covered by the case homogeneous censoring probabilities with $p = 1$.

Acknowledgements

The authors would like to thank the editor and two referees for their constructive comments which served to improve the manuscript. Matthias Mielke acknowledges financial support from the Georg Lichtenberg program "Applied Statistics & Empirical Methods".

References

[1] WMA. World medical association declaration of helsinki. ethical principles for medical research involving human subjects. http://www.wma.net/e/policy/b3.htm.

[2] B. Jones, P. Jarvis, J.A. Lewis, and A.F. Ebbutt. Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal*, 313:36–39, 1996.

[3] J. Röhmel. Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine*, 17:1703–1714, 1998.

[4] R.B. D'Agostino, J.M. Massaro, and L.M. Sullivan. Non-inferiority trials: design concepts and issues-the encounters of academic consultants in statistics. *Statistics in Medicine*, 22:169–186, 2003.

[5] S. Senn. *Statistical Issues in Drug Development*. John Wiley & Sons, 1997.

[6] ICH. ICH Harmonized Tripartite Guideline. Choice of Control Group and Related Issues in Clinical Trials (E10). *CPMP/ICH/364/96*.

[7] R.H. Dworkin, J. Katz, and M.J. Gitlin. Placebo response in clinical trials of depression and its implications for research on chronic neuropathic pain. *Neurology*, 65:7–19, 2005.

[8] B.T. Walsh, S.N. Seidman, and R. Sysko. Placebo response in studies of major depression - variable, substantial and growing. *Journal of the American Medical Association*, 287:1840–1847, 2002.

[9] M. Rothmann, N. Li, G. Chen, G. Chi, R. Temple, and H.-H. Tsou. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine*, 22:239–264, 2003.

[10] R. Temple and S. Ellenberg. Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: Ethical and scientific issues. *Annals of Internal Medicine*, 133:455–463, 2000.

[11] A. Koch and J. Röhmel. Hypothesis testing in the "gold standard" design for proving the efficacy of an experimental treatment relative to placebo and a reference. *Journal of Biopharmaceutical Statistics*, 14:315–325, 2004.

[12] I. Pigeot, J. Schäfer, J. Röhmel, and D. Hauschke. Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*, 22:883–899, 2003.

[13] M.-L. Tang and N.-S. Tang. Tests of noninferiority via rate difference for three-arm clinical trials with placebo. *Journal of Biopharmaceutical Statistics*, 14:337–347, 2004.

[14] M. Kieser and T. Friede. Planning and analysis of three-arm non-inferiority trial with binary enpoints. *Statistics in Medicine*, 26:253–273, 2007.

[15] A.A. Nierenberg and E.C. Wright. Evolution of remission as the new standard in the treatment of depression. *Journal of Clinical Psychiatry*, 60 (suppl. 22):7–11, 1999.

[16] G. Yadid, A. Zangen, A. Dmitrochenko, D.H. Overstreet, and J. Zohar. Screening for new antidepressants with fast onset and long-lasting action. *Drug Development Research*, 50:392–399, 2000.

[17] G. Freitag. Methods for assessing noninferiority with censored data. *Biometrical Journal*, 47:88–98, 2005.

[18] CPMP. Note for guideline on clinical investigation of medicinal products for the treatment of depression. CPMP/EWP/518/97 rev.1, 2002.

[19] M. Mielke, A. Munk, and A. Schacht. Technical report on "the assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints". http://www.stochastik.math.uni-goettingen.de, 2008.

[20] S. Lange and G. Freitag. Choice of delta: Requirements and reality - results of a systematic review. *Biometrical Journal*, 47:12–27, 2005.

[21] J. Hung, S.-J. Wang, and R. O'Neil. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal*, 47:28–36, 2005.

[22] T.A. Schwartz and J.S. Denne. A two-stage sample size recalculation procedure for placebo- and active-controlled non-inferiority trials. *Statistics in Medicine*, 45:3396–3406, 2006.

[23] A. Koch. Discussion on "establishing efficacy of a new experimental treatment in the "gold standard" design". *Biometrical Journal*, 47:792–793, 2005.