STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:847–858 (DOI: 10.1002/sim.1449)

Identifying effective and/or safe doses by stepwise confidence intervals for ratios

Frank Bretz[1,*,†], Ludwig A. Hothorn[1] and Jason C. Hsu[2]

[1]*Bioinformatics Unit, University of Hannover, Herrenhäuser Str. 2, D-30419 Hannover, Germany*
[2]*Department of Statistics, Ohio State University, Columbus, OH 43210, U.S.A.*

SUMMARY

Typical randomized clinical dose-finding studies consist of the comparison of several doses of a drug versus a placebo. Interest lies in estimating relevant doses among those under investigation for efficacy and safety variables, such as the minimum effective dose or the maximum safe dose (or estimating both doses simultaneously). Step-down procedures have been proposed for comparing the standardized differences of the dose groups against placebo. In this paper we consider the ratio of population means and propose stepwise confidence intervals for these ratios. These confidence intervals do not require multiplicity adjustments and yield the same decisions as the associated test procedures. In addition, several power concepts are investigated within the present framework. The results allow sample size determination in the design phase of a study for the probability of estimating correctly the dose of interest. Auxiliary results of a numerical study show the range of application of these methods. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: stepwise confidence intervals; ratios; therapeutic window; partitioning principle; multiple testing

1. INTRODUCTION

One main goal of randomized clinical dose-finding studies is to assess the dose range where a certain compound or drug exerts its effects. In efficacy studies the investigator is interested in determining the minimum dose which still reveals a clinically relevant effect. In safety studies, the interest lies in detecting those doses which are safe within a predetermined safety margin. A therapeutic window is then defined as the proper range of doses, which are both safe and efficient [1, 2].

Numerous papers exist on the identification of the minimum effective dose (MED) or the maximum safe dose (MSD) using contrasts or order restricted tests [3, 4]. Mostly they

*Correspondence to: Dr. Frank Bretz, LG Bioinformatik, Universität Hannover, Herrenhäuser Str. 2, D-30419 Hannover, Germany.
†E-mail: bretz@bioinf.uni-hannover.de

Contract/grant sponsor: Deutsche Forschungsgemeinschaft; contract/grant number: BR 2202/1.
Contract/grant sponsor: Drug Information Association.

propose stepwise procedures, which test the appropriate hypotheses in a fixed sequence, each at full level α. Such an approach is justified by the closed testing principle [5, 6]. In the setting of stepwise procedures the derivation of correct and meaningful confidence intervals is sometimes very difficult. However, it is common sense that the use of confidence intervals should be preferred to the traditional quotation of p-values. Stepwise confidence intervals yield more information about the differences of the parameters under investigation than a standard p-value approach does. In addition, once stepwise confidence intervals are available, it is straightforward to show that directional errors (that is, correct side decisions at each step) are simultaneously controlled with the familywise error rate. Hsu and Berger [7] acknowledged this fact and proposed a stepwise confidence interval approach for the successive comparison of the dose groups against a control. Their approach is based on a natural partition of the parameter space and ensures that the same decisions are made as with the corresponding tests, while providing a meaningful guarantee against incorrect decisions. As discussed later in more detail, this approach does not assume any response shape to control the overall significance level. Even non-monotone shapes with downturns at high doses are allowed. In contrast, many current procedures [3, 4] rely on specific monotonicity assumptions, such as of increasing effects with increasing doses. Violations of these assumptions may result in incorrect decisions, which are not controlled any more [8].

Tests on differences of means are routinely used. However, their interpretation in absolute terms or in relation to the standard deviation seems to be difficult and of restricted use in some scenarios [9, 10]. This paper considers the case when the relevance shifts are defined in terms of ratio of population means and the original, untransformed data are normally distributed. If the observations follow a log-normal distribution, taking the logarithms transforms the ratio problem into a problem involving differences of means and current methods can be used. However, there are many situations in clinical trials for which the normality assumption of the original variable is justified. Examples include the assessment of therapeutic equivalence for two inhalers applied for the relief of asthma attacks using the morning peak expiratory flow rate as a measure of airflow obstructions [11] and the pharmacokinetic characteristic AUC for topical dermatologic corticosteroids [12].

The purpose of the paper is to provide stepwise confidence intervals for identifying effective and/or safe doses when the percentage ratio of treatment versus placebo is of main interest. The basic procedure is described in detail for normally distributed data. Extensions to dichotomous data are discussed briefly. In a second step we investigate different power concepts for the determination of MED/MSD. The discussion involves complete and average power as well as a decision-theoretic approach. We derive exact methods for the calculation of power (and hence sample sizes) for all these approaches in the case of normally distributed data. As in the case of deriving the test procedure, extensions of the power calculations to dichotomous data based on the asymptotic normal theory are possible.

2. AN EXAMPLE OF A COMBINED EFFICACY AND SAFETY ANALYSIS

Knapp *et al.* [13] examined the efficacy and safety of Simvastatin and Colesevelam in a double-blind placebo-controlled multi-centre non-factorial combination study. Subjects with hypercholesterolaemia were randomly assigned to receive daily doses of placebo, Simvastatin 10 or 20 mg or Simvastatin 20 mg with Colesevelam 2.3 g. It can be assumed *a priori* that

Table I. Effect of different Simvastatin therapies or placebo on LDL cholesterol level and percentages of patients with adverse events after 6 weeks.

Group	Sample sizes	Means (LDL cholesterol mg/dl)	Standard deviation	Per cent of patients with adverse events
Placebo	33	177	30	71
Simvastatin 10 mg	35	136	31	61
Simvastatin 20 mg	39	119	26	59
Simvastatin 20 mg and Colesevelam 2.3 g	37	111	37	68

the combination therapy is at least as effective as the highest dose of the single compound therapy [13]. The primary efficacy variable was the reduction in serum LDL cholesterol after 42 days. For the safety analysis the percentage of total adverse events was used. The summary data are given in Table I. The question arises, which dose(s) are both efficient and safe. Note that there is a marked non-parallel effect between the efficacy and the safety variable. The LDL level decreases monotonically with increasing regimes, but the frequency of patients with adverse events is highest for both the placebo group and the combination therapy and no monotonicity assumption seems to hold for the safety variable.

Many other dose-finding trials involving the simultaneous assessment of one single efficacy and one single safety parameter can be found [14, 15]. In general, the scales of efficacy and safety endpoints are different, including continuous, ordered categorical or dichotomous data. It transpires that usually the efficacy endpoint is known in advance, but the safety endpoint is only *post hoc* defined. In addition, all examples reviewed by us considered the frequencies of patients showing main events as the safety endpoint. The response shapes were different from case to case, even a downturn could occur at higher doses.

3. STEPWISE CONFIDENCE INTERVALS

Assume that k doses $D_1, \ldots, D_k$ are tested against a placebo D_0. Let $\gamma_i = \mu_i/\mu_0$ and $\lambda_i = \tau_i/\tau_0$ be the ratios of interest, $i = 1, \ldots, k$, where the μ_i's (τ_i's) are the location parameters of the efficacy (safety) variable. Without loss of generalization we present the results for the variables showing an increasing effect for higher doses. The hypotheses of interest are then given by

$$H_i^{\mathrm{E}}\colon\ \gamma_i \leqslant \theta^{\mathrm{E}} \quad \text{versus} \quad K_i^{\mathrm{E}}\colon\ \gamma_i > \theta^{\mathrm{E}}$$

and

$$H_i^{\mathrm{S}}\colon\ \lambda_i \geqslant \theta^{\mathrm{S}} \quad \text{versus} \quad K_i^{\mathrm{S}}\colon\ \lambda_i < \theta^{\mathrm{S}}$$

with $\theta^{\mathrm{E}}, \theta^{\mathrm{S}} > 0$ being the corresponding relevance margins for efficacy and safety, respectively. Usually, θ^{E} and θ^{S} are chosen to be larger than one. This leads to a test on relevant superiority for the efficacy variable. For the safety assessment, such a choice would allow for a marginal, but clinically irrelevant, deterioration. In the complete normal set-up, we further assume that the pairs of observation (X_{ij}, Y_{ij}) are i.i. $\mathrm{N}_2((\mu_i, \tau_i)', \Sigma)$ distributed with covariance matrix Σ, $j = 1, \ldots, n_i$, $i = 0, \ldots, k$. We require the variances σ_X^2 and σ_Y^2 to be homogeneous for all

dose groups within the efficacy and the safety variables X and Y, but possibly $\sigma_X^2 \neq \sigma_Y^2$. If the covariance $\mathrm{cov}(X_{ij}, Y_{ij})$ is unknown, approximate results are obtained by using a sample estimate. Otherwise, setting $\mathrm{cov}(X_{ij}, Y_{ij}) = 0$ leads to a conservative procedure, as discussed later. If only one of the variables is normal distributed, the corresponding marginal model is assumed. The MED is commonly defined as $\min\{i\colon\ \gamma_i > \theta^{\mathrm{E}}\}$, that is, it is the smallest among the investigated doses to show a clinically relevant effect. Similarly, the MSD is given by $\max\{i\colon\ \lambda_i < \theta^{\mathrm{S}}\}$. The therapeutic window $\{i\colon\ \mathrm{MED} \leqslant i \leqslant \mathrm{MSD}\}$ is thus characterized by those doses being efficient and safe simultaneously.

Consider first the problem of determining the MED. Individual $(1-\alpha)100$ per cent confidence intervals $C_{i,1-\alpha}^{\mathrm{E}}$ for γ_i are obtained by using Fieller's [16] method. Let $\bar{X}_i$ be the arithmetic mean of the ith group, $i = 0, 1, \ldots, k$, and let s_X^2 be the joint variance estimator with $v = \sum_{i=0}^{k} n_i - k - 1$ degrees of freedom. The confidence intervals of interest are derived from the random variable

$$T_i^{\mathrm{E}} = \frac{\bar{X}_i - \theta^{\mathrm{E}}\bar{X}_0}{s_X\sqrt{\{\frac{1}{n_i} + \frac{(\theta^{\mathrm{E}})^2}{n_0}\}}} \tag{1}$$

which has a t-distribution with parameter v. The lower confidence bound is obtained as the smaller root of the quadratic equation $(T_i^{\mathrm{E}})^2 = t_{1-\alpha,v}^2$, with T_i^{E} thought of as a function in θ^{E}. Thus

$$C_{i,1-\alpha}^{\mathrm{E}} = \left(\frac{\bar{X}_i\bar{X}_0 - \sqrt{(a_0\bar{X}_i^2 + a_i\bar{X}_0^2 - a_0a_i)}}{\bar{X}_0^2 - a_0}, \infty\right)$$

where $a_i = s_X^2 t_{1-\alpha,v}^2/n_i$, $i = 1, \ldots, k$ and $a_0 = s_X^2 t_{1-\alpha,v}^2/n_0$. As discussed by Fieller [16], these confidence intervals are valid as long as $\bar{X}_0^2 > a_0$, in which case (1) and $C_{i,1-\alpha}^{\mathrm{E}}$ lead to the same test decision. The latter inequality can be regarded as a test on μ_0 being significantly larger than 0 and is a necessary condition for the ratio testing approach of Fieller. The stepwise confidence interval procedure to determine the MED then takes the following form:

Step 1: If $C_{k,1-\alpha}^{\mathrm{E}} \not\subset (\theta^{\mathrm{E}}, \infty)$ then conclude $\gamma_k \in C_{k,1-\alpha}^{\mathrm{E}}$ and stop. Otherwise conclude $\gamma_k \in (\theta^{\mathrm{E}}, \infty)$ and continue.

Step m: If $C_{k-m+1,1-\alpha}^{\mathrm{E}} \not\subset (\theta^{\mathrm{E}}, \infty)$ then conclude $\gamma_{k-m+1} \in C_{m,1-\alpha}^{\mathrm{E}}$ and stop. Otherwise conclude $\gamma_{k-m+1} \in (\theta^{\mathrm{E}}, \infty)$ and continue.

Step k: If $C_{1,1-\alpha}^{\mathrm{E}} \not\subset (\theta^{\mathrm{E}}, \infty)$ then conclude $\gamma_1 \in C_{1,1-\alpha}^{\mathrm{E}}$ and stop. Otherwise conclude $\gamma_1 \in (\theta^{\mathrm{E}}, \infty)$ and continue.

Step k+1: Conclude $\min_{i-1,\ldots,k}\gamma_i > \min_{i-1,\ldots,k}(\bar{X}_i\bar{X}_0 - \sqrt{(a_0\bar{X}_i^2 + a_i\bar{X}_0^2 - a_0a_i)})/(\bar{X}_0^2 - a_0)$

The MSD can be identified analogously for the safety endpoint by reverting the direction of the stepwise procedure and start testing H_1^{S}. Let

$$C_{i,1-\alpha}^{\mathrm{S}} = \left(-\infty, \frac{\bar{Y}_i\bar{Y}_0 + \sqrt{(b_0\bar{Y}_i^2 + b_i\bar{Y}_0^2 - b_0b_i)}}{\bar{Y}_0^2 - b_0}\right), \quad i = 1, \ldots, k$$

be the associated confidence interval for λ_i. Note that the $b_i = s_Y^2 t_{1-\alpha,\nu}^2/n_i$ here are defined in terms of s_Y. Then the general step $1 \leqslant m \leqslant k$ is given by:

Step m: If $C^{\mathrm{S}}_{m,1-\alpha} \not\subset (-\infty, \theta^{\mathrm{S}})$ then conclude $\lambda_m \in C^{\mathrm{S}}_{m,1-\alpha}$ and stop. Otherwise conclude $\lambda_m \in (-\infty, \theta^{\mathrm{S}})$ and continue.

If all k hypotheses $H_1^{\mathrm{S}}, \ldots, H_k^{\mathrm{S}}$ have been rejected, the final conclusion consists of:

Step k+1: Conclude $\max_{i=1,\ldots,k} \lambda_i < \max_{i=1,\ldots,k}(\bar{Y}_i\bar{Y}_0 + \sqrt{(b_0\bar{Y}_i^2 + b_i\bar{Y}_0^2 - b_0 b_i)})/(\bar{Y}_0^2 - b_0)$

In addition, both approaches can be combined to estimate the therapeutic window. Doses which are both effective and safe can be identified by splitting the familywise error rate $\alpha = \alpha_{\mathrm{E}} + \alpha_{\mathrm{S}}$[1]. Each of both hierarchical procedures above is then applied separately at its familywise level α_{E} and α_{S}, respectively. For each dose lying in the therapeutic window, this procedure yields simultaneous confidence intervals for both of its efficacy and safety endpoints. The remaining doses are then considered to be either effective but not safe, or to be safe but not effective. For some (possibly all) of these doses simultaneous confidence intervals for at least one of the two endpoints are provided. By taking into account correlation between efficacy and toxicity endpoints, such as using the bootstrap technique in Tamhane and Logan [2], more powerful methods may be possible.

The above procedure can be generalized in a natural way to other data conditions as long as individual $(1-\alpha)100$ per cent confidence intervals $C_{i,1-\alpha}$ for the ratio of two location parameters are available. In the case of binomial data, for example, Katz *et al.* [17] proposed Fieller-type asymptotic confidence intervals for the ratio of two success rates π_i/π_0. A simple solution is based on the test statistic $T_i = \hat{\pi}_i - \theta^{\mathrm{E}}\hat{\pi}_0$ with its estimated variance $(T_i) = \hat{\pi}_i(1-\hat{\pi}_i)/n_i + (\theta^{\mathrm{E}})^2\hat{\pi}_0(1-\hat{\pi}_0)/n_0$. Analogously to the normal case sketched above, the confidence limits are obtained as the roots of the quadratic equation $T_i^2 = z_{1-\alpha}^2 V(T)$ in θ^E, where $z_{1-\alpha}$ is the $(1-\alpha)$ standard normal quantile. Other valid confidence intervals may be used instead, also it is known that asymptotic methods in connection with binomial problems may not behave very well in general [18]. We refer to Gart and Nam [19] and Agresti and Min [20], who compared several alternatives to the Fieller-type confidence intervals and investigated their small sample size behaviour.

4. DISCUSSION AND FURTHER REMARKS ON THE PROCEDURE

A formal proof of the validity of the procedure in Section 3 is given by applying the partitioning principle [7, 21]. Thus, the familywise error is controlled at level α. Figure 1 visualizes the idea for the special case $k = 2$. Let $\Theta = \mathbb{R}^2$ be the entire parameter space, $\Theta_1 = H_2^{\mathrm{E}}$, $\Theta_2 = H_1^{\mathrm{E}} \cap K_2^{\mathrm{E}}$ and $\Theta^\delta = \{\min(\gamma_1, \gamma_2) = \theta^{\mathrm{E}} + \delta\}$, $\delta > 0$. Note that $\Theta = \Theta_1 \cup \Theta_2 \cup \bigcup_\delta \Theta^\delta$ is a partition and that the true parameter vector $\gamma = (\gamma_1, \gamma_2)^{\mathrm{T}}$ lies in one and only one of the disjoint subsets. Applying local level α tests on each of these subsets thus leads to a multiple test procedure which controls the familywise error at level α. A confidence interval for γ is then obtained by intersecting the complementary regions of those hypotheses, which have been rejected.

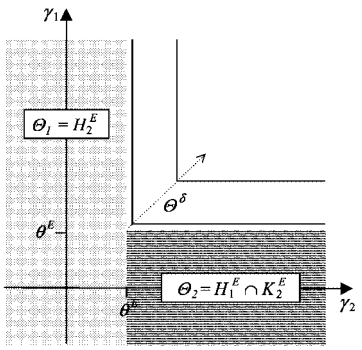

Figure 1. Graphical representation of the stepwise confidence interval procedure.

In words, for MED the null hypotheses are

$$\begin{aligned}
\tilde{H}_k^{\mathrm{E}}&: \quad \text{dose } k \text{ is ineffective}\\
\tilde{H}_{k-1}^{\mathrm{E}}&: \quad \text{dose } k \text{ is effective but dose } k-1 \text{ is ineffective}\\
&\vdots\\
\tilde{H}_i^{\mathrm{E}}&: \quad \text{doses } i+1,\ldots,k \text{ are effective but dose } i \text{ is ineffective}\\
&\vdots\\
\tilde{H}_1^{\mathrm{E}}&: \quad \text{doses } 2,\ldots,k \text{ are effective but dose 1 is ineffective}
\end{aligned}$$

The interesting thing is, in testing $\tilde{H}_j^{\mathrm{E}},\ j=1,\ldots,k$, simultaneously no multiplicity adjustment is needed to control the probability of rejecting any true null hypothesis. This is because at most one null hypothesis can be true; it cannot be the case that dose 6 is ineffective ($\tilde{H}_6^{\mathrm{E}}$) and that dose 6 is effective but dose 5 is ineffective ($\tilde{H}_5^{\mathrm{E}}$), for example. We thus test each $\tilde{H}_j^{\mathrm{E}},\ j=1,\ldots,k$, at level α. (In this form, the proof follows from the partitioning principle instead of the closed testing principle.)

The result of testing $\tilde{H}_i^{\mathrm{E}},\ i=1,\ldots,k$, simultaneously needs to be interpreted with care. For example, suppose $k=6$ and $\tilde{H}_4^{\mathrm{E}}$ and $\tilde{H}_6^{\mathrm{E}}$ are the only null hypotheses rejected, then we infer dose 6 is efficacious but beyond that the interpretation is not clear. However, for any integer i, if $\tilde{H}_j^{\mathrm{E}},\ j=i,\ldots,k$, are all rejected, then the logical inference is doses $i,\ldots,k$ are all effective: $\mu_i/\mu_0 > \theta^{\mathrm{E}},\ j=i,\ldots,k$. For example, since the union of

$$\tilde{H}_6^{\mathrm{E}}: \quad \text{dose 6 is ineffective}$$

and

$$\tilde{H}_5^{\mathrm{E}}: \quad \text{dose 6 is effective but dose 5 is ineffective}$$

is 'either dose 6 or dose 5 is not efficacious', the rejection of $\tilde{H}_5^{\mathrm{E}}$ and $\tilde{H}_6^{\mathrm{E}}$ implies 'both dose 6 and dose 5 are effective'.

A common misconception is that the validity of this step-down method depends on an assumption that the true response is monotonically non-decreasing as dose increases. Actually,

the stepwise method is valid without any assumption on the response curve. It is valid in the sense that the familywise error rate is controlled at a prespecified level α. Again, this is because the union of the null hypotheses $\tilde{H}_i^{\mathrm{E}},\ i=1,\ldots,k$, together with the part of the parameter in which all doses are efficacious, exhaust the entire parameter space of all possible response curves. However, non-monotone dose–response relationships can have a huge impact on the power. This will be demonstrated in the numerical study in Section 6.

5. ANALYSIS OF THE HYPERCHOLESTEROLAEMIA EXAMPLE

We illustrate the methods by analysing the data example of Section 2. The pooled estimate of the standard deviation for efficacy is 31. We set $\alpha_{\mathrm{E}}=\alpha_{\mathrm{S}}=0.025$, which corresponds to the usual Bonferroni approach. The efficacy endpoint is tested first. The order of testing is determined through the ordering of the treatments. We start testing the combination therapy against placebo. If the null hypothesis is rejected, the high dose of Simvastatin is tested next. If we achieve significance again, the low dose of Simvastatin is tested. The non-rejection at any step renders further testing unnecessary. The rationale is that if we fail to show efficacy for the combination therapy or the high dose group, we would not be interested in the remaining comparisons. Since the direction of a positive effect is given through decreasing values, we look at the upper confidence intervals at each step. The resulting confidence intervals are summarized in Table II. We assume *a priori* that $\theta^{\mathrm{E}}=0.9$, that is, a 10 per cent decrease of LDL cholesterol at the end of the study is assumed to be of clinical relevance. The results in Table II show that at each step the upper confidence interval $C_{i,1-\alpha}^{\mathrm{E}}$ is lower than the relevance threshold. Thus, all three comparisons are declared to be significant. According to the final step of the procedure given in Section 3, we would assess that the true maximum ratio of any of the three treatment groups to placebo lies below 84.7 per cent. Thus, either of the Simvastatin therapies reduces the serum LDL cholesterol by a clinically relevant and statistically significant amount.

For the safety variable, we assume *a priori* that $\theta^{\mathrm{S}}=1.2$. Thus, we allow a 20 per cent increase of total adverse events, which would be regarded as a clinically non-relevant increase. The order of testing is reversed and we start assessing the safety for the lowest dose of Simvastatin. The rationale is that if we fail to show safety for the low dose or the high dose group, we would not be interested in the remaining comparisons. The results in Table II show that both single treatment groups with Simvastatin are safe, while the combination therapy cannot be regarded to be safe. Thus we conclude that the true maximum ratio of the single treatment groups with Simvastatin to placebo lies below 120 per cent.

Table II. Results of the hypercholesterolaemia study.

Comparison	Upper 97.5 per cent confidence interval			
	Efficacy		Safety	
	$C_{i,1-\alpha}^{\mathrm{E}}$	Final decision	$C_{i,1-\alpha}^{\mathrm{S}}$	Final decision
Simvastatin 10 mg versus placebo	0.847	0.847	1.199	1.200
Simvastatin 20 mg versus placebo	0.744	0.847	1.156	1.200
Simvastatin/Colesevelam versus placebo	0.698	0.847	1.297	1.297

6. POWER CALCULATIONS

Frequently, the confidence interval approach for the analysis of a clinical study is not sufficient. Power consideration for adequate sample size determinations is inherent to any well designed study. Again, we consider the correct determination of the MED for normally distributed data. Formulae for the binomial case follow similarly assuming asymptotic normality. It is well known that there is no unique power concept in multiple decision situations. From a decision-theoretic point of view, we are interested in the correct estimation of the MED D_i for a fixed $i \in \{1,\ldots,k\}$. If in the design phase it is difficult to specify the correct MED among the doses, the least favourable case is obtained by setting $i=1$. The MED is correctly specified, if and only if $C^{\mathrm{E}}_{j,1-\alpha} \subset (\theta^{\mathrm{E}},\infty)$ and $C^{\mathrm{E}}_{i-1,1-\alpha} \not\subset (\theta^{\mathrm{E}},\infty)$, $j=i,\ldots,k$. Thus

$$P(\mathrm{MED}=D_i)=P\left(\bigcap_{j=i}^{k}\{T_j^{\mathrm{E}}>t_{1-\alpha,\nu}\}\cap\{T_{i-1}^{\mathrm{E}}\leqslant t_{1-\alpha,\nu}\}\right) \tag{2}$$

From the multiple hypotheses testing point of view, however, power is defined solely in terms of rejecting the incorrect null hypotheses. Expression (2) is therefore reduced to

$$P(\text{reject } H_i^{\mathrm{E}},\ldots,H_k^{\mathrm{E}})=P\left(\bigcap_{j=i}^{k}\{T_j^{\mathrm{E}}>t_{1-\alpha,\nu}\}\right) \tag{3}$$

Note that this concept is closely related to the 'all-pairs' power introduced by Ramsey [22], since all true differences of interest are to be detected. Expression (3) can be interpreted as the probability that any dose $D_1,\ldots,D_i$ is estimated to be the true MED. The stepwise procedure of Section 3 nevertheless controls the familywise error rate. Thus, the probability of incorrectly estimating any of the $D_1,\ldots,D_{i-1}$ to be the true MED is still controlled. In fact, expressions (2) and (3) will usually give very similar answers. A third power concept is given by the possible need for calculating the expected average probability to estimate the true MED. This approach can be formalized as the weighted sum of each probability to reject an element of the hypotheses' family of interest. In the present context this probability is seen to be

$$\frac{1}{k-i+1}\sum_{j=i}^{k}P(\text{reject } H_j^{\mathrm{E}})=\frac{1}{k-i+1}\sum_{j=i}^{k}P\left(\bigcap_{m=j}^{k}\{T_m^{\mathrm{E}}>t_{1-\alpha,\nu}\}\right) \tag{4}$$

The probability expressions (2)–(4) are all evaluated using existing numerical methods for the computation of non-central multivariate t-probabilities [23]. Expression (3), for example, involves a $(k-i+1)$-variate non-central t-distribution with ν degrees of freedom and correlations $\mathrm{corr}(T_{j_1}^{\mathrm{E}},T_{j_2}^{\mathrm{E}})=\lambda_{j_1}\lambda_{j_2}$, $\lambda_j=\theta^{\mathrm{E}}/\sqrt{\{(\theta^{\mathrm{E}})^2+\frac{n_0}{n_j}\}}$, $i\leqslant j_1<j_2\leqslant k$. The non-centrality parameters are given by

$$\Delta_j^{\theta^{\mathrm{E}}}=\frac{\theta_j-\theta^{\mathrm{E}}\theta_0}{\sigma\sqrt{\{\frac{1}{n_j}+\frac{(\theta^{\mathrm{E}})^2}{n_0}\}}},\quad j=i,\ldots,k$$

A product-type correlation matrix therefore holds.

In total analogy to expressions (2)–(4), similar power formulae are also derived for detecting the MSD. The computations are a little different for a therapeutic window of both safe

and effective doses D_j. Let $i_1 \leqslant j \leqslant i_2$, where $D_{i_1}(D_{i_2})$ is the MED (MSD), respectively. The probability under consideration is

$$P\left(\bigcap_{j=i_1}^{k} \{T_j^{\mathrm{E}} > t_{1-\alpha_{\mathrm{E}},\nu}\} \cap \bigcap_{j=1}^{i_2} \{T_j^{\mathrm{S}} > t_{1-\alpha_{\mathrm{S}},\nu}\} \right)$$

where T_j^{S} are the pivotal test statistics for the safety point similar to (1). Since the true correlation between both endpoints is virtually unknown, we propose to calculate the power under the conservative assumption of independence, that is, $\boldsymbol{\Sigma} = \boldsymbol{I}_2$ (the conservativeness basically follows from the Bonferroni inequality). Thus, $\mathrm{corr}(T_{j_1}^{\mathrm{E}}, T_{j_2}^{\mathrm{S}}) = 0$ and the correlation matrix is simplified to a block-diagonal structure. Under this assumption the given probability term leads to a product of a two multivariate non-central t-probabilities. Note that this assumption also allows an individual standardization for each endpoint, such that the need of a joint mean square error for both endpoints is not given.

7. NUMERICAL STUDY

We performed a numerical study in order to assess the influence of several parameters on the power to estimate the true MED (similar results hold also for the other cases and are therefore omitted). We investigated the case $k = 3$, $\sigma = 1$ and $n_i = 100$ for all i. Power computations are based on formula (3) using algorithms provided by Genz and Bretz [23]. The power values are accurate to three significant digits. Table III shows the results for different dose–response shapes, relevance margins and MEDs.

In the upper left part of the table the highest dose is the MED. Here, the power consists of the probability of rejecting at least the comparison of dose 3 versus placebo. As expected, the power increases with increasing effect (that is, larger values of δ) and the power decreases for higher clinical relevance margins (that is, larger values of θ^{E}). These results are found to be qualitatively similar for the upper right and the lower left part of the table (the MED being the middle and the lowest dose, respectively). Comparing the power for fixed δ and θ^{E} across the parts of the table, it transpires that the power diminishes drastically as the MED is one of the lower doses. This result is expected due to the general nature of stepwise testing. The lower right part of Table III shows the performance when the highest dose group shows a lower effect than the middle dose, thus leading to a non-monotone dose–response shape. The power to estimate the MED correctly is reduced substantially (compared to the case, when the high and the middle dose have similar effects). Better strategies than just testing from the highest dose on downward are discussed in the next section.

8. CONCLUSIONS

The method described in this paper gives a general way of obtaining stepwise confidence intervals for the ratio of means. In many standard situations tests on differences are not the appropriate way to investigate the data. Ratios, instead, are often easier to interpret and the natural measure of efficacy and/or safety. Trials investigating the ratio of population means are getting more and more popular. Using stepwise confidence intervals unites two important

Table III. Power results to estimate the MED correctly for different scenarios.

δ	θ^{E}				θ^{E}			
	1.05	1.1	1.15	1.2	1.05	1.1	1.15	1.2
	$\mu_0 = \mu_1 = \mu_2 = 1, \mu_3 = \mu_0 + \delta$				$\mu_0 = \mu_1 = 1, \mu_3 = \mu_2 = \mu_0 + \delta$			
0.3	0.530	0.381	0.254	0.157	0.369	0.232	0.133	0.071
0.4	0.778	0.644	0.497	0.357	0.660	0.498	0.344	0.218
0.5	0.927	0.851	0.742	0.607	0.876	0.764	0.619	0.464
0.6	0.984	0.957	0.904	0.819	0.971	0.925	0.843	0.724
0.7	0.998	0.991	0.975	0.940	0.996	0.984	0.956	0.899
	θ^{E}				θ^{E}			
	1.05	1.1	1.15	1.2	1.05	1.1	1.15	1.2
	$\mu_0 = 1, \mu_3 = \mu_2 = \mu_1 = \mu_0 + \delta$				$\mu_0 = \mu_1 = 1, \mu_2 = \mu_0 + \delta, \mu_3 = \mu_2 - 0.1$			
	0.286	0.166	0.088	0.043	0.214	0.120	0.062	0.030
	0.584	0.414	0.268	0.158	0.477	0.324	0.202	0.116
	0.837	0.703	0.544	0.385	0.749	0.601	0.445	0.303
	0.959	0.899	0.798	0.661	0.918	0.832	0.709	0.563
	0.994	0.978	0.940	0.867	0.982	0.952	0.892	0.796

aspects. First, stepwise procedures keep the price for multiplicity fairly low. Especially in dose-finding studies, the assumption of pre-ordering the hypotheses prior to the clinical trial is not painful. Second, the use of confidence intervals is a tool which often has been acknowledged to be superior to the crude citation of p-values and should therefore be preferred. The proposed methods yield stepwise confidence intervals, which all control the familywise error rate. In addition, the probability of a correct rejection of a null hypothesis in the wrong direction (directional errors) is simultaneously controlled with the familywise error rate by the present methods. However, the converse is not true, that if the null hypotheses are not well formulated, then it is not easy to tell what assumption on the parameter space is required for the familywise error rate to translate to probability of an incorrect decision [8].

It is noteworthy that two levels of prioritization exist. First, typical tests on dose response require a certain sort of monotonicity assumption of the dose response in order to control the significance level. Violations of these assumptions can inflate the error rates badly. Secondly, the hypothesis itself might be ordered *a priori*, according to one's interests and beliefs. The procedures proposed in this paper basically rely on pairwise comparisons and thus do not require the first assumption. A good prespecification of the hypotheses, however, may result in more powerful procedures. Indeed, it is not necessary to start with dose k, then $k-1,\ldots,$ in doing MED, or dose 1, then $2,\ldots,$ in doing MSD. The procedures will work for any prespecified sequence. Assume, for example, that an inverted U-shape cannot be excluded prior to the study and one has some experience with the response shape. In this case, if one expects the peak to be around dose 3 or 4, it may be more advantageous to prespecify dose 3, then dose 4, then dose 2, and so on. In fact, adequate modifications of the power formulae earlier in this paper could help to compare different sequences and sort out the most promising ones. Such an approach would tailor the ultimate design choice to one's beliefs.

ACKNOWLEDGEMENTS

The work of Frank Bretz and Ludwig Hothorn is supported by the Deutsche Forschungsgemeinschaft, grant BR 2202/1. Jason Hsu's work is partially supported by a grant from the Drug Information Association.

REFERENCES

1. Bauer P, Brannath W, Posch M. Multiple testing for identifying effective and safe treatments. *Biometrical Journal* 2001; **43**:605–616.
2. Tamhane AC, Logan B. Multiple test procedures for identifying the minimum effective and maximum safe dose of a drug. *Journal of the American Statistical Association* 2002; **97**:293–301.
3. Tamhane AC, Hochberg Y, Dunnett CW. Multiple test procedures for dose finding. *Biometrics* 1996; **52**:21–37.
4. Tamhane AC, Dunnett CW, Green JW, Wetherington JD. Multiple test procedures for identifying the maximum safe dose. *Journal of the American Statistical Association* 2001; **96**:835–843.
5. Marcus R, Peritz E, Gabriel KB. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
6. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
7. Hsu JC, Berger RL. Stepwise confidence intervals without multiplicity adjustment for dose–response and toxicity studies. *Journal of the American Statistical Association* 1999; **94**:468–482.
8. Bauer P. A note on multiple testing procedures in dose finding. *Biometrics* 1997; **53**:1125–1128.
9. Röhmel J. Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine* 1998; **17**: 1703–1714.
10. Hothorn LA, Hauschke D. Identifying the maximum safe dose: a multiple testing approach. *Journal of Biopharmaceutical Statistics* 2000; **10**:15–30.

11. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* 1996; **313**:36–39.
12. Food and Drug Administration. Guidance – topical dermatologic corticosteroids: in vivo bioequivalence. Food and Drug Administration, 1995.
13. Knapp HH, Schrott H, Ma P, Knopp R, Chin B, Gaziano JM, Donovan JM, Burke SK, Davidson MH. Efficacy and safety of combination simvastatin and colesevelam in patients with primary hypercholesterolemia. *American Journal of Medicine* 2001; **110**:352–360.
14. Fischer S, Hanefeld M, Spengler M, Boehme K, Temelkova-Kurktschiev T. European study on dose-response relationship of acarbose as a first-line drug in non-insulin-dependent diabetes mellitus: efficacy and safety of low and high doses. *Acta Diabetologica* 1998; **35**:34–40.
15. Turri M, Stein G. The determination of practically useful doses of new drugs: some methodological considerations. *Statistics in Medicine* 1986; **5**:449–457.
16. Fieller EC. Some problems in interval estimation. *Journal of the Royal Statistical Society*, *Series B* 1954; **16**:175–185.
17. Katz D, Baptista J, Azen SP, Pike MC. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 1978; **34**:469–474.
18. Brown LD, Cai TT, DasGupta, A. Interval estimation for a binomial proportion. *Statistical Sciences* 2001; **16**:101–133.
19. Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Biometrics* 1988; **44**:323–328.
20. Agresti A, Min Y. On small sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001; **57**:963–971.
21. Finner H, Strassburger K. The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics* 2002; **30**:1194–1213.
22. Ramsey PH. Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association* 1978; **73**:479–487.
23. Genz A, Bretz F. Methods for the computation of multivariate *t*-probabilities. *Journal of Computational and Graphical Statistics* 2002; **11**:950–971.

