Simultaneous confidence intervals for ratios with application to the gold standard design with more than one experimental treatment

Sylvia Schmidt | Werner Brannath

Kompetenzzentrum für Klinische Studien Bremen, Universität Bremen, Bremen, Germany

Correspondence Werner Brannath, Kompetenzzentrum für Klinische Studien Bremen, Universität Bremen, 28359 Bremen, Germany. Email: brannath@uni-bremen.de

Funding information Deutsche Forschungsgemeinschaft, Grant/Award Number: BR 3737/1-1

Considering a study design with two experimental treatments, a reference treatment and a placebo, we extend a previous approach considering the ratios of effects to a procedure for analyzing multiple ratios. The technical framework for constructing tests and compatible simultaneous confidence intervals is set in a general manner. Besides a single step procedure and its extension to a stepdown procedure, also, an informative stepwise procedure in the spirit of our previous work is developed. The latter is especially interesting, because noninferiority studies require informative confidence intervals to infer more information than just noninferiority at the prespecified margin. Results from a simulation study for the three methods are shown. We also argue that an extension to more than two experimental treatments is straightforward.

KEYWORDS
gold standard design, ratios, simultaneous confidence intervals

1 | INTRODUCTION AND NOTATIONS

Consider the gold standard design for clinical trials, which includes a new treatment (called experimental treatment), a standard treatment for the given indication (called reference) and a placebo. This design is particularly recommended if the reference effect has proven unstable in earlier trials and if the inclusion of a placebo group is ethically acceptable. To establish the efficacy of the experimental treatment, noninferiority to the reference has to be shown. Furthermore, the assay sensitivity of the clinical trial has to be established. This means that noninferiority to the reference is not sufficient, because it can already be shown if neither the reference nor the experimental treatment have any effect over placebo. Therefore, the placebo group is included in the trial.

Three-armed clinical trials have been proposed in the works of Pigeot et al and Koch and Röhmel. The first uses the ratio of the effects over placebo of the experimental and the reference treatment as parameter of interest, thus requiring the superiority of reference over placebo. The latter proposes a hierarchical strategy, testing first the superiority of the experimental treatment over placebo and then the noninferiority of the experimental treatment to the reference. Both the superiority of the reference and the superiority of the experimental treatment over placebo assure assay sensitivity.

We will in the following pursue the approach by Pigeot et al, ie, we consider the null hypothesis

\[ H_0 : \frac{\mu_E - \mu_P}{\mu_R - \mu_P} \leq \theta, \quad \theta \in \mathbb{R}, \]

where \( \mu_E, \mu_R, \mu_P \) are the effects of the experimental treatment, reference, and placebo, respectively. Rejection of \( H_0 \) means showing that the effect of the experimental treatment (compared to placebo) is at least \( \theta \) times the effect of the reference...
(compared to placebo). Such a proof of efficacy is useful, since the clinical relevant effect of the experimental treatment over placebo is frequently described as a proportion of the reference’s effect (see the work of Hauschke and Pigeot\(^1\)). Non-inferiority of the experimental treatment to the reference is shown if \(H_0\) is rejected for a suitable noninferiority margin \(g\), which is defined in the planning of the clinical trial. Normally, \(g\) is some value smaller than 1, but larger than 1/2. Note that the superiority \(\mu_R > \mu_P\) is needed as prerequisite for this approach. Due to the hierarchical order, \(H_0\) can be tested at full level after the reference’s superiority has been established.

Especially in noninferiority studies, knowledge about the quantity of the actual effect is more important than sheer rejection of the null hypothesis. Establishing a confidence interval is a useful possibility to test all \(H_0\) simultaneously and thus gain information without the need of adjusting the significance level. For example, rejection of \(H_1\) would be a proof of the superiority of the experimental treatment over the reference. The set of all \(\theta \in \mathbb{R}\), for which the corresponding null hypothesis \(H_0\) is not rejected, constitutes a natural left-sided confidence interval for the ratio \(\gamma = (\mu_E - \mu_P)/(\mu_R - \mu_P)\). In the work of Pigeot et al,\(^1\) the construction of this confidence interval is performed using Fieller’s method.\(^4\)

In this paper, we want to extend the idea of Pigeot et al\(^1\) to situations, where two experimental treatments are to be tested simultaneously in a gold standard trial. For example, one might want to examine two different doses of a new drug. Then, an adjustment of the multiple type I error is needed. Since we are particularly interested in confidence intervals, our goal is to construct simultaneous confidence intervals (SCIs) for

\[
\gamma_1 = \frac{\mu_{E_1} - \mu_P}{\mu_R - \mu_P} \quad \text{and} \quad \gamma_2 = \frac{\mu_{E_2} - \mu_P}{\mu_R - \mu_P},
\]

where \(\mu_{E_i}\) is the effect of treatment \(i\) for \(i = 1, 2\). We therefore construct multiple test procedures for the hypotheses \(H_{1,\theta_i} : \gamma_i = \theta_i\) versus the alternatives \(A_{1,\theta_i} : \gamma_i > \theta_i, i = 1, 2\). For the sake of lucidity, we will restrict ourselves to the case of two experimental treatments, although a generalization to more than two treatments is straightforward (see Remark 2 as follows). The construction of SCIs for multiple ratios has been considered in the work of Dilba et al\(^6\) for the special case of the comparison of several treatments with a control, and in the work of Dilba et al\(^6\) for the general case.

We assume normally distributed observations and denote by \(X_s\), the observed mean effect in group \(s\) for \(s \in \{E_1, E_2, P, R\}\). Variance homogeneity within the four groups is assumed. A natural test statistic for \(H_{1,\theta_i}\) is

\[
T_{1,\theta_i} = \frac{X_{E_i} - X_P - \theta_i(X_R - X_P)}{S\sqrt{v_i(\theta_i)}},
\]

where \(S\) is the pooled variance estimator from the four groups and \(v_i(\theta_i)\) is the variance of the numerator of \(T_{1,\theta_i}\) divided by the variance of each single observation. Explicitly,

\[
v_i(\theta_i) = \frac{1}{n_{E_i}} + \frac{\theta_i^2}{n_R} + \frac{(1 - \theta_i)^2}{n_P}, \quad i = 1, 2,
\]

where \(n_s\) is the number of observations in group \(s\). Under \(H_{1,\theta_i}\), the test statistic \(T_{1,\theta_i}\) has a \(t\)-distribution with \(\nu = n_{E_i} + n_R + n_P - 4\) degrees of freedom.

The SCI estimation for \(\gamma_1\) and \(\gamma_2\) is complicated by the fact that the correlation of \(T_{1,\theta_i}\) and \(T_{2,\theta_i}\) depends on the unknown parameters. There are several proposals to replace this correlation by an observed quantity; see the work of Dilba et al\(^6\) and Section 4 for details.

It is well-known that stepwise procedures are more powerful than single-step procedures.\(^7\) Hence, if the major interest of the clinical trial consists in establishing noninferiority of preferably both new treatments, then a stepwise procedure might be preferred. However, in noninferiority studies, confidence intervals are normally indispensable, and it is not trivial to obtain SCIs with nice properties, which are compatible with the rejection decisions of a stepwise multiple test procedure. To our knowledge, all existing SCIs for ratios are derived in a single step procedure, ie, the confidence bound for \(\gamma_i\) does not depend on \(X_{E_i}\) with \(i \neq j\). Only the work of Bretz et al\(^9\) considered stepwise SCIs for ratios; however, they restrict themselves to dose-finding studies, where a natural hierarchical ordering is present. Besides, the hierarchical SCIs provide information only if all hypotheses are rejected. The latter point might be solved by applying the SCIs of Schmidt and Brannath.\(^9\) However, in our application to the gold standard design with two experimental treatments, there is no clear hierarchy of testing; therefore, we need another approach. Strassburger and Bretz\(^10\) proposed a construction of SCIs for a large class of closed testing procedures. The drawback of these SCIs is that the confidence interval is often equal to the
alternative of the null hypothesis if it is rejected, thus giving no information on the quantity of the effect. To overcome the problem of such noninformative rejections, Brannath and Schmidt\textsuperscript{11} derived SCIs, which constitute a flexible compromise between the powerful stepwise SCIs and the informative single-step SCIs. In this paper, we give a general approach to obtain SCIs, which is based on the partitioning lemma.\textsuperscript{12} It includes the standard single-step SCI, a stepwise extension, and a weighted approach inspired by Brannath and Schmidt.\textsuperscript{11} We will compare these three alternative constructions by simulations in the context of a gold standard trial with two experimental treatments.

Our article is arranged as follows. In Section 2, we explain in a very general manner how to construct SCIs for two parameters. We present stepdown SCIs similar to the work of Strassburger and Bretz\textsuperscript{10} and weighted SCIs in the spirit of Brannath and Schmidt.\textsuperscript{11} In Section 3, we construct SCIs for the gold standard design with two experimental treatments giving concrete formulas for the confidence bounds. We extend the approach of Dilba et al\textsuperscript{8} who constructed single-step SCIs for ratios of linear combinations of treatment means, to stepwise tests and compatible SCIs. In Section 4, we analyze the three methods by simulations in clinical trials with different settings. In Section 5, we discuss our results and mention possible extensions of the methods presented here. In particular, we explain how our approach is generalized to more than two treatments.

## 2 SINGLE-STEP AND STEPWISE SCIS

We consider the general situation of testing two hypotheses $H_i : \gamma_i \leq g$, $i = 1, 2$, against their alternatives $A_i : \gamma_i > g$ at significance level $\alpha \in (0, 1)$, via a SCI $(L_1, \infty) \times (L_2, \infty)$ with coverage probability $1 - \alpha$. That is, we reject $H_i$ if and only if $L_i \geq g$. Note that we choose for simplicity the same margin $g$ for both hypotheses. If there is a practical reason to use different margins $g_1$ and $g_2$, this can be realized by obvious replacements of $g$ by $g_1$ or $g_2$ in the formulas as follows.

By the duality of tests and confidence sets,\textsuperscript{13} we can construct individual tests $\phi(\theta) \in \{0, 1\}$ with significance level $\alpha$ for every $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$. More precisely, $\phi(\theta) = 1$ means that $H_{1,\theta} \cap H_{2,\theta}$ is rejected (with $H_{1,\theta} : \gamma_1 = \theta_1$). If $P_\theta(\phi(\theta) = 1) \leq \alpha$ for all $\theta$, then $C = \{\theta \in \mathbb{R}^2 : \phi(\theta) = 0\}$ is a $(1 - \alpha)$-confidence set for $(\gamma_1, \gamma_2)$. A SCI with coverage probability not smaller than $1 - \alpha$ is given by the smallest rectangle $(L_1, \infty) \times (L_2, \infty)$ containing $C$. Frequently, the tests $\phi(\theta)$ are constructed in a way that $C$ already forms a rectangle and thus is equal to the SCI.

To define the tests $\phi(\theta)$ in the following derivation of the single-step, stepwise, and weighted SCI, we need the one-dimensional tests for $H_{1,\theta}$ at any significance level $\delta \in (0, 1)$, which we denote by $\phi_i^\alpha(\theta_i)$. It is natural to assume that every $\phi_i^\alpha$ is monotonically decreasing in $\theta_i$ and that $\phi_i^{\alpha/2} \leq \phi_i^\alpha$, i.e., a parameter value that is rejected at level $\alpha/2$, is rejected at level $\alpha$ as well.

### Single-step SCI

We choose individual tests of the form

$$\phi(\theta) = \max \left\{ \phi_1^{\alpha/2}(\theta_1), \phi_2^{\alpha/2}(\theta_2) \right\}.$$ 

Then, $\phi(\theta)$ is an $\alpha$-test by the Bonferroni inequality. We obtain $C = (L_1, \infty) \times (L_2, \infty)$, where $L_i = \inf\{\theta_i \in \mathbb{R} : \phi_i^{\alpha/2}(\theta_i) = 0\}$. The confidence bound for $\gamma_i$ does not depend on the confidence bound for $\gamma_j$ with $j \neq i$; hence, we obtain a single-step SCI. We also note that this SCI does not depend on the value of the noninferiority margin $g$, which is an important difference to the stepwise SCI and weighted SCI considered next.

### Stepwise SCI

In order to improve the probabilities to reject $H_1$ and $H_2$, we partition the parameter space $\mathbb{R}^2$ in four disjoint regions and consider different individual tests in each region. Let $\mathbb{R}_I$ be the region of parameters, for which $I$ is the index set of true hypotheses. More precisely,

$$\mathbb{R}_{\{1,2\}} = H_1 \cap H_2 = \{\theta \in \mathbb{R}^2 : \theta_1 \leq g, \theta_2 \leq g\},$$

$$\mathbb{R}_{\{1\}} = H_1 \cap A_2 = \{\theta \in \mathbb{R}^2 : \theta_1 \leq g, \theta_2 > g\},$$

$$\mathbb{R}_{\{2\}} = A_1 \cap H_2 = \{\theta \in \mathbb{R}^2 : \theta_1 > g, \theta_2 \leq g\},$$

$$\mathbb{R}_{\emptyset} = A_1 \cap A_2 = \{\theta \in \mathbb{R}^2 : \theta_1 > g, \theta_2 > g\}.$$
The stepwise SCI is determined by different individual \( \alpha \)-tests \( \phi_I \) in each region \( \mathbb{R}_I \), which are given by

\[
\phi_{(1)}(\theta) = \varphi_{1}^*(\theta), \\
\phi_{(2)}(\theta) = \varphi_{2}^*(\theta), \\
\phi_{(1,2)}(\theta) = \phi_\emptyset(\theta) = \max\{ \varphi_{1}^*(\theta_1), \varphi_{2}^*(\theta_2) \}.
\]

By our choice of \( \phi_{(1)} \) and \( \phi_{(2)} \), we put more emphasis on the relevant null hypotheses and increase the power to reject them. In the regions \( \mathbb{R}_{1,2} \) and \( \mathbb{R}_\emptyset \), where both \( \theta_1 \) and \( \theta_2 \) belong either to the null or to the alternative hypothesis, equal weight is given to both parameters. Note that \( H_i \) is rejected if and only if \( \phi_{(i)}(g) = 1 \) and \( \phi_{(1,2)}(g) = 1 \).

For \( i = 1, 2 \), denote by \( L^i_1 \) the confidence bound in region \( \mathbb{R}_I \). Formally,

\[
L^i_1 = \inf \{ x \in \mathbb{R} : \min_{\theta \in \mathbb{R}_I \setminus \{ x \}} \phi_I(\theta) = 0 \}, \quad \text{where} \quad \inf \emptyset = \infty,
\]

ie, the bound \( L^i_1 \) excludes those values \( \theta_i \), where all \( \theta \in \mathbb{R}_I \) with this component value are rejected by \( \phi_I \). The confidence set is now a union of all confidence sets within the four regions, ie, the final bound \( L_1 \) is the lowest of the four \( L^i_1 \) (see Figure 1).

Due to the monotonicity properties of the one-dimensional tests \( \varphi_i^* \), we can explicitly determine the SCI \( (L_1, \infty) \times (L_2, \infty) \). The bound \( L_1 \) is explicitly given as

\[
L_1 = \begin{cases} 
L^{(1,2)}_1 & \text{in Case 1,} \\
L^{(1)}_1 & \text{in Case 2,} \\
g & \text{in Case 3,} \\
\max\{L^{\emptyset}_1, g\} & \text{in Case 4.}
\end{cases}
\]
We first describe the construction of a one-dimensional confidence interval with coverage probability 1 − \(\alpha\) for the parameters \(\gamma_1\) and \(\gamma_2\) in (1), i.e., for the ratio of the effects of two experimental treatments (compared to placebo) relative to the effect of the reference (compared to placebo). In general, confidence intervals for ratios can be obtained by Fieller’s method.\(^4\) As we have mentioned before, the test statistic in (2) has, under \(H_{i,\theta_i}\), a \(t\)-distribution with \(v = n_{E_i} + n_{E_i} + n_R + n_P - 4\) degrees of freedom.

## One-dimensional confidence interval

We first describe the construction of a one-dimensional confidence interval with coverage probability 1 − \(\delta\). It can be found in the work of Dilba et al.\(^6\) for general ratios and in the work of Pigeot et al.\(^1\) for the given situation. By the duality of tests and confidence intervals, a one-sided one-dimensional confidence set with coverage probability 1 − \(\delta\) for \(\gamma_i\) is given as

\[
C_i = \{ \theta \in \mathbb{R} : T_{i,\theta_i} \geq g_{i,1-\delta} \},
\]
where \( q_{t,1-\delta} \) denotes the \((1-\delta)\)-quantile of the \( t \)-distribution with \( \nu \) degrees of freedom. Because \( T_{i,\theta} \) is not monotone in \( \theta_i \), we need the following assumption, which states that the reference is significantly superior (at significance level \( \alpha/2 \)) to the placebo.

**Assumption 1** (Reference superiority).

\[
\frac{X_R - X_P}{\sqrt{1/n_R + 1/n_P}} \geq q_{t,1-\alpha/2}.
\]

It was stated in the work of Dilba et al\(^6\) that, under Assumption 1, the \((1-\delta)\)-confidence set (where \( \delta \geq \alpha/2 \)) is an interval \( C_i = (\ell_i(\theta), \infty) \), where \( \ell_i(\theta) \) is the solution \( \theta_i \) of the equation \( T_{i,\theta} = q_{t,1-\delta} \). An explicit formula of \( \ell_i(\delta) \) is given in the Supporting Information (formula (1.3) there). Lower and upper bounds can also be calculated by the R package *mratios*. If Assumption 1 is not satisfied, then the confidence interval might be a union of two disjoint open intervals; see the discussion at the end of Section 1 in the Supporting Information.

In the following, we use the one-dimensional tests \( \varphi_i^\delta(\theta_i) \) to describe the single-step, stepdown, and weighted SCI for \( \gamma_1 \) and \( \gamma_2 \). In terms of Section 2, these tests have the form

\[
\varphi_i^\delta(\theta_i) = \begin{cases} 
1 & \text{if } \theta_i \leq \ell_i(\delta) \ (\Leftrightarrow T_{i,\theta_i} \geq q_{t,1-\delta}) \\
0 & \text{otherwise.}
\end{cases}
\]

We note that our assumptions of Section 2 are satisfied, namely, \( \varphi_i^\delta(\theta_i) \) is decreasing in \( \theta_i \) and \( \varphi_i^{\alpha/2} \leq \varphi_i^\alpha \), because under Assumption 1, the function \( \ell_i(\delta) \) is increasing in \( \delta \) for \( \delta \geq \alpha/2 \) (see the Supporting Information for an illustration).

**Single-step SCI**

The single-step SCI with coverage probability \((1-\alpha)\) is the set of all \( \theta \in \mathbb{R}^2 \) that are not rejected by the multiple test \( \phi(\theta) = \max\{\varphi_i^{\alpha/2}(\theta_1), \varphi_i^{\alpha/2}(\theta_2)\} \). This simply leads to the single-step SCI \((L_1, \infty) \times (L_2, \infty)\) with \( L_i = \ell_i(\alpha/2) \) for \( i = 1, 2 \).

**Remark 1.** Instead of using Bonferroni’s inequality, it would of course be more efficient to work with the joint distribution of \( T_{1,\theta} \) and \( T_{2,\theta} \). As is pointed out in the work of Dilba et al\(^6\) the problem is here that the correlation of the two test statistics depends on the unknown parameter \( \theta \). Less conservative methods than Bonferroni are proposed to replace this correlation by an observed quantity, such that the coverage probability is conserved exactly or approximately. For better comparison with the other methods introduced here, we will only describe the Bonferroni approach in detail. However, in our simulations in Section 4, we will compare our SCIs also with the single-step SCIs obtained by two alternative approaches explained in the work of Dilba et al\(^6\).

**Stepdown SCI**

The stepdown SCI for \( \gamma_1 \) and \( \gamma_2 \) is obtained by inserting the one-dimensional tests (8) into formula (5). Now, the four cases (4) are described by conditions for the one-dimensional confidence bounds \( \ell_1(\delta) \) and \( \ell_2(\delta) \) for \( \delta \in \{\alpha/2, \alpha\} \). More precisely, we have

- **Case 1:** \( \ell_1(\alpha/2) < g \) and \( \ell_2(\alpha/2) < g \) (both null hypotheses accepted).
- **Case 2:** \( \ell_1(\alpha) < g \) and \( \ell_2(\alpha/2) \geq g \) (only \( H_2 \) rejected).
- **Case 3:** \( \ell_1(\alpha/2) \geq g \) and \( \ell_2(\alpha) < g \) (only \( H_1 \) rejected).
- **Case 4:** \( \ell_1(\alpha) \geq g \) and \( \ell_2(\alpha) \geq g \) and \( \max\{\ell_1(\alpha/2), \ell_2(\alpha/2)\} \geq g \) (both null hypotheses rejected).

These cases can also be described by the test statistics, because \( \ell_i(\delta) < g \) is equivalent to the corresponding relation \( T_{ig} < q_{t,1-\delta} \). Figure 3 illustrates the four cases within the dimensions of \( T_{1g} \) and \( T_{2g} \). The resulting expression for the
confidence bounds is

\[ L_1 = \begin{cases} \ell_1(a/2) & \text{in Case 1,} \\ \ell_1(a) & \text{in Case 2,} \\ g & \text{in Case 3,} \\ \max\{\ell_1(a/2), g\} & \text{in Case 4,} \end{cases} \]  

(9)

and, respectively, \( L_2 \) by interchanging the indices 1 and 2.

**Weighted SCI**

We now consider the approach (6), where the individual tests use different weights for different hypotheses \( H_{i,\theta} \). These weights depend continuously on the parameter \( \theta \) via the function given in (7).

The procedure to obtain the weighted SCI is following the ideas of Brannath and Schmidt. We find a suitable one-dimensional set \( M \subseteq \mathbb{R}^2 \) and functions \( f_i \) (for \( i = 1, 2 \)) that are continuous and increasing in \( \theta_i \) on \( M \) with \( \inf_M f_i \leq a \leq \sup_M f_i \). Furthermore, we show that the simultaneous confidence bound \( L = (L_1, L_2) \) satisfies \( L \in M \) and \( f_1(L) = f_2(L) = a \). Then, we can construct an easy bisection search to determine the weighted SCI. We explain this approach in more detail in the Supporting Information.

Due to the nonmonotonicity of the test statistics \( T_{i,\theta} \), we cannot directly apply the approach of Brannath and Schmidt, which supposes monotone p-values. Therefore, we need to restrict the search for the confidence bounds on a two-dimensional interval, where the test statistics are decreasing. Since the derivation of such an interval is rather technical, we defer it to the Appendix. We only mention that the following restriction for the sample sizes is required.

**Assumption 2** (Sample sizes). Let \( g \in (0, 1) \) be the noninferiority margin for the relative effect. Then, we assume

\[ \frac{n_F}{n_R} \geq \frac{1 - g}{g}. \]

The best strategy to satisfy Assumption 2 is to choose equality here, because for ethical reasons, one wants not more patients in the placebo group than necessary. Actually, Pigeot et al suggested exactly this ratio as optimal allocation quotient between \( n_F \) and \( n_R \). Hence, Assumption 2 is not seriously restrictive.

**4 SIMULATION STUDY**

We conducted a simulation for a study with two experimental treatments, a reference and a placebo. The R programs for this study are available from the corresponding author upon request. We set the standard deviation of all single observations to 1. The effects of the placebo and reference were 0 and 1, respectively. The noninferiority margin \( g \) was set to 0.8. Treatment \( E_1 \) was assumed to be as effective as the reference (effect 1), while we varied the effect of treatment \( E_2 \) from noninferior to the reference (effect 0.8) to superior to the reference (effect 1.4).
We chose the one-sided significance level $\alpha = 0.025$. Since we have multiple hypotheses, we used $\alpha/2$ for the sample size calculation to obtain sufficient power for the single-step procedure with Bonferroni correction. We calculated the sample size for the experimental treatments according to formula (26) of the work of Pigeot et al$^1$ such that noninferiority can be shown with probability $1 - \beta = 0.8$ if their actual effect is 1 and the reference effect is 1 as well:

$$n_{E_1} = n_{E_2} = \frac{2}{(1 - g)^2} \left( z_{1 - \alpha/2} + z_{1 - \beta} \right)^2 = 476,$$

where $z_{\delta}$ is the $\delta$-quantile of the standard normal distribution. According to formulas (27) and (28) of the work of Pigeot et al$^1$ and to assure that Assumption 2 is satisfied, we further put

$$n_R = gn_{E_1} = 381 \quad \text{and} \quad n_F = (1 - g)n_{E_1} = \frac{1-g}{g}n_R = 96.$$

These sample sizes led to an effective power of 1 to show superiority of the reference over placebo, ie, Assumption 1 was satisfied in all 10 000 simulation runs.

We consider the three procedures with Bonferroni correction that we have described in the previous section (single-step, stepdown, and weighted procedure with exponential weight function). Dilba et al$^7$ have proposed alternative single-step procedures involving the multivariate distribution of $T_1$ and $T_2$, where the unknown correlation is estimated in different ways. The resulting procedure works like the single-step test with Bonferroni correction, but with different quantiles than $t_{1,1 - \alpha/2}$ as critical points. The “MtI” approach uses the quantiles of the multivariate $t_\nu$-distribution, where the identity matrix is taken as correlation matrix. The “plug-in” method replaces the unknown correlation between $T_1$ and $T_2$ by an estimate obtained from the data. It guarantees the desired coverage probability only asymptotically. For comparison, we show results also for these two alternatives of the single-step procedure and an according adaption of the stepdown procedure.

Results for treatment $E_1$ if $\mu_{E_2} = 1$

Table 1 shows the simulation results for the case, where both experimental treatments are as effective as the reference. Results for $E_1$ are indicated; of course, the results for $E_2$ are very similar. Coverage is good for all procedures and slightly smaller for the plug-in approaches as expected. The stepdown procedure can show noninferiority in more cases than the other procedures; however, the confidence bound then equals often exactly the noninferiority margin $g$. Hence, the probability for what we call informative noninferiority, ie, a confidence bound indicating a positive effect above the noninferiority margin, is even lower than for the single-step procedures. The weighted approach does not have this problem and exhibits a marginal power gain compared to the single-step procedure. The median confidence bounds (and also the means) are quite similar for all methods.

Results for treatments $E_1$ and $E_2$ in dependence of $\mu_{E_2}$

Tables 2 and 3 show the probabilities to prove noninferiority, informative noninferiority, and superiority for the second treatment, where the true effect $\mu_2$ of $E_2$ is varying. The weighted method is compared to the Bonferroni-corrected single-step and stepdown procedure. Only the latter has different results for noninferiority and informative noninferiority. While the power to show noninferiority is the largest for the stepdown approach for all values of $\mu_2$, its probability to obtain informative noninferiority is much smaller. Accordingly, superiority can be shown in less cases, even if the effect of $E_2$ is very large. The weighted intervals yield uniformly larger probabilities to obtain informative noninferiority than

| Coverage | Noninferiority | Informative NI | Median |
|----------|----------------|----------------|--------|
| L₁ ≥ g   | L₁ > g         | L₁ > g         |        |
| Single-Step Bonf. | 97.9 | 78.8 | 78.8 | 0.853 |
| Single-Step MtI | 97.9 | 78.9 | 78.9 | 0.853 |
| Single-Step Plug | 97.6 | 79.8 | 79.8 | 0.855 |
| Steadown Bonf. | 97.9 | 83.9 | 72.3 | 0.849 |
| Steadown MtI | 97.9 | 83.9 | 72.4 | 0.849 |
| Steadown Plug | 97.7 | 84.2 | 73.1 | 0.851 |
| Weighted Bonf. | 97.9 | 79.0 | 79.0 | 0.853 |

### Table 1

Results of the simulation study with two experimental treatments $E_1$ and $E_2$, reference $R$ and placebo $P$. Simultaneous confidence intervals are calculated with the three methods described in Section 3 and different methods for estimating the correlation between the test statistics. The table shows coverage, power for noninferiority (NI) and informative NI, and median for $L₁ := L_{E₁,R}$ in the case that $\mu_{E_1} = \mu_{E_2} = \mu_R = 1$. 
TABLE 2 Results of the simulation study with two experimental treatments $E_1$ and $E_2$, reference $R$ and placebo $P$. Simultaneous confidence intervals are calculated with the three methods described in Section 3. The table shows the power for noninferiority (NI) and informative NI (Inf. NI) for $L_2 := L_{E_j - R}$ in the case that $\mu_{E_1} = \mu_R = 1$ and $\mu_{E_2}$ is varying. Note that, for the weighted and the single-step method, the probabilities for NI ($L_2 \geq g$) and for Inf. NI ($L_2 > g$) coincide.

| $\mu_{E_2}$ | Weighted NI | Single-Step NI | Stepdown NI |
|-------------|-------------|---------------|-------------|
| 0.8         | 1.2         | 1.1           | 2.4         | 1.1 |
| 0.9         | 24.4        | 24.0          | 32.6        | 23.4 |
| 1.0         | 79.7        | 79.6          | 84.1        | 72.9 |
| 1.1         | 99.0        | 99.2          | 99.2        | 86.4 |
| 1.2         | 100.0       | 100.0         | 100.0       | 86.4 |

TABLE 3 Setting as for Table 2. The table shows the power for superiority (Sup.: $L_2 \geq 1$) for $L_2 := L_{E_j - R}$.

| $\mu_{E_2}$ | Weighted Sup. | Single-Step Sup. | Stepdown Sup. |
|-------------|---------------|------------------|---------------|
| 1.1         | 20.5          | 21.1             | 20.6          |
| 1.2         | 73.1          | 74.1             | 68.2          |
| 1.3         | 98.1          | 98.3             | 85.6          |
| 1.4         | 100.0         | 100.0            | 86.4          |

FIGURE 4 Setting as for Table 2. The figure compares the three approaches for the simultaneous confidence bound $L_1 := L_{E_1 - R}$ with respect to their power to show noninferiority ($L_1 \geq g$) and their mean $L_1$.

The gold standard design is very relevant for practical applications and the question of multiple adjustment if more than one experimental treatment is involved arises naturally. In this paper, we gave a first insight on how SCIs can be...
constructed in this case. We extended the existing single-step procedure for ratios, first, to a more powerful method in the spirit of Strassburger and Bretz\textsuperscript{11} and second, to a more powerful and informative method in the spirit of Brannath and Schmidt.\textsuperscript{11}

As a theoretical novelty, we have extended the Fieller confidence intervals for multiple ratios, as proposed in the work of Dilba et al\textsuperscript{6} in a very general manner to a stepdown procedure. Since the corresponding SCIs of Strassburger and Bretz\textsuperscript{10} are often noninformative, which is rarely acceptable in noninferiority trials, we have investigated an alternative construction, which is based on a more flexible weight function. The new approaches were analyzed for the gold standard design with two experimental treatments. As the following remark indicates, a generalization to more than two experimental treatments is straightforward.

Remark 2. Assume that we have \( m \) experimental treatments and thus need to construct an SCI for the \( m \)-dimensional parameter \( \gamma = (\gamma_1, \ldots, \gamma_m) \) with \( \gamma_i \) as in (1). If we replace the quantile \( q_{i, 1-a/m} \) in Assumption 1 by \( q_{i, 1-a/\gamma_i} \), then all three procedures considered here can easily be applied to this general case.

The single-step interval construction uses for \( i = 1, \ldots, m \) the one-dimensional intervals \( C_i = (L_i, \infty) \), where \( L_i = \mathcal{C} \delta \) with \( \delta = a/\gamma_i \) is the solution \( \theta_i \) of the equation \( T_{i, \theta_i} = q_{i, 1-\delta} \).

The stepdown interval construction applies the partitioning principle to the Bonferroni-Holm test for the hypotheses \( H_i : \gamma_i \leq \gamma \), \( i = 1, \ldots, m \). This leads to the generalization of (9), which is given by formula (10) of the work of Strassburger and Bretz.\textsuperscript{10} More precisely,

\[
L_i = \begin{cases} 
\mathcal{C}(\alpha/|R|) & \text{if } i \notin R, \\
g & \text{if } i \in R, R \neq \{1, \ldots, m\}, \\
\max\{\mathcal{C}(\alpha/m), g\} & \text{if } R = \{1, \ldots, m\},
\end{cases}
\]

where \( R \) is the set of rejected hypotheses.

The weighted interval construction generalizes the weight function (7) as follows. Let \( E(x) \) be any continuous and decreasing function on \( \mathbb{R} \) with \( E(x) = 1 \) for \( x \leq \gamma \) and \( E(x) \to 0 \) as \( x \to \infty \). Then,

\[
w_i(\theta) = \frac{E(\theta_i)}{E(\theta_1) + \ldots + E(\theta_m)}, \quad \theta = (\theta_1, \ldots, \theta_m).
\]

Each \( \theta \in \mathbb{R}^m \) is rejected if and only if any of the \( m \) individual hypotheses \( H_i, \theta : \gamma_i = \theta_i \) is rejected at level \( w_i(\theta) \). The weighted SCI is the smallest rectangle containing all \( \theta \) that are not rejected. A bisection search to explicitly find the confidence bounds is derived with the same ideas as explained above and in the Appendix. Since the intervals where the test statistics are monotone are constructed individually for each \( i = 1, \ldots, m \), the generalization of this approach is straightforward.

We have seen that the stepdown procedure considerably increases power to reject the noninferiority hypotheses, but the confidence bound itself is often useless because it gives no more information on the quantity of the effects. The weighted procedure performs well and shows marginal improvement of the single-step approach in our application. The weight function (7) has a very general form and depends on the choice of the continuous decreasing function \( E(x) \). In this paper, we work with an exponential function. We also made simulations with different choices of \( E(x) \); however, the results were substantially the same.

Further applications and scenarios could be considered to make full use of the theoretical strength of the weighted approach. In addition, this procedure might be extended by incorporating the common distribution of the test statistics, as for the MtI- and plug-in-options in the single-step and stepdown case. An extension to settings with more than one reference is possible as well.

ACKNOWLEDGEMENT

This research was supported by the DFG grant BR 3737/1-1.

DATA AVAILABILITY STATEMENT

The R programs for the simulation study are available from the corresponding author upon request.
REFERENCES

1. Pigeot I, Schäfer J, Röhmel J, Hauschke D. Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statist Med.* 2003;22(6):883-899.

2. Koch A, Röhmel J. Hypothesis testing in the “gold standard” design for proving the efficacy of an experimental treatment relative to placebo and a reference. *J Biopharm Stat.* 2004;14(2):315-325. https://doi.org/10.1081/BIP-120037182

3. Hauschke D, Pigeot I. Establishing efficacy of a new experimental treatment in the “gold standard” design. *Biom J.* 2005;47(6):782-786. https://doi.org/10.1002/bimj.200510169

4. Fieller EC. Some problems in interval estimation. *J R Stat Soc Ser B.* 1954;16(2):175-185.

5. Dilba G, Bretz F, Guiard V, Hothorn LA. Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. *Methods Inf Med.* 2004;43(5):465-469.

6. Dilba G, Bretz F, Guiard V. Simultaneous confidence sets and confidence intervals for multiple ratios. *J Stat Plan Inference.* 2006;136(8):2640-2658. https://doi.org/10.1016/j.jspi.2004.11.009

7. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63(3):655-660.

8. Bretz F, Hothorn LA, Hsu JC. Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Statist Med.* 2003;22(6):847-858.

9. Strassburger K, Bretz F, Hochberg Y. Compatible confidence intervals for intersection union tests involving two hypotheses. In: Benjamini Y, Bretz F, Sarkar S, eds. *Recent Developments in Multiple Comparison Procedures.* Beachwood, OH: Institute of Mathematical Statistics. 2004:129-142. https://doi.org/10.1214/lmms/1196285631

10. Lehmann E. *Testing Statistical Hypotheses.* 2nd ed. New York, NY: Wiley; 1986.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Schmidt S, Brannath W. Simultaneous confidence intervals for ratios with application to the gold standard design with more than one experimental treatment. *Statistics in Medicine.* 2019;38:5350–5360. https://doi.org/10.1002/sim.8365