Biometrika (2013), *pp.* 1–12
Printed in Great Britain
doi: 10.1093/biomet/ast035

Simultaneous confidence intervals that are compatible with closed testing in adaptive designs

BY D. MAGIRR, T. JAKI,

Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K.
d.magirr@lancaster.ac.uk t.jaki@lancaster.ac.uk

M. POSCH AND F. KLINGLMUELLER

Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, 1090 Wien, Austria
martin.posch@meduniwien.ac.at florian.klinglmueller@meduniwien.ac.at

SUMMARY

We describe a general method for finding a confidence region for a parameter vector that is compatible with the decisions of a two-stage closed test procedure in an adaptive experiment. The closed test procedure is characterized by the fact that rejection or nonrejection of a null hypothesis may depend on the decisions for other hypotheses and the compatible confidence region will, in general, have a complex, nonrectangular shape. We find the smallest cross-product of simultaneous confidence intervals containing the region and provide computational shortcuts for calculating the lower bounds on parameters corresponding to the rejected null hypotheses. We illustrate the method with an adaptive phase II/III clinical trial.

Some key words: Closed testing principle; Combination test; Conditional error; Multiple comparisons; Simultaneous inference.

1. INTRODUCTION

For experiments designed to make inference about a parameter vector $\theta = (\theta_1, \ldots, \theta_K)$, it is common to find confidence intervals for all of the individual θ_k such that the simultaneous coverage probability is at least $1 - \alpha$. Sometimes, though, an experimenter will only attempt to assert that an individual parameter exceeds a specific value, say $\theta_k > \delta_k$. If this cannot be achieved in such a way that the probability of making at least one incorrect rejection in a family of hypotheses $H_k = \{\theta_k \leqslant \delta_k\}$ $(k = 1, \ldots, K)$ is no greater than α, the experimenter will not assert anything about θ_k. The latter method of inference is used in so-called closed test procedures (Marcus et al., 1976), and its advantage is often greater power.

For experiments conducted in a single stage, Hayter & Hsu (1994) showed how simultaneous $100(1 - \alpha)\%$ confidence intervals can be constructed to be compatible with some commonly used closed test procedures, in the sense that a null hypothesis H_k is rejected at familywise level α if and only if the confidence interval for θ_k excludes all values for which H_k is true. Often, these intervals are scarcely more informative than the test decisions. For example, for one-sided problems where larger parameter values are more beneficial, no $100(1 - \alpha)\%$ lower confidence bound for any individual θ_k can exceed δ_k unless all hypotheses $H_1, \ldots, H_K$ can be rejected at familywise level α.

In this article we derive confidence intervals for adaptive experiments. Our motivating example is a seamless phase II/III clinical trial, although the method is not limited to this setting. Such trials consist of a first stage in which K experimental treatments, indexed by $T_1 = \{1, \ldots, K\}$, are compared with a common control and, after an interim analysis, a second stage in which only a subset of treatments, indexed by $T_2 \subseteq T_1$, are compared with the control. The state-of-the-art methodology for this problem (Bauer & Kieser, 1999; Posch et al., 2005; Bretz et al., 2009) is a hybrid of the closure principle of Marcus et al. (1976) and a p-value combination which goes back to Fisher (1932). This methodology allows any subset of treatments to be chosen at interim, based on all trial data and external factors. Other adaptations, such as sample size re-estimation, are also possible. A serious concern, though, is that there is no established method for constructing confidence intervals. As emphasized in the International Conference on Harmonisation's E9 guideline (ICH E9 Expert Working Group, 1999, p. 1932), 'Estimates of treatment effect should be accompanied by confidence intervals, whenever possible, and the way in which these will be calculated should be identified.'

Posch et al. (2005) proposed $100(1-\alpha)\%$ simultaneous confidence intervals following such a trial. Unfortunately, their intervals are not guaranteed to be compatible with the closed test procedure. Here, we construct intervals that are compatible. As in the one-stage case, an inevitable shortcoming of these intervals is that they are not always substantially more informative than the original test decisions. We will show that this problem is mitigated to some extent by the adaptive nature of the experiment.

2. FUNDAMENTAL METHODOLOGY

2·1. *Closure principle*

The closure principle of Marcus et al. (1976) is a general method for multiple hypothesis testing. A formal description is given in Finner & Strassburger (2002), and we adopt similar notation here. Let $\mathcal{P} = \{P_{\theta^*} : \theta^* \in \Theta\}$ be a family of probability measures defined on a common sample space $(\Omega, \mathcal{F})$, where Θ is a multi-dimensional parameter space. Suppose that we wish to test a family of null hypotheses $\mathcal{H} = \{H_i : i \in \mathcal{I}\}$, where $H_i \subset \Theta$ for each i in some index set $\mathcal{I}$. Let $\psi = \{\psi_i : i \in \mathcal{I}\}$ denote a multiple test of $\mathcal{H}$, with each component ψ_i taking value 0 or 1 corresponding to nonrejection or rejection of H_i, respectively. It is often desirable to ensure that

$$\sup_{\theta^* \in \Theta} P_{\theta^*}\left(\bigcup_{i \in I(\theta^*)} \{\psi_i = 1\}\right) \leqslant \alpha, \tag{1}$$

where $I(\theta^*) = \{i \in \mathcal{I} : \theta^* \in H_i\}$ is the index set of true hypotheses under θ^*. In other words, the probability of rejecting at least one true null hypothesis is bounded by α. This is known as strong control of the familywise error rate. The closure principle can be used to ensure (1). We are required to find, for each $I \subseteq \mathcal{I}$ such that $H_I = \bigcap_{i \in I} H_i$ is nonempty, a local level-α test φ_I for the intersection hypothesis H_I; that is, we require

$$\sup_{\theta^* \in H_I} P_{\theta^*}(\varphi_I = 1) \leqslant \alpha, \tag{2}$$

where φ_I takes values in $\{0, 1\}$ with the usual interpretation. If we define $\psi_i = \min_{I : H_I \neq \varnothing, H_I \subseteq H_i}(\varphi_I)$, then (1) holds. This can be very useful, as in many applications it is easy to find tests satisfying (2), whereas validating (1) directly is hard.

2·2. *Combination test*

Fisher (1932) discussed combining independent p-values to test a single null hypothesis. For convenience and brevity, we will only consider two-stage designs. We define a p-value combination function $Q : [0, 1]^2 \mapsto [0, 1]$ that is left-continuous and nondecreasing in both its arguments and is uniformly distributed provided that both arguments are themselves independent and uniformly distributed. An example is

$$Q(u, v) = 1 - \Phi\left[2^{1/2}\left\{\Phi^{-1}(1-u) + \Phi^{-1}(1-v)\right\}\right], \tag{3}$$

where Φ denotes the standard normal distribution function.

Such a combination function lends itself to a two-stage adaptive closed test, ψ, for a family of null hypotheses, $\mathcal{H}$. An important application, discussed in Bretz et al. (2009), is a seamless phase II/III confirmatory clinical trial. We henceforth restrict attention to a parameter $\theta = (\theta_1, \ldots, \theta_K)$ taking values in parameter space $\Theta = \mathbb{R}^K$ and a family of null hypotheses $\mathcal{H} = \{H_k : k \in T_1\}$ where $T_1 = \{1, \ldots, K\}$ and $H_k = \{\theta_k \leqslant \delta_k\}$ $(k \in T_1)$ for some constants $\delta_1, \ldots, \delta_K \in \mathbb{R}$. The θ_k $(k \in T_1)$ might correspond to the mean effects of K different treatments, for example. By defining local tests φ_I $(I \subseteq T_1)$ via a combination function Q, it is possible to make data-dependent modifications to the trial design at an interim analysis (cf. Bauer & Kieser, 1999; Hommel, 2001; Brannath et al., 2002). For instance, attention can be focused on a subset $T_2 \subseteq T_1$ of the initial hypotheses of interest; changes can be made to sample sizes, allocation ratios, etc.

2·3. *Two-stage closed test procedure*

Assume that the full first-stage trial data are represented by a random vector $X \in \mathbb{R}^n$ with distribution function $G(x; \theta)$. Prior to starting the trial, one must specify a combination function Q and, for each $I \subseteq T_1$, a first-stage test of $H_I = \bigcap_{i \in I} H_i$ with an associated p-value function $p_I^{(1)} : \mathbb{R}^n \to [0, 1]$ that satisfies $\sup_{\theta^* \in H_I} \int_{\{p_I^{(1)}(x) \leqslant u\}} \mathrm{d}G(x; \theta^*) \leqslant u$ for all $u \in [0, 1]$. The second-stage design is unspecified.

At the interim analysis, the experimenter defines a second-stage design, d, by choosing a subset of the original hypotheses, indexed by $T_2 \subseteq T_1$, to continue studying in the second stage, along with second-stage sample sizes and, for each $I \subseteq T_1$, a second-stage hypothesis test for H_I. See below for a proposal for choosing second-stage tests for H_I where $I \nsubseteq T_2$. We assume that the design d is allowed to depend on the unblinded first-stage data x without prespecifying an adaptation rule. Let Y denote the data collected at the second stage, taking values in $\mathbb{R}^m$, and let $p_{I,x,d}^{(2)}(y)$ $(I \subseteq T_1)$ denote the p-value functions of the second-stage tests. Because the tests used in the second stage depend on the first-stage data x and the chosen design d, the p-value functions will in general depend on both.

Let $F_{x,d}(y; \theta)$ denote the distribution function of the second-stage data, given the chosen design d and interim data x. We assume that for all x, d and $I \subseteq T_1$, the second-stage p-values $p_{I,x,d}^{(2)}$ satisfy $\sup_{\theta^* \in H_I} \int_{\{p_{I,x,d}^{(2)}(y) \leqslant u\}} \mathrm{d}F_{x,d}(y; \theta^*) \leqslant u$ for all $u \in [0, 1]$. The distribution $F_{x,d}$ is assumed to be known, i.e., not merely specified up to a null set, for all x and d, a condition that can be formalized by assuming an appropriate regression model (Brannath et al., 2012). See § 3·2 for a numerical example.

At the final analysis, for each $I \subseteq T_1$, the test decision is $\varphi_I = 1$ if and only if $Q\{p_I^{(1)}, p_{I,x,d}^{(2)}\} \leqslant \alpha$. As shown in Brannath et al. (2012), this combination test for H_I controls the Type I error rate at level α.

We assume that only data for the hypotheses indexed by T_2 are collected in the second stage and propose setting $p_I^{(2)} = p_{I\cap T_2}^{(2)}$ for $I \not\subseteq T_2$, where we drop the indices x and d for simplicity and set $p_\varnothing^{(2)} = 1$ by convention. Such second-stage p-values have the required distribution under $H_{I\cap T_2}$ and hence also under H_I.

We emphasize that while Type I error control is guaranteed even if the second-stage design is initially open-ended, in the design of actual clinical trials it is crucial to perform detailed planning based on likely first-stage outcomes. The added flexibility is necessary because it is impossible to foresee all eventualities in extremely complex areas such as clinical drug development.

3. CONFIDENCE REGIONS

3·1. *Partitioning the parameter space*

A standard approach to deriving a $100(1-\alpha)\%$ confidence set for θ is to perform a level-α test of each elementary hypothesis $\{\theta = \theta^*\}$ ($\theta^* \in \Theta$) and include all θ^* corresponding to nonrejected hypotheses (see, e.g., Lehmann, 1986, p. 90). To ensure compatibility with closed testing, the key idea (Stefansson et al., 1988; Hayter & Hsu, 1994; Finner & Strassburger, 2002) is to partition the parameter space into disjoint regions

$$\Theta_I = \{\theta^* \in \Theta : \theta_i^* \leqslant \delta_i,\ i \in I;\ \theta_i^* > \delta_i,\ i \in T_1 \setminus I\} \quad (I \subseteq T_1)$$

and apply different tests in each of the disjoint Θ_I. If, for each $I \subseteq T_1$, we let $\{\varphi_I(\theta^*) : \theta^* \in \Theta\}$ denote a family of tests with

$$\inf_{\theta^* \in \Theta} P_{\theta^*}\{\varphi_I(\theta^*) = 0\} \geqslant 1 - \alpha, \tag{4}$$

where $\varphi_I(\theta^*)$ takes values in $\{0, 1\}$ with the usual interpretation, we can apply the following general result from Hsu (1996, p. 234).

LEMMA 1. *A level-*$100(1-\alpha)\%$ *confidence set for* θ *is*

$$C = \bigcup_{I \subseteq T_1} \big[\{\theta^* \in \Theta : \varphi_I(\theta^*) = 0\} \cap \Theta_I\big]. \tag{5}$$

Our aim is to find families of tests such that C is compatible with the two-stage closed test procedure. This requires us to augment our specification of $p_{I\cap T_j}^{(j)}$ ($j = 1, 2$; $I \subseteq T_1$) with a family of p-values $\{p_{I\cap T_j}^{(j)}(\theta^*) : \theta^* \in \Theta\}$ where, under $\{\theta = \theta^*\}$, the distribution of $p_I^{(1)}(\theta^*)$ and $p_{I\cap T_2}^{(2)}(\theta^*)$ meet conditions as outlined for $p_I^{(1)}$ and $p_{I\cap T_2}^{(2)}$ in § 2·3. Additionally, if we treat the data as fixed and view each family as a function $p_{I\cap T_j}^{(j)} : \Theta \to [0, 1]$, then unless $I \cap T_j = \varnothing$, $p_{I\cap T_j}^{(j)}(\theta^*)$ is constant in all arguments θ_i^* such that $i \notin I \cap T_j$, and is left-continuous and nondecreasing in all arguments θ_i^* such that $i \in I \cap T_j$, with $p_{I\cap T_j}^{(j)}(\theta^*) = p_{I\cap T_j}^{(j)}$ for any θ^* such that $\theta_i^* = \delta_i$ for all $i \in I \cap T_j$. Furthermore, we assume that

$$\lim_{\theta_i^* \to \infty,\ i \in T_2} p_\varnothing^{(2)}(\theta^*) = 1. \tag{6}$$

PROPOSITION 1. *Inserted into* (5)*, the following families of hypothesis tests give rise to a* $100(1-\alpha)\%$ *confidence set for* θ*, denoted by* C*, that is compatible with the two-stage closed*

test procedure, i.e., $\psi_k = 1$ if and only if $H_k \cap C = \varnothing$: for $\varnothing \neq I \subseteq T_1$ and $\theta^* \in \Theta$,

$$\varphi_I(\theta^*) = \begin{cases} 1, & Q\{p_I^{(1)}(\theta^*), p_{I\cap T_2}^{(2)}(\theta^*)\} \leqslant \alpha, \\ 0, & Q\{p_I^{(1)}(\theta^*), p_{I\cap T_2}^{(2)}(\theta^*)\} > \alpha, \end{cases} \tag{7}$$

and $\{\varphi_\varnothing(\theta^*) : \theta^* \in \Theta\}$ is any family of tests satisfying (4)*.*

Proof. See the Appendix. □

There will be no unique collection of families of p-values satisfying the aforementioned distributional and monotonicity constraints. Rather, the families must be specified in a two-stage procedure in an analogous way to the p-values in § 2·3. As will become clear from the example below, for many commonly encountered scenarios and when $I \cap T_j \neq \varnothing$, the choice of $\{p_{I\cap T_j}^{(j)}(\theta^*) : \theta^* \in \Theta\}$ will be obvious from the choice of $p_{I\cap T_j}^{(j)}$. As a simple example, suppose that $p_{\{k\}}^{(j)}$ is the p-value from a one-sided z-test of the null hypothesis $\{\theta_k \leqslant \delta_k\}$ using the stage-j data only. Then the natural choice for $p_{\{k\}}^{(j)}(\theta^*)$ is the one-sided p-value from a standard z-test of $\{\theta_k \leqslant \theta_k^*\}$ using the same stage-j data.

While for $I \cap T_j \neq \varnothing$ there will often be a natural choice for $p_{I\cap T_j}^{(j)}(\theta^*)$, it is unclear how $\varphi_\varnothing(\theta^*)$ and $p_\varnothing^{(2)}(\theta^*)$ should be chosen. A reasonable suggestion is given below.

Corollary 1. *Define $p_\varnothing^{(j)}(\theta^*) = p_{T_j}^{(j)}(\theta^*)$ for $j = 1, 2$. The following is a $100(1-\alpha)\%$ confidence region for θ that is compatible with the two-stage closed test procedure:*

$$C_1 = \bigcup_{I \subseteq T_1} \left[\theta^* \in \Theta_I : Q\left\{p_I^{(1)}(\theta^*), p_{I\cap T_2}^{(2)}(\theta^*)\right\} > \alpha\right]. \tag{8}$$

The properties of a region defined by (8) are best illustrated by a specific example.

3·2. *Example*

Posch et al. (2005) considered a clinical trial where three active treatments, indexed by $T_1 = \{A, B, C\}$, are compared with a placebo using a two-stage adaptive design. The individual null hypotheses of interest are $H_k = \{\theta_k \leqslant 0\}$ $(k \in T_1)$, where $\theta_k = \pi_k - \pi_0$ denotes the difference between the success probabilities of treatment k and placebo. Denote the observed success rate of treatment k in stage j by $\hat{\pi}_{k,j}$ $(k \in T_1 \cup \{0\};\ j = 1, 2)$, where treatment 0 corresponds to a placebo.

At the design stage, the inverse normal combination function (3) is specified and $n_1 = 140$ first-stage patients are recruited to each treatment arm. Approximately, the $\hat{\theta}_{k,1} = \hat{\pi}_{k,1} - \hat{\pi}_{0,1}$ $(k \in T_1)$ are multivariate normal with $E(\hat{\theta}_{k,1}) = \theta_k$, $\mathrm{var}(\hat{\theta}_{k,1}) = \{\hat{\pi}_{k,1}(1-\hat{\pi}_{k,1}) + \hat{\pi}_{0,1}(1-\hat{\pi}_{0,1})\}/n_1$ and positive correlations. Based on this assumption, Simes (1986) tests are used for each intersection hypothesis; that is, $p_{\{k\}}^{(1)} = 1 - \Phi[\hat{\theta}_{k,1}\{\mathrm{var}(\hat{\theta}_{k,1})\}^{-1/2}]$ for $k \in T_1$ and, for $|I| > 1$, $p_I^{(1)} = \min_{k\in I} p_{\{k\}}^{(1)}|I|/R(k, I)$, where $R(k, I)$ denotes the rank of $p_{\{k\}}^{(1)}$

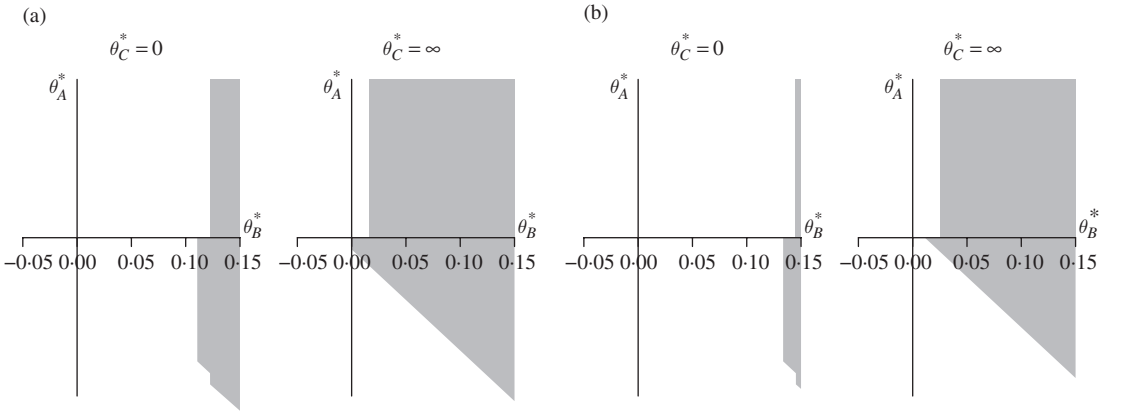

Fig. 1. Cross-sections of confidence regions of the form (9) for making inference on the second-stage parameter of interest, θ_B, in the example of § 3·2: (a) two cross-sections of the 97·5% confidence region; (b) two cross-sections of the 95% confidence region.

among $\{p_{\{i\}}^{(1)} : i \in I\}$. The natural way of augmenting these p-values is to define $p_{\{k\}}^{(1)}(\theta^*) = 1 - \Phi[(\hat{\theta}_{k,1} - \theta_k^*)\{\mathrm{var}(\hat{\theta}_{k,1})\}^{-1/2}]$ for $k \in T_1$ and $p_I^{(1)}(\theta^*) = \min_{k \in I} p_{\{k\}}^{(1)}(\theta^*)|I|/R(k, I, \theta^*)$ for $|I| > 1$, where $R(k, I, \theta^*)$ denotes the rank of $p_{\{k\}}^{(1)}(\theta^*)$ among $\{p_{\{i\}}^{(1)}(\theta^*) : i \in I\}$.

Suppose that the unblinded first-stage results are $\hat{\pi}_{0,1} = 0{\cdot}21$, $\hat{\pi}_{A,1} = 0{\cdot}22$, $\hat{\pi}_{B,1} = 0{\cdot}3$ and $\hat{\pi}_{C,1} = 0{\cdot}36$. The experimenter decides that treatments A and C are not to be considered in the second stage owing to lack of efficacy and safety concerns, respectively. A further $n_2 = 140$ patients are recruited to both treatment B and placebo. A family of p-values with $p_{\{B\}}^{(2)}(\theta^*) = 1 - \Phi[(\hat{\theta}_{B,2} - \theta_B^*)\{\mathrm{var}(\hat{\theta}_{B,2})\}^{-1/2}]$ is chosen, where $\hat{\theta}_{B,2} = \hat{\pi}_{B,2} - \hat{\pi}_{0,2}$.

Now suppose that the second-stage results are $\hat{\pi}_{0,2} = 0{\cdot}19$ and $\hat{\pi}_{B,2} = 0{\cdot}31$. The p-values from the elementary hypotheses are $p_{\{A\}}^{(1)} = 0{\cdot}419$, $p_{\{B\}}^{(1)} = 0{\cdot}0412$, $p_{\{C\}}^{(1)} = 0{\cdot}00241$ and $p_{\{B\}}^{(2)} = 0{\cdot}00961$. Therefore $p_{\{A,B,C\}}^{(1)} = 3p_{\{C\}}^{(1)}$, $p_{\{A,B\}}^{(1)} = 2p_{\{B\}}^{(1)}$ and $p_{\{B,C\}}^{(1)} = 2p_{\{C\}}^{(1)}$. As $\min_{I \subseteq T_1, B \in I} Q(p_I^{(1)}, p_B^{(2)}) \leqslant 0{\cdot}025$, H_B can be rejected at familywise level 0·025. Both H_A and H_C fail to be rejected, as $Q\{p_{\{k\}}^{(1)}, 1\} = 1$ for $k = A, C$. A compatible 97·5% confidence region for θ is given by

$$\bigcup_{I \subseteq T_1} \left\{ \theta^* \in \Theta_I : Q\{p_I^{(1)}(\theta^*), p_B^{(2)}(\theta^*)\} > 0{\cdot}025 \right\}, \tag{9}$$

where $p_{\varnothing}^{(1)}(\theta^*)$ is defined as $p_{T_1}^{(1)}(\theta^*)$ for all $\theta^* \in \Theta$.

The region (9) will have a complicated three-dimensional shape. However, in terms of making inference on θ_B, its crucial features can be seen by taking two cross-sections, as displayed in Fig. 1. As $p_I^{(1)}(\theta^*)$ is nondecreasing in θ_C^* for all $I \subseteq T_1$, we know that for any $\gamma \in (-\infty, 0)$, the cross-section at $\theta_C^* = \gamma$ is contained in the cross-section at $\theta_C^* = 0$. Similarly, for any $\gamma \in (0, \infty)$, the cross-section at $\theta_C^* = \gamma$ is contained in the limit of the cross-section of the region as $\theta_C^* \to \infty$. One can see immediately from Fig. 1 that for any $\epsilon > 0$, the 97·5% confidence region fails to exclude all parameter vectors θ^* such that $\theta_B^* \leqslant \epsilon$. In other words, the lower confidence bound on θ_B provides no more information than the decision of the closed test procedure.

For confidence intervals that are compatible with single-stage closed test procedures (Hayter & Hsu, 1994; Strassburger & Bretz, 2008; Guilbaud, 2008), a necessary condition for obtaining informative lower confidence bounds for parameters corresponding to the rejected

null hypotheses is that $\psi_k = 1$ for all $k \in T_1$. In the adaptive setting, this is no longer a necessary condition. For example, repeating the above test procedure at level $\alpha = 0{\cdot}05$, the compatible 95% confidence region analogous to (9) is also summarized in Fig. 1. Here it appears, and indeed can be verified by considering all values of θ_A^*, that there does exist some $\epsilon > 0$ such that the confidence region excludes all parameter vectors θ^* for which $\theta_B^* \leqslant \epsilon$. We will show that for the two-stage adaptive setting, a necessary condition for informative lower confidence bounds on parameters corresponding to the rejected null hypotheses is that $\psi_k = 1$ for all $k \in T_2$. However, as can be seen from Fig. 1, this condition is not sufficient.

3·3. *A two-stage, single-step confidence region*

Posch et al. (2005) proposed the following $100(1-\alpha)\%$ confidence region:

$$C_2 = \left\{\theta^* \in \Theta : Q\left\{p_{T_1}^{(1)}(\theta^*), p_{T_2}^{(2)}(\theta^*)\right\} > \alpha\right\}. \tag{10}$$

They note that the resulting confidence intervals are not compatible with the closed test procedure described in § 2·3 (Posch et al., 2005, p. 3702). Nevertheless, the region (10) can be used to generate an alternative multiple test. More generally, any $1-\alpha$ confidence set C generates a multiple test for a family of hypotheses $\mathcal{H}$, whereby $H_k \in \mathcal{H}$ is rejected if and only if $H_k \cap C = \varnothing$. This guarantees strong control of the familywise error rate (1). The multiple test generated by (10) can be thought of as single-step in the sense that rejection or nonrejection of a null hypothesis does not take into account the decision for any other hypothesis. If H_k is rejected, informative lower bounds will be available for θ_k regardless of the test decisions for all other hypotheses.

4. Computation of confidence intervals

4·1. *Least-favourable parameter configurations*

In the above example, marginal inference on θ_B was achieved by considering least-favourable parameter configurations for θ_k, $k \in T_1 \setminus \{B\}$. This idea can be generalized to find $100(1-\alpha)\%$ simultaneous confidence intervals containing (8) or (10).

Definition 1. *For $j = 1, 2$, $k \in T_1$ and $I \subseteq T_j$, the locally least-favourable jth-stage p-value function for H_k in Θ_I, $p_{k,I}^{(j)} : \mathbb{R} \to [0, 1]$, is defined for $I \neq \varnothing$ as $p_{k,I}^{(j)}(\vartheta) = p_I^{(j)}(\xi)$, where $\xi = (\xi_1, \ldots, \xi_K)$ with $\xi_i = \delta_i$ for $i \neq k$ and $\xi_k = \vartheta$. Additionally, for $j = 1, 2$,*

$$p_{k,\varnothing}^{(j)}(\vartheta) = \lim_{\xi_i \to \infty, i \in T_j \setminus \{k\}} p_{T_j}^{(j)}(\xi) \quad (\xi_k = \vartheta). \tag{11}$$

Proposition 2. *The smallest Cartesian product of intervals, $\times_{k \in T_1}(l_k, \infty)$, that contains the confidence region (8) has $l_k = \min_{I \subseteq T_1} l_{k,I}$, where for $k \in I$,*

$$l_{k,I} = \begin{cases} \infty & (\varphi_I = 1), \\ \sup\left\{\vartheta : Q\{p_{k,I}^{(1)}(\vartheta), p_{k,I \cap T_2}^{(2)}(\vartheta)\} \leqslant \alpha\right\} & (\varphi_I = 0), \end{cases} \tag{12}$$

and for $k \notin I$,

$$l_{k,I} = \max\left(\delta_k, \sup\left\{\vartheta : Q\{p_{k,I}^{(1)}(\vartheta), p_{k,I \cap T_2}^{(2)}(\vartheta)\} \leqslant \alpha\right\}\right). \tag{13}$$

Furthermore, these intervals are compatible with the two-stage closed test procedure, i.e., $\psi_k = 1$ if and only if $H_k \cap \times_{k \in T_1}(l_k, \infty) = \varnothing$.

Proof. See the Appendix. □

In general, to find each interval requires one-dimensional root finding for each $I \subseteq T_1$, a calculation that is $O(2^K)$. However, substantial shortcuts are available for reducing the computational burden.

4·2. *Efficient computation of confidence bounds*

There are two possible scenarios at the end of the closed test procedure: either $\psi_k = 1$ for all $k \in T_2$, or at least one H_k ($k \in T_2$) fails to be rejected. In the latter case, there exists some $I \subseteq T_1$ with $I \cap T_2 \neq \varnothing$ such that for any $k \in T_2$,

$$\alpha < Q\big(p_I^{(1)}, p_{I\cap T_2}^{(2)}\big) = Q\big\{p_{k,I}^{(1)}(\delta_k), p_{k,I\cap T_2}^{(2)}(\delta_k)\big\}$$

and therefore $l_k \leqslant l_{k,I} \leqslant \delta_k$. Due to the compatibility of the intervals with the closed test procedure, if $\psi_k = 1$, then $l_k = \delta_k$; if $\psi_k = 0$, then $l_k < \delta_k$.

If $\psi_k = 1$ for all $k \in T_2$, then $l_k \geqslant \delta_k$ for all $k \in T_2$. Additionally, we can use the fact that for all $k \in T_2$ and $I \subseteq T_1$ with $I \cap T_2 \neq \varnothing$, we know from (12) and (13) that $l_{k,I} = \infty$; so, when finding $l_k = \min_{I\subseteq T_1} l_{k,I}$ in Proposition 2, the minimum can be taken over a much smaller number of $l_{k,I}$. The following algorithm finds the lower bounds for all parameters corresponding to the rejected hypotheses.

Step 1. Perform the closed test procedure. If $\psi_{k'} = 0$ for some $k' \in T_2$, then $l_k = \delta_k$ for $\psi_k = 1$ and $l_k < \delta_k$ for $\psi_k = 0$. If $\psi_k = 1$ for all $k \in T_2$, go to Step 2.

Step 2. Find $p_{\mathrm{M}} = \max_{\varnothing \neq I\subseteq T_1\setminus T_2} p_I^{(1)}$. If $T_1 \setminus T_2 = \varnothing$, then $p_{\mathrm{M}} = 0$.

Step 3. For $k \in T_2$,

$$l_k = \max\left[\delta_k,\ \sup\left\{\vartheta : Q\left[\max\{p_{\mathrm{M}},\ p_{k,\varnothing}^{(1)}(\vartheta)\},\ p_{k,\varnothing}^{(2)}(\vartheta)\right] \leqslant \alpha\right\}\right].$$

The cost of computing the intervals for θ_k ($k \in T_2$) in Step 3 is linear in the number of parameters. In general, Step 1 is $O(2^{|T_1|})$, but a shortcut of $O(|T_1|^2)$ is given in Brannath & Bretz (2010). Step 2 is $O(2^{|T_1\setminus T_2|})$, but a shortcut of size $|T_1 \setminus T_2|$ is available, provided there exists an ordering $i_1, \ldots, i_k$ of $T_1 \setminus T_2$ such that for each $u \in \{1, \ldots, k\}$, $p_J^{(1)} \leqslant p_L^{(1)}$ for all $J \subseteq L \subseteq \{i_u, \ldots, i_k\}$ with $i_u \in J$. This is because we only have to check $p_{\{i_u,\ldots,i_k\}}^{(1)}$ for $u = 1, \ldots, k$. Many common multiple test procedures, such as those based on Dunnett (1955) tests or weighted Bonferroni tests, satisfy this condition, with the ordering $i_1, \ldots, i_k$ following the ordering of the univariate test statistics or the weighted elementary p-values (Brannath & Bretz, 2010).

4·3. *Lower bounds for parameters corresponding to retained hypotheses*

Consider $k \in T_2$ such that $\psi_k = 0$. We know that $l_k < \delta_k$, and therefore we need only consider $l_{k,I}$ such that $k \in I$. However, since in general $l_{k,I} < \infty$, finding the minimum such lower bound will still have a computational cost that is exponential in the number of parameters.

For $k \in I \subseteq T_1 \setminus T_2$, we have $p_{k,I\cap T_2}^{(2)}(\vartheta) = p_{k,\varnothing}^{(2)}(\vartheta)$ and know from (11) and (6) that this is equal to 1. Many commonly used combination functions, including (3), have the property that $v = 1$ implies $Q(u, v) = 1$. In this case, $l_k = -\infty$ for all $k \in T_1 \setminus T_2$.

4·4. *Lower bounds for the two-stage single-step procedure*

Posch et al. (2005) showed that the region (10) is contained in a rectangle, $\times_{k\in T_1}(\bar{l}_k, \infty)$, where

$$\bar{l}_k = \sup\left\{\vartheta : Q\{p^{(1)}_{k,\varnothing}(\vartheta), p^{(2)}_{k,\varnothing}(\vartheta)\} \leqslant \alpha\right\}. \tag{14}$$

The computation of each interval requires only a one-dimensional search for a root, and overall computation will be linear in the number of parameters.

4·5. *Example continued*

Recall from § 3·2 that $T_2 = \{B\}$ and $\psi_B = 1$. Proceeding to Step 2 of the above algorithm, $p_{\mathrm{M}} = 0{\cdot}419$. In this case we need just one iteration in Step 3, because

$$Q\left[\max\{0{\cdot}419,\ p^{(1)}_{B,\varnothing}(0)\},\ p^{(2)}_{B,\varnothing}(0)\right] = 0{\cdot}0360 > 0{\cdot}025,$$

and therefore the 97·5% confidence interval for θ_B is $(0, \infty)$, consistent with Fig. 1. This example emphasizes that there is a price to pay for the additional power of the closed test as opposed to the single-step procedure of § 3·3 with, by (14),

$$\bar{l}_B = \sup\left\{\vartheta\ :\ Q\{p^{(1)}_{B,\varnothing}(\vartheta), p^{(2)}_{B,\varnothing}(\vartheta)\} \leqslant 0{\cdot}025\right\} = 0{\cdot}0159.$$

While this agrees with the assertion $\theta_B > 0$ in this specific case, it is invalid to claim it as a 97·5% lower confidence bound if the closed test procedure of § 2·3 had been planned. One can see that for any $\alpha > 0{\cdot}036$, the $100(1-\alpha)$% confidence interval for treatment B that is compatible with the closed test procedure has a positive lower bound. For example, the 95% lower confidence bound is $l_B = 0{\cdot}0112$, consistent with Fig. 1. Again, if the region (10) had been specified pre-trial, the 95% lower confidence bound (14) would have been $\bar{l}_B = 0{\cdot}0252$.

5. Confidence bounds for closed tests based on the conditional error rate

Consider again the two-stage closed test procedure of § 2·3. As an alternative to combination tests, Koenig et al. (2008) used the conditional error approach (Proschan & Hunsberger, 1995) to derive local tests φ_I $(I \subseteq T_1)$. The only difference is that instead of prespecifying a combination function Q and first-stage p-value $p^{(1)}_I$, one must prespecify a measurable conditional error function $A_I : \mathbb{R}^n \to [0, 1]$ such that

$$\sup_{\theta^*\in H_I} \int_{\mathbb{R}^n} A_I(x)\,\mathrm{d}G(x;\theta^*) \leqslant \alpha$$

and, at the final analysis, $\varphi_I = 1$ if and only if $p^{(2)}_{I\cap T_2} \leqslant A_I(x)$.

To produce a compatible $100(1-\alpha)$% confidence region for θ, each A_I $(I \subseteq T_1)$ must be augmented with a family of conditional error functions $\{A_I(\theta^*) : \theta^* \in \Theta\}$ such that $\int_{\mathbb{R}^n} A_I(\theta^*)(x)\,\mathrm{d}G(x;\theta^*) \leqslant \alpha$ and, for fixed $x \in \mathbb{R}^n$, $A_I(\theta^*)$ is constant in all arguments θ^*_i with $i \notin I$ and is left-continuous and nonincreasing in all arguments θ^*_i with $i \in I$. Furthermore, $A_I(\theta^*) = A_I$ for all $\theta^* \in \Theta$ such that $\theta^*_i = \delta_i$ for $i \in I$. The second-stage p-values $p^{(2)}_{I\cap T_2}$ $(I \subseteq T_1)$ must be augmented with a family $\{p^{(2)}_{I\cap T_2}(\theta^*) : \theta^* \in \Theta\}$ as described in § 3·1.

Müller & Schäfer (2004) propose defining $A_I = \sup_{\theta^*\in H_I} E_{\theta^*}(\phi_I \mid X)$, where ϕ_I is a pre-planned fixed sample level-α test for H_I. In many situations the natural choice for $A_I(\theta^*)$ will

be obvious from A_I. For example, if ϕ_I is the decision function for a Dunnett (1955) test of $H_I = \bigcap_{k\in I}\{\theta_k \leqslant \delta_k\}$, then it is natural to choose $A_I(\theta^*) = E_{\theta^*}(\phi_{I,\theta^*} \mid X)$ where ϕ_{I,θ^*} is the decision function for a Dunnett test of $\bigcap_{k\in I}\{\theta_k \leqslant \theta_k^*\}$, which can be derived via a corresponding translation of the test statistics.

Using the arguments of Propositions 1 and 2, it can be shown that, analogously to (8), a compatible $100(1-\alpha)\%$ confidence region for θ is

$$\bigcup_{I\subseteq T_1}\left\{\theta^* \in \Theta_I : p^{(2)}_{I\cap T_2}(\theta^*) > A_I(\theta^*)\right\},$$

where $p^{(2)}_{\varnothing}(\theta^*)$ and $A_{\varnothing}(\theta^*)$ are set equal to $p^{(2)}_{T_2}(\theta^*)$ and $A_{T_1}(\theta^*)$, respectively. Also, the largest compatible $100(1-\alpha)\%$ confidence lower bounds are $l_k = \min_{I\subseteq T_1} l_{k,I}$, where for $k \in I$,

$$l_{k,I} = \begin{cases} \infty & (\varphi_I = 1), \\ \sup\{\vartheta : p^{(2)}_{k,I\cap T_2}(\vartheta) \leqslant A_{k,I}(\vartheta)\} & (\varphi_I = 0), \end{cases}$$

and for $k \notin I$, $l_{k,I} = \max[\delta_k, \sup\{\vartheta : p^{(2)}_{k,I\cap T_2}(\vartheta) \leqslant A_{k,I}(\vartheta)\}]$ with $A_{k,I}(\vartheta)$ defined analogously to $p^{(1)}_{k,I}(\vartheta)$ $(k \in T_1;\, I \subseteq T_1)$ in Definition 1.

6. CONCLUDING REMARKS

The lower confidence bounds (12)–(13) provide more information about the location of θ than the decisions of the closed test procedure of § 2·3. The utility of this additional information will depend strongly on the context. In practice, the primary concern will often be to find lower bounds for the components of θ corresponding to the rejected null hypotheses. As this can be achieved using an algorithm that is $O(K^2)$, application to large-scale simultaneous inference problems is, in principle, feasible. However, these lower bounds will only be informative if all hypotheses considered in the second stage of testing are rejected, and even this may be insufficient. In practice, therefore, the lower bounds (12)–(13) are only likely to be useful in relatively small-scale problems. Furthermore, in situations where informative lower confidence bounds are deemed to be more important than the possibility of rejecting as many individual null hypotheses as possible, it would be sensible to use the intervals (14) instead of applying the closed test procedure. For large-scale simultaneous inference problems, an approach based on controlling the false coverage-statement rate (Benjamini & Yekutieli, 2005) may be more appropriate than aiming for a high simultaneous coverage probability.

Extensions to more than two stages and to allow early rejection of hypotheses are straightforward with an appropriate combination function in place of (3). An open question is how best to choose $\varphi_{\varnothing}(\theta^*)$ and $p^{(2)}_{\varnothing}(\theta^*)$. The tests we use in region (8) are a natural choice but may not be the most powerful.

ACKNOWLEDGEMENT

This work was supported by the National Institute for Health Research and the Austrian Science Fund. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health.

Appendix

Proof of Proposition 1. With the assumptions in § 3·1, all tests of the form (7) satisfy condition (4), and therefore C is a $100(1-\alpha)\%$ confidence set for θ. By the monotonicity conditions imposed on the p-values, we have $p^{(j)}_{I\cap T_j}(\theta^*) \leqslant p^{(j)}_{I\cap T_j}$ for all $\theta^* \in \Theta_I$ $(j = 1, 2;\ I \neq \varnothing;\ I \subseteq T_1)$, so that $\Theta_I \cap C = \varnothing$ if and only if $Q(p^{(1)}_I, p^{(2)}_{I\cap T_2}) \leqslant \alpha$. Therefore, $\psi = 1$ if and only if $\min_{I\subseteq T_1, k\in I} Q(p^{(1)}_I, p^{(2)}_{I\cap T_2}) \leqslant \alpha$ if and only if $\bigcup_{I\subseteq T_1, k\in I} \Theta_I \cap C = \varnothing$. Since $\bigcup_{I\subseteq T_1, k\in I} \Theta_I = H_k$, we have compatibility. □

Proof of Proposition 2. First, note the key property that $p^{(j)}_{k,I\cap T_j}(\vartheta) \geqslant p^{(j)}_{I\cap T_j}(\theta^*)$ for all $\theta^* \in \Theta_I$ with $\theta^*_k \leqslant \vartheta$ $(I \subseteq T_1;\ k \in T_1;\ j = 1, 2)$.

To show that $C_1 \subseteq \times_{k\in T_1}(l_k, \infty)$, consider any $\theta^* \in \Theta \setminus \times_{k\in T_1}(l_k, \infty)$. We must have $\theta^* \subseteq \Theta_I$ for some $I \subseteq T_1$ and $\theta^*_k \leqslant l_k$ for some $k \in T_1$. If $k \in I$, then $\theta^*_k \leqslant \min(\delta_k, l_{k,I})$, and (12) implies that $\alpha \geqslant Q\{p^{(1)}_{k,I}(\theta^*_k), p^{(2)}_{k,I\cap T_2}(\theta^*_k)\} \geqslant Q\{p^{(1)}_I(\theta^*), p^{(2)}_{I\cap T_2}(\theta^*)\}$. The same inequality follows from $l_{k,I} \geqslant \theta^*_k > \delta_k$ and (13) if $k \notin I$. Therefore, $\theta^* \notin C_1$ and $C_1 \subseteq \times_{k\in T_1}(l_k, \infty)$.

To show that no smaller interval $(l_k + \epsilon, \infty)$ is possible for any $\epsilon > 0$, we must find some $\theta^* \in C_1$ with $\theta^*_k \in (l_k, l_k + \epsilon)$. Consider a subset $I \subseteq T_1$ such that $l_k = l_{k,I}$ and therefore $Q\{p^{(1)}_{k,I}(\vartheta), p^{(2)}_{k,I\cap T_2}(\vartheta)\} > \alpha$ for all $\vartheta > l_k$. If $k \in I$ or, equivalently, $l_k < \delta_k$, take any $\theta^*_k \in (l_k, \min\{\delta_k, l_k + \epsilon\})$. If $k \notin I$ or, equivalently, $l_k \geqslant \delta_k$, take any $\theta^*_k \in (l_k, l_k + \epsilon)$. Now consider a parameter vector $\xi^{I,k} = (\xi^{I,k}_1, \ldots, \xi^{I,k}_K)$, where $\xi^{I,k}_k = \theta^*_k$, $\xi^{I,k}_i = \delta_i$ for $k \neq i \in I$, and $\xi^{I,k}_i > \delta_i$ for $i \notin I \cup \{k\}$. All such parameter vectors $\xi^{I,k}$ are contained in Θ_I, and

$$\alpha < Q\left\{p^{(1)}_{k,I}(\theta^*_k), p^{(2)}_{k,I\cap T_2}(\theta^*_k)\right\} = \lim_{\xi^{I,k}_i \to \infty, i\notin I\cup\{k\}} Q\left\{p^{(1)}_I(\xi^{I,k}), p^{(2)}_{I\cap T_2}(\xi^{I,k})\right\}.$$

Thus there exists some such $\xi^{I,k} \in C_1$, and hence C_1 is not contained in this smaller product of intervals.

Finally, $H_k \cap \times_{k\in T_1}(l_k, \infty) = \varnothing$ if and only if $l_{k,I} \geqslant \delta_k$ for $I \subseteq T_1$, if and only if $Q\{p^{(1)}_{k,I}(\delta_k), p^{(2)}_{k,I\cap T_2}(\delta_k)\} = Q\{p^{(1)}_I, p^{(2)}_{I\cap T_2}\} \leqslant \alpha$ for $I \subseteq T_1$ and $k \in I$, if and only if $\psi_k = 1$. □

References

Bauer, P. & Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statist. Med.* **18**, 1833–48.

Benjamini, Y. & Yekutieli, Y. (2005). False discovery rate controlling confidence intervals for selected parameters. *J. Am. Statist. Assoc.* **100**, 71–80.

Brannath, W. & Bretz, F. (2010). Shortcuts for locally consonant closed test procedures. *J. Am. Statist. Assoc.* **105**, 660–9.

Brannath, W., Gutjahr, G. & Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *J. Am. Statist. Assoc.* **107**, 824–32.

Brannath, W., Posch, M. & Bauer, P. (2002). Recursive combination tests. *J. Am. Statist. Assoc.* **97**, 236–44.

Bretz, F., Koenig, F., Brannath, W., Glimm, E. & Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statist. Med.* **28**, 1181–217.

Dunnett, C. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Am. Statist. Assoc.* **50**, 1096–121.

Finner, H. & Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Ann. Statist.* **30**, 1194–213.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver and Boyd, 4th ed.

Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biomet. J.* **50**, 678–92.

Hayter, A. J. & Hsu, J. C. (1994). On the relationship between stepwise decision procedures and confidence sets. *J. Am. Statist. Assoc.* **89**, 128–36.

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biomet. J.* **43**, 581–9.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. London: Chapman and Hall.

ICH E9 Expert Working Group (1999). Statistical principles for clinical trials: ICH harmonized tripartite guideline. *Statist. Med.* **18**, 1905–42.

Koenig, F., Brannath, W., Bretz, F. & Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statist. Med.* **27**, 1612–25.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. New York: Wiley, 2nd ed.

MARCUS, R., PERITZ, E. & GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–60.

MÜLLER, H. H. & SCHÄFER, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statist. Med.* **23**, 2497–508.

POSCH, M., KOENIG, F., BRANSON, M., BRANNATH, W., DUNGER-BALDAUF, C. & BAUER, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statist. Med.* **24**, 3697–714.

PROSCHAN, M. & HUNSBERGER, S. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–24.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–4.

STEFANSSON, G., KIM, W. & HSU, J. (1988). On confidence sets in multiple comparisons. In *Statistical Decision Theory and Related Topics IV*, S. S. Gupta & J. O. Berger, eds. New York: Springer, pp. 89–104.

STRASSBURGER, K. & BRETZ, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statist. Med.* **27**, 4914–27.

[*Received February* 2012. *Revised May* 2013]