The Annals of Statistics
2010, Vol. 38, No. 6, 3782–3810
DOI: 10.1214/10-AOS829

THE SEQUENTIAL REJECTION PRINCIPLE OF FAMILYWISE ERROR CONTROL

BY JELLE J. GOEMAN[1] AND ALDO SOLARI

Leiden University Medical Center and University of Padova

Closed testing and partitioning are recognized as fundamental principles of familywise error control. In this paper, we argue that sequential rejection can be considered equally fundamental as a general principle of multiple testing. We present a general sequentially rejective multiple testing procedure and show that many well-known familywise error controlling methods can be constructed as special cases of this procedure, among which are the procedures of Holm, Shaffer and Hochberg, parallel and serial gatekeeping procedures, modern procedures for multiple testing in graphs, resampling-based multiple testing procedures and even the closed testing and partitioning procedures themselves. We also give a general proof that sequentially rejective multiple testing procedures strongly control the familywise error if they fulfill simple criteria of monotonicity of the critical values and a limited form of weak familywise error control in each single step. The sequential rejection principle gives a novel theoretical perspective on many well-known multiple testing procedures, emphasizing the sequential aspect. Its main practical usefulness is for the development of multiple testing procedures for null hypotheses, possibly logically related, that are structured in a graph. We illustrate this by presenting a uniform improvement of a recently published procedure.

1. Introduction. Well-known multiple testing procedures that control the familywise error are often sequential, in the sense that rejection of some of the hypotheses may make rejection of the remaining hypotheses easier. A famous example is Holm's (1979) procedure, in which the alpha level for rejection of each null hypothesis depends on the number of previously rejected hypotheses. Other classical examples of sequentially rejective multiple testing procedures include various types of gatekeeping procedures [Dmitrienko and Tamhane (2007)], which can be explicitly constructed as sequential, and the closed testing procedure [Marcus, Peritz and Gabriel (1976)], in which rejection of a null hypothesis can only occur after all implying intersection null hypotheses have been rejected. Other modern multiple testing procedures, such as the exact resampling-based method of Romano and Wolf (2005), as well as recent methods for multiple testing in graphs of logically related hypotheses [Goeman and Mansmann (2008), Meinshausen (2008)], can also be viewed as sequentially rejective procedures.

Received December 2009; revised April 2010.
[1]Supported by NWO Veni Grant 91676122.
AMS 2000 subject classifications. Primary 62H15; secondary 62J15.
Key words and phrases. Multiple testing, familywise error rate, graph.

This paper presents a unified approach to the class of sequentially rejective multiple testing procedures, emphasizing the sequential aspect. A general sequentially rejective procedure will be constructed as a sequence of single-step methods, determined by a rule for setting the rejection regions for each null hypothesis based on the current collection of unrejected null hypotheses. The general sequentially rejective procedure encompasses all of the methods mentioned above, as well as many others. Our work continues along the path set out by Romano and Wolf (2005), who wrote of stepwise procedures that "an ideal situation would be to proceed at any step without regard to previous rejections, in the sense that once a hypothesis is rejected, the remaining hypotheses are treated as a new family, and testing for this new family proceeds independent of past decisions." We extend the work of Romano and Wolf (2005, 2010) to logically related hypotheses and show that past decisions can even make the tests in each new family easier, as the tests for each new family may assume that all rejections in previous families were correct rejections, as in Shaffer's (1986) procedure. By emphasizing the role of logical relationships between hypotheses, we are able to demonstrate the versatility of sequential rejection as an approach to multiple testing.

We give a general proof that sequentially rejective multiple testing procedures strongly control the familywise error. The proof shows that, for any sequentially rejective multiple testing procedure that fulfills a simple monotonicity requirement, strong familywise error control of the sequential procedure follows from a limited form of weak familywise error control at each single step. This property, which can be used to turn a single-step familywise error controlling procedure into a sequential one, is a very general principle of familywise error control. We refer to this principle as the *sequential rejection principle*. It does not depend in any way on the method of familywise error control imposed in the single steps and it does not require any additional assumptions on the joint distribution of the test statistics, aside from what is needed for single-step familywise error control.

It is a notable feature of the sequential rejection principle that control of the familywise error at each single step is only necessary with respect to those data distributions for which all previous rejections have been correct rejections. As a consequence, the principle facilitates the design of sequentially rejective multiple testing procedures in situations in which there are logical relationships between null hypotheses. Also, in other cases, the principle may aid the design of multiple testing procedures since, by the principle, proof of familywise error control of the sequential procedure can be achieved by checking monotonicity and proving weak familywise error control of single steps, which, typically, is relatively easy to do.

Earlier generalizations of sequentially rejective testing were formulated by Romano and Wolf (2005) and Hommel, Bretz and Maurer (2007), both on the basis of the closed testing procedure Marcus, Peritz and Gabriel (1976). Our procedure can be seen as an extension of these procedures, encompassing both as a special case, as well as some other procedures that these earlier generalizations do not

encompass. The procedure of Hommel, Bretz and Maurer is limited to Bonferroni-based control at each single step. The procedure of Romano and Wolf was originally limited to have identical critical values for all hypotheses, although this was generalized by Romano and Wolf (2010). Neither Romano and Wolf (2005, 2010) nor Hommel, Bretz and Maurer (2007) explicitly considered the issue of logically related hypotheses.

This paper is organized as follows. We first formulate the general sequentially rejective multiple testing procedure and a set of sufficient conditions under which such a procedure guarantees strong control of the familywise error. Together with the formal statements, much attention will be given to the development of the intuitions behind the principle. The remaining part of the paper is devoted to a review of well-known multiple testing procedures, in which we show how important procedures such as Shaffer, Hochberg, closed testing, partitioning and gatekeeping procedures can be viewed as examples of the general sequentially rejective procedure. The majority of sequentially rejective procedures use some version of Bonferroni, modified by Shaffer's (1986) treatment of logically related null hypotheses, in their single-step control of the familywise error. We go into this specific class of procedures in more detail in Section 3. We also give examples of non-Bonferroni-based procedures, such as resampling-based multiple testing [Romano and Wolf (2005)] and the step-up method of Hochberg (1988), that can be viewed as special cases of the general sequentially rejective multiple testing procedure, demonstrating that the sequential rejection principle is not restricted to Bonferroni–Shaffer-based methods. Next, we demonstrate how the sequential rejection principle might be used to improve existing procedures by presenting a uniform improvement of the method of Meinshausen (2008) for tree-structured hypotheses. Finally, we show how to calculate multiplicity-adjusted p-values for the general sequentially rejective procedure.

2. Sequential rejection. The formulation of the general sequentially rejective procedure and its proof require formal notation. We suppose that we have a statistical model, a set $\mathbb{M}$ for which each $M \in \mathbb{M}$ indexes a probability measure P_M, defined on a common outcome space Ω. We also suppose that we have a collection $\mathcal{H}$ of null hypotheses of interest, each of which is a proper submodel of $\mathbb{M}$, that is, $H \subset \mathbb{M}$ for every $H \in \mathcal{H}$. Depending on P_M, some or all of the hypotheses in $\mathcal{H}$ may be true null hypotheses. For each $M \in \mathbb{M}$, we define the collection of true null hypotheses as $\mathcal{T}(M) = \{H \in \mathcal{H} : M \in H\} \subseteq \mathcal{H}$ and the collection of false null hypotheses as $\mathcal{F}(M) = \mathcal{H} \setminus \mathcal{T}(M)$. If desired, the collection $\mathcal{H}$ may contain an infinite number of hypotheses. Collections such as $\mathcal{H}$ are collections of sets. We use the shorthand

$$\bigcup \mathcal{A} = \bigcup_{A \in \mathcal{A}} A$$

when working with such collections of sets (both for unions and for intersections). We use the phrase *almost surely* for statements that hold with probability 1 for every M in $\mathbb{M}$.

2.1. *The principle.* We first present the sequential rejection principle in a general set-theoretic form that does not involve test statistics and critical values.

In general, we define a sequentially rejective multiple testing procedure of the hypotheses in $\mathcal{H}$ by choosing a random and measurable function $\mathcal{N}$ that maps from the power set $2^{\mathcal{H}}$ of all subsets of $\mathcal{H}$ to itself. We call $\mathcal{N}$ the *successor* function and interpret $\mathcal{N}(\mathcal{R})$ as saying what to reject in the next step of the procedure, after having rejected $\mathcal{R}$ in the previous step.

The sequentially rejective procedure based on $\mathcal{N}$ iteratively rejects null hypotheses in the following manner. Let $\mathcal{R}_i \subseteq \mathcal{H}$, $i = 0, 1, \ldots$, be the collection of null hypotheses rejected after step i. The procedure is defined by

$$\begin{aligned} \mathcal{R}_0 &= \varnothing, \\ \mathcal{R}_{i+1} &= \mathcal{R}_i \cup \mathcal{N}(\mathcal{R}_i). \end{aligned} \tag{1}$$

In short, a sequentially rejective procedure is a procedure that sequentially chooses hypotheses to reject, based on the collection of hypotheses that have previously been rejected. Let $\mathcal{R}_\infty = \lim_{i\to\infty} \mathcal{R}_i$ be the final set of rejected null hypotheses. Two simple conditions on $\mathcal{N}$ are sufficient for the procedure (1) to strongly control the familywise error. These are given in Theorem 1.

THEOREM 1 (Sequential rejection principle). *Suppose that for every $\mathcal{R} \subseteq \mathcal{S} \subset \mathcal{H}$, almost surely,*

$$\mathcal{N}(\mathcal{R}) \subseteq \mathcal{N}(\mathcal{S}) \cup \mathcal{S} \tag{2}$$

and that for every $M \in \mathbb{M}$, we have

$$\mathrm{P}_M\big(\mathcal{N}(\mathcal{F}(M)) \subseteq \mathcal{F}(M)\big) \geq 1 - \alpha. \tag{3}$$

Then, for every $M \in \mathbb{M}$,

$$\mathrm{P}_M\big(\mathcal{R}_\infty \subseteq \mathcal{F}(M)\big) \geq 1 - \alpha. \tag{4}$$

A simple outline of the proof of Theorem 1, given below, will give an intuitive explanation of the familywise error control of sequentially rejective multiple testing procedures. On one hand, condition (3), the *single-step condition*, guarantees familywise error control in the "critical case" in which we have rejected all false null hypotheses and none of the true ones. On the other hand, condition (2), the *monotonicity condition*, guarantees that no false rejection in the critical case also implies no false rejection in situations with fewer rejections than in the critical case so that type I error control in the critical case is sufficient for overall familywise error control of the sequential procedure.

PROOF OF THEOREM 1. Choose any $M \in \mathbb{M}$, use the shorthand $\mathcal{T} = \mathcal{T}(M)$, $\mathcal{F} = \mathcal{F}(M) = \mathcal{H} \setminus \mathcal{T}$ and let E be the event $E = \{\mathcal{N}(\mathcal{F}) \subseteq \mathcal{F}\}$. By the single-step condition (3), we have $\mathrm{P}_M(E) \geq 1 - \alpha$. Suppose that the event E is realized. We now use induction to prove that, in this case, $\mathcal{R}_i \subseteq \mathcal{F}$. Obviously, $\mathcal{R}_0 \subseteq \mathcal{F}$. Now, suppose that $\mathcal{R}_i \subseteq \mathcal{F}$. By the monotonicity assumption, we have, almost surely,

$$\mathcal{R}_{i+1} \cap \mathcal{T} = \mathcal{N}(\mathcal{R}_i) \cap \mathcal{T} \subseteq \mathcal{N}(\mathcal{F}) \cap \mathcal{T} = \varnothing.$$

Therefore, E implies that $\mathcal{R}_i \subseteq \mathcal{F}$ for all i. Hence, for all i,

$$\mathrm{P}_M(\mathcal{R}_i \subseteq \mathcal{F}) \geq \mathrm{P}(E) \geq 1 - \alpha.$$

The corresponding result for $\mathcal{R}_\infty$ follows from the dominated convergence theorem. □

A simple and general admissibility criterion can be derived from Theorem 1 in the case of *restricted combinations* [Shaffer (1986)]. Restricted combinations occur if, for some $\mathcal{R} \subseteq \mathcal{H}$, there is no model $M \in \mathbb{M}$ such that $\mathcal{R} = \mathcal{F}(M)$. A standard example concerns testing pairwise equality of means in a one-way ANOVA model: if any single null hypothesis is false, it is not possible that all other null hypotheses are simultaneously true. Let $\Phi = \{\mathcal{F}(M) : M \in \mathbb{M}\}$, the collection of subsets of $\mathcal{H}$ that can actually be a collection of false null hypotheses. For collections $\mathcal{R} \notin \Phi$, the single-step condition sets no restrictions on $\mathcal{N}(\mathcal{R})$, so $\mathcal{N}(\mathcal{R})$ is only constrained by the monotonicity condition. Without loss of familywise error control, we may, therefore, set $\mathcal{N}(\mathcal{R})$ to be the maximal set allowed by monotonicity, setting

$$\mathcal{N}(\mathcal{R}) = \bigcap \{\mathcal{S} \cup \mathcal{N}(\mathcal{S}) : \mathcal{R} \subset \mathcal{S} \in \Phi\} \qquad \text{for every } \mathcal{R} \notin \Phi, \tag{5}$$

interpreting this as $\mathcal{N}(\mathcal{R}) = \mathcal{H}$ if there is no $\mathcal{S} \in \Phi$ for which $\mathcal{R} \subset \mathcal{S}$. Any sequential rejection procedure that does not fulfill (5) is inadmissible and can be uniformly improved by redefining $\mathcal{N}$ such that (5) holds.

2.2. *Using test statistics and critical values.* We generally think of a multiple testing procedure as a procedure that involves test statistics and critical values. To understand the principle, it is helpful to reformulate the principle in such terms, even when that makes it slightly less general. Assume that we have a test statistic $S_H : \Omega \to \mathbb{R}$ for each null hypothesis $H \in \mathcal{H}$, for which large values of S_H indicate evidence against H. In that case, we can construct a successor function $\mathcal{N}$ by choosing a critical value function $\mathbf{c} = \{c_H\}_{H \in \mathcal{H}}$ for which each c_H maps from the power set $2^{\mathcal{H}}$ of all subsets of $\mathcal{H}$ to $\mathbb{R} \cup \{-\infty, \infty\}$. The function $\mathbf{c}$ may be either fixed and chosen in advance before data collection, or it may be random, possibly even depending on the data, as in permutation testing or other resampling-based testing. Choosing

$$\mathcal{N}(\mathcal{R}) = \{H \in \mathcal{H} \setminus \mathcal{R} : s_H \geq c_H(\mathcal{R})\}, \tag{6}$$

the function $c_H(\mathcal{R})$ gives the critical value for hypothesis H after the hypotheses in $\mathcal{R}$ have been rejected. Only the values of $c_H(\mathcal{R})$ for $H \notin \mathcal{R}$ are relevant; the values for $H \in \mathcal{R}$ do not influence the procedure in any way.

The sequentially rejective procedure based on (6) is a sequence of single-step procedures. At each single step, the critical values for all null hypotheses are determined by the set $\mathcal{R}_i$ of rejected null hypotheses in the previous step, or, equivalently, by the set $\mathcal{H} \setminus \mathcal{R}_i$ of remaining hypotheses. After every step, the procedure adjusts the critical values on the basis of the new rejected set.

The *monotonicity* condition (2) for the choice (6) of $\mathcal{N}(\mathcal{R})$ is equivalent to the requirement that for every $\mathcal{R} \subseteq \mathcal{S} \subset \mathcal{H}$ and every $H \in \mathcal{H} \setminus \mathcal{S}$, we have

$$c_H(\mathcal{R}) \geq c_H(\mathcal{S}). \tag{7}$$

In the case where the critical value function $\mathbf{c}$ is random (see Section 5.2), the condition (7) should hold almost surely. The condition requires that as more hypotheses are rejected, the critical values of unrejected null hypotheses never increase, so that, generally, more rejections in previous steps allow reduced critical values in subsequent steps. It follows immediately from condition (7) that for every $H \in \mathcal{H} \setminus \mathcal{R}_i$,

$$c_H(\mathcal{R}_{i+1}) \leq c_H(\mathcal{R}_i), \tag{8}$$

so a sequentially rejective procedure that fulfils the monotonicity condition (7) must have nonincreasing critical values at every step. It is important to realize, however, that the statement (8) is a substantially weaker statement than the condition (7) itself. In fact, as an alternative condition, the statement (8) is too weak to guarantee familywise error control. We show this with a counterexample in Appendix A. This counterexample shows that the condition (7) must also hold for sets $\mathcal{R}$ and $\mathcal{S}$ that can never appear as members of the same sequence $\mathcal{R}_0, \mathcal{R}_1, \ldots$ of sets of rejected hypotheses. Romano and Wolf (2005) also provide an interesting example illustrating the importance of monotonicity in sequentially rejective procedures (Example 6 in that paper).

The *single-step* condition (3) translates to the requirement that for every $\mathcal{R} \subset \mathcal{H}$ and every $M \in \mathbb{M}$ for which $\mathcal{R} = \mathcal{F}(M)$, we have

$$\mathrm{P}_M\Bigg(\bigcup_{H \in \mathcal{T}(M)} \{S_H \geq c_H(\mathcal{R})\} \Bigg) \leq \alpha. \tag{9}$$

The condition (9) requires a limited form of weak familywise error control at each individual step. The most notable feature of this limited form of control is that it is not necessary to control the familywise error for all possible data distributions in $M \in \mathbb{M}$, but only for those distributions for which $\mathcal{R} = \mathcal{F}(M)$. This clause relaxes the required control in two useful ways. On one hand, we may assume that $\mathcal{R} \supseteq \mathcal{F}(M)$, which implies that all nonrejected null hypotheses are true. Therefore, the required familywise error control of condition (9) is no more than weak control.

On the other hand, we may assume that $\mathcal{R} \subseteq \mathcal{F}(M)$, which implies that all rejected hypotheses are false. This latter aspect of the single-step condition is relevant in the case of logical relations or substantial overlap between null hypotheses, and makes it easy to exploit such relationships, for example, in the manner of Shaffer (1986). In the case of restricted combinations, the admissibility condition (5) can be used, which translates to

$$c_H(\mathcal{R}) = \max_{\mathcal{S} \in \Phi: H \notin \mathcal{S}, \mathcal{R} \subset \mathcal{S}} c_H(\mathcal{S}),$$

as a condition on critical values.

Because of the exploitation of relationships between hypotheses, the required control of condition (9) is very limited: it is even weaker than weak control. In this context, it is important to note that the "local test" that is implicit in the single-step condition (9), which rejects if $S_H \geq c_H(\mathcal{R})$ for any $H \in \mathcal{H} \setminus \mathcal{R}$, is not generally a valid local test of the intersection hypothesis $\bigcap(\mathcal{H} \setminus \mathcal{R})$ in the sense of the closed testing procedure. As condition (9) only needs to control the familywise error for those $M \in \mathbb{M}$ for which $\mathcal{T}(M) \cap \mathcal{R} = \varnothing$, that is, only for $M \notin \bigcup \mathcal{R}$, the test only needs to be a valid test for the more restricted hypothesis $\{\bigcap(\mathcal{H} \setminus \mathcal{R})\} \setminus \bigcup \mathcal{R}$. The latter hypothesis is part of the partitioning of $\mathcal{H}$ rather than of its closure (see Section 4). As the single-step condition only needs to control the probability of falsely rejecting this more restricted hypothesis, it has potential for a gain in power over closed testing-based procedures.

As with the monotonicity condition, the single-step condition must hold for every $\mathcal{R}$ for which $\mathcal{R} = \mathcal{F}(M)$ for some $M \in \mathbb{M}$, even if it can never appear as a member of an actual sequence $\mathcal{R}_0, \mathcal{R}_1, \ldots$ of sets of rejected hypotheses.

As a side note, it can be remarked that it is conventional, but not necessary, to use closed rejection sets in (6), rejecting when $S_H \geq c_H(\mathcal{R})$. We may just as well define an analogous sequentially rejective multiple testing procedure based on open rejection sets, defining

$$\mathcal{N}(\mathcal{R}) = \{H \in \mathcal{H} \setminus \mathcal{R} : S_H > k_H(\mathcal{R})\} \tag{10}$$

for some critical value function $\mathbf{k} = \{k_H\}_{H \in \mathcal{H}}$. This open-set-based procedure will be important in Section 5.2.

3. Bonferroni–Shaffer-based methods. There is a large class of sequentially rejective multiple testing procedures that fulfil the single-step condition through an inequality we call the *Bonferroni–Shaffer inequality*: the Bonferroni inequality combined with Shaffer's (1986) treatment of logically related hypotheses. In this section, we review examples of such methods and show that they all conform to the general sequentially rejective multiple testing procedure described in the previous section.

All Bonferroni–Shaffer-based methods start from raw p-values $\{p_H\}_{H \in \mathcal{H}}$ for each hypotheses, which have the property that for every $H \in \mathcal{T}(M)$ and every

$\alpha \in [0, 1]$,

$$\mathrm{P}_M(p_H \le \alpha) \le \alpha. \tag{11}$$

We may define a sequentially rejective multiple testing procedure directly for the raw p-values. Analogous to choosing the critical value function $\mathbf{c}$, choose some function $\boldsymbol{\alpha} = \{\alpha_H\}_{H\in\mathcal{H}}$, for which $\alpha_H : 2^{\mathcal{H}} \to [0, 1]$ for every $H \in \mathcal{H}$, and set

$$\mathcal{N}(\mathcal{R}) = \{H \in \mathcal{H} \setminus \mathcal{R} : p_H \le \alpha_H(\mathcal{R})\}. \tag{12}$$

It will be helpful to restate some of the inequalities of the previous section in terms of $\{p_H\}_{H\in\mathcal{H}}$ and $\boldsymbol{\alpha}(\cdot)$. It follows from Theorem 1 that the procedure based on (12) controls the familywise error if it fulfils the *monotonicity condition* that

$$\alpha_H(\mathcal{R}) \le \alpha_H(\mathcal{S})$$

for every $\mathcal{R} \subseteq \mathcal{S} \subset \mathcal{H}$ and every $H \in \mathcal{H} \setminus \mathcal{S}$, and the *single-step condition* that

$$\mathrm{P}_M\Bigg(\bigcup_{H\in\mathcal{T}(M)} \{p_H \le \alpha_H(\mathcal{R})\}\Bigg) \le \alpha$$

for every $\mathcal{R} \subset \mathcal{H}$ and for every $M \in \mathbb{M}$ for which $\mathcal{R} = \mathcal{F}(M)$.

The Bonferroni–Shaffer-based methods make use of the following inequality in the single-step condition. If $\mathcal{R} \subset \mathcal{H}$ and $\mathcal{T}(M) \cap \mathcal{R} = \varnothing$, we have

$$\mathrm{P}_M\Bigg(\bigcup_{H\in\mathcal{T}(M)} \{p_H \le \alpha_H(\mathcal{R})\}\Bigg) \le \sum_{H\in\mathcal{T}(M)} \alpha_H(\mathcal{R}) \le \sum_{H\in\mathcal{H}\setminus\mathcal{R}} \alpha_H(\mathcal{R}) \tag{13}$$

and we can control the left-hand side by controlling either the right-hand side term (the classical Bonferroni inequality) or the middle term (Shaffer's improvement). The difference between the middle term and the right-hand side term of (13) is important in the case of logical implications between null hypotheses.

Many well-known multiple testing procedures make use of the inequality (13) for their single-step condition. These methods have exact familywise error control if the p-values they are based on conform to (11) exactly and asymptotic control if the p-values conform to (11) asymptotically. We review a number of them briefly below. The methods described in Section 6 and even those in Section 4 can also be seen as Bonferroni–Shaffer-based.

Holm's procedure is explicitly sequential, as the title of his paper (1979) clearly states. Let $|\cdot|$ indicate the cardinality of a set and suppose that $|\mathcal{H}|$ is finite. The critical value function of Holm's procedure is given by

$$\alpha_H(\mathcal{R}) = \frac{\alpha}{|\mathcal{H} \setminus \mathcal{R}|}.$$

The monotonicity condition holds because $|\mathcal{H} \setminus \mathcal{R}| \ge |\mathcal{H} \setminus \mathcal{S}|$ if $\mathcal{R} \subseteq \mathcal{S}$, and the single-step condition follows immediately from the Bonferroni inequality (13). This construction trivially extends to the weighted version of Holm's procedure.

In the case of logical relationships between procedures, we may obtain uniformly more powerful procedures by setting $\alpha_H(\mathcal{R}) = \alpha/|\mathcal{H} \setminus \mathcal{R}|$ for all $\mathcal{R} \in \Psi$, as in Holm's procedure, and use the condition (5) to obtain improved critical values for $\mathcal{R} \notin \Psi$. We set

$$\alpha_H(\mathcal{R}) = \min_{\mathcal{S} \in \Phi: H \notin \mathcal{S}, \mathcal{R} \subset \mathcal{S}} \alpha_H(\mathcal{S})$$

for all $\mathcal{R} \notin \Psi$, which results in the critical value function

$$\alpha_H(\mathcal{R}) = \min_{M \in H: \mathcal{T}(M) \cap \mathcal{R} = \varnothing} \frac{\alpha}{|\mathcal{T}(M)|}.$$

This is the so-called "P3" procedure of Hommel and Bernhard (1999). This procedure is a uniform improvement over the earlier "S2" procedure of Shaffer (1986), which has critical value function

$$\alpha_H(\mathcal{R}) = \min_{M: \mathcal{T}(M) \cap \mathcal{R} = \varnothing} \frac{\alpha}{|\mathcal{T}(M)|}. \tag{14}$$

Shaffer's procedure may be obtained by taking

$$\alpha_H(\mathcal{R}) = \min_{\mathcal{S} \in \Phi: \mathcal{R} \subset \mathcal{S}} \frac{\alpha}{|\mathcal{H} \setminus \mathcal{R}|}$$

for all $\mathcal{R} \notin \Psi$, using a weaker version of condition (5). The monotonicity and single-step conditions for the "S2" and "P3" procedures may also be checked directly from Theorem 1. Monotonicity is trivial and single-step control follows immediately from the Bonferroni–Shaffer inequality (13).

4. Closed testing and partitioning. The general closed testing [Marcus, Peritz and Gabriel (1976)] and partitioning procedures [Finner and Strassburger (2002)] are fundamental principles of multiple testing in their own right. Still, as we shall show in this section, even in their most general formulation, both principles can be derived as special cases of the sequentially rejective procedure and the Bonferroni–Shaffer inequality, provided that the collection of hypotheses $\mathcal{H}$ is extended to include the closure or the partitioning of these hypotheses, respectively.

Even though we can view closed testing and partitioning as special cases of sequential Bonferroni–Shaffer methods in this way, the procedures are different from the Bonferroni–Shaffer-based procedures described earlier. They ensure that, before a false rejection has been made, there is never more than one true null hypothesis $H \in T(M)$ that has $\alpha_H(\mathcal{R}) > 0$. Therefore, they control the sum in (13) through the number of terms, rather than through their magnitude. This makes closed testing and partitioning less conservative than some other methods, which is illustrated by the fact that, unlike most Bonferroni–Shaffer-based procedures, these methods never give multiplicity-adjusted p-values (see Section 7) that are exactly 1 unless there are raw p-values which are exactly 1.

It is interesting to note that the relationship described in this section between closed testing and partitioning on the one hand, and sequential methods on the

other, is reversed relative to the traditional one. It has often been observed that sequential methods such as Holm's can be derived as special cases of closed testing or partitioning. Here, we show, conversely, that closed testing and partitioning procedures, in their most general forms, can be derived as special cases of sequential rejection methods.

4.1. *Closed testing*. The closed testing procedure was formulated by Marcus, Peritz and Gabriel (1976). It requires that the set $\mathcal{H}$ of null hypotheses be closed with respect to intersection, that is, for every $H \in \mathcal{H}$ and $J \in \mathcal{H}$, we must have $H \cap J \in \mathcal{H}$, unless $H \cap J = \varnothing$. If necessary, the set $\mathcal{H}$ may be recursively extended to include all nonempty intersection hypotheses. Define $i(H) = \{J \in \mathcal{H} : J \subset H\}$ as the set of all implying null hypotheses of H.

We consider the most general form of the closed testing procedure here, placing no restrictions on the local test statistic S_H used to obtain the marginal p-values p_H for each intersection hypothesis $H \in \mathcal{H}$. The closed testing procedure is sequential. It starts by testing all hypotheses which have no implying null hypotheses in $\mathcal{H}$ (typically, this is only $\bigcap \mathcal{H}$, the intersection of all null hypotheses). If at least one of these hypotheses is rejected, then the procedure continues to test all null hypotheses for which all implying null hypotheses have been rejected, until no more rejection occurs. All tests are done at level α. In terms of the general sequentially rejective procedure, the critical value function is given by

$$\alpha_H(\mathcal{R}) = \begin{cases} \alpha, & \text{if } i(H) \subseteq \mathcal{R}, \\ 0, & \text{otherwise.} \end{cases}$$

The closed testing procedure conforms to the conditions of Theorem 1. The monotonicity is immediate from the definition of the critical value function. The single-step condition follows from the Shaffer inequality (13) in the following way. Assume that $\mathcal{R} \cap \mathcal{T}(M) = \varnothing$. Consider $T = \bigcap \mathcal{T}(M)$, the intersection of all true null hypotheses. As $M \in T$, T is not empty and, by the closure assumption, $T \in \mathcal{H}$ and even $T \in \mathcal{T}(M)$, so $T \notin \mathcal{R}$. For every $H \in \mathcal{T}(M)$ for which $H \neq T$, we have $T \in i(H)$ and therefore $i(H) \not\subseteq \mathcal{R}$. Hence,

$$\sum_{H \in \mathcal{T}(M)} \alpha_H(\mathcal{R}) \leq \alpha_T(\mathcal{R}) \leq \alpha,$$

which proves the single-step condition.

The practical value of this construction is algorithmic. The sequentially rejective view of closed testing emphasizes that it is not usually required to calculate all intersection hypotheses tests, but only those for which all implying hypothesis have been rejected in previous steps. At the cost of some bookkeeping, this may greatly reduce the number of tests which must be performed.

4.2. *Partitioning.* The closed testing principle of Marcus, Peritz and Gabriel (1976) has been a cornerstone of multiple hypotheses testing for decades. However, Stefansson, Kim and Hsu (1988) introduced what is now called the *partitioning principle*, and Finner and Strassburger (2002) showed that the partitioning principle gives a multiple testing procedure that is at least as powerful as closed testing and which may be more powerful in some situations.

The main idea is to partition the union of the hypotheses of interest into disjoint sub-hypotheses such that each hypothesis can be represented as the disjoint union of some of them. We refer to the collection of these sub-hypotheses as the partitioning $\mathcal{P}$ and include it in $\mathcal{H}$. Formally, we assume that $\mathcal{P} \subseteq \mathcal{H}$, where $\mathcal{P}$ is such that for any J and K in $\mathcal{P}$ with $J \neq K$, we have $J \cap K = \varnothing$, and for each $H \in \mathcal{H}$, $H = \bigcup \mathcal{K}$ for some $\mathcal{K} \subseteq \mathcal{P}$. The set $\mathcal{H}$ may have to be extended by its partitioning to make $\mathcal{P} \subseteq \mathcal{H}$ hold true.

As in closed testing above, we put no restrictions on the test statistics used; however, the procedure only actually uses the marginal p-values p_H for $H \in \mathcal{P}$. In terms of the general sequentially rejective procedure, the critical value function for the hypotheses in $\mathcal{H} \setminus \mathcal{R}$ is given by

$$\alpha_H(\mathcal{R}) = \begin{cases} \alpha, & \text{if } H \in \mathcal{P}, \\ 1, & \text{if } H \in \mathcal{H} \setminus \mathcal{P} \text{ and } H \subseteq \bigcup \mathcal{R}, \\ 0, & \text{otherwise.} \end{cases}$$

As a sequentially rejective procedure, the partitioning procedure never requires more than two steps. In the first step, the procedure rejects only those hypotheses that are part of the partitioning and in the second, it rejects any hypotheses implied by the union of the rejected partitioning hypotheses.

To prove familywise error control through the sequential rejection principle, we check the monotonicity and single-step conditions. Monotonicity is trivial. Let $\mathcal{T}(M) \cap \mathcal{R} = \varnothing$. The single-step condition follows trivially from the Shaffer inequality (13) by writing

$$\sum_{H \in \mathcal{T}(M)} \alpha_H(\mathcal{R}) = \sum_{H \in \mathcal{T}(M) \cap \mathcal{P}} \alpha_H(\mathcal{R}) + \sum_{H \in \mathcal{T}(M) \setminus \mathcal{P}} \alpha_H(\mathcal{R}). \tag{15}$$

We have $|\mathcal{T}(M) \cap \mathcal{P}| \leq 1$ because the hypotheses in $\mathcal{P}$ are disjoint, and $\alpha_H(\mathcal{R}) = 0$ for every $H \in \mathcal{T}(M) \setminus \mathcal{P}$ since $H \in \mathcal{T}(M)$ with $\mathcal{T}(M) \cap \mathcal{R} = \varnothing$ implies $H \not\subseteq \bigcup \mathcal{R}$. The right-hand side of (15) is therefore bounded by α.

As for the relationship between sequential rejection and partitioning, it is interesting to remark that it is possible to construct an alternative proof of Theorem 1 that constructs the sequential rejection principle as a partitioning procedure with shortcuts [see Calian, Li and Hsu (2008) for the definition of shortcuts in the partitioning procedure]. Combined with the result of this subsection, this suggests an interesting duality between sequential rejection and partitioning: sequential rejection is partitioning with shortcuts, while partitioning is sequential rejection based on an augmented collection of hypotheses. The same alternative construction of

sequential rejection based on shortcuts also makes it easier to compare the sequential rejection principle with earlier treatments of sequential testing, such as that of Hommel, Bretz and Maurer (2007), which are constructed using shortcuts in the closed testing procedure. In contrast to these methods, the sequential rejection procedure can exploit some of the additional power of partitioning relative to closed testing [Finner and Strassburger (2002)], especially in the case of logical relationships or overlap between hypotheses.

A simple example may serve to illustrate the relationships between partitioning, closed testing and sequential rejection. Let $\Delta > 0$ and let $H_1 : \theta \leq \Delta$, $H_2 : \theta \geq -\Delta$ and $H_{12} = H_1 \cap H_2$ be the three hypotheses of interest. Closed testing would start by testing H_{12} at level α and proceed to test H_1 and H_2, both at α, once H_{12} is rejected. Sequential rejection may similarly start testing H_{12} at level α. After rejection of H_{12} at the second step, however, it may assume, for all subsequent tests, that H_{12} is truly false. As a consequence, it may simultaneously test H_1 using a test for $H_1' : \theta < -\Delta$ and H_2 using a test for $H_2' : \theta > \Delta$, performing both tests at level α because H_1' and H_2' are disjoint. The latter tests may be more powerful than the original tests for H_1 and H_2. Note that the partitioning procedure would start immediately by constructing H_1' and H_2', and would come to exactly the same qualitative conclusion regarding the hypotheses of interest as the sequential rejection principle.

5. Non-Bonferroni-based methods. The Bonferroni–Shaffer inequality allows control of the familywise error with only assumptions on the marginal distribution of each test statistic and no additional assumptions on their joint distribution. Implicitly, the methods based on that inequality (except closed testing and partitioning) assume the worst possible joint distribution for familywise error control, which is the distribution for which all rejection regions are disjoint. If the joint distribution is more favorable, Bonferroni–Shaffer-based methods may be conservative, controlling the familywise error at a level lower than the nominal α level. Improved results may be obtained for distributions that are more favorable than the worst case of Bonferroni–Shaffer, but only at the cost of additional assumptions.

The sequential rejection principle is not limited to methods based on the Bonferroni–Shaffer inequality (13), but may also be used in combination with other methods to ensure the single-step familywise error condition.

5.1. *Šidák's inequality.* For example, we may be willing to assume that Šidák's (1967) inequality,

$$\mathrm{P}_M\Bigg(\bigcap_{H\in\mathcal{T}(M)} S_H \leq s_H\Bigg) \geq \prod_{H\in\mathcal{T}(M)} \mathrm{P}_M(S_H \leq s_H),$$

holds for every $M \in \mathbb{M}$ and for all constants $\{s_H\}_{H\in\mathcal{H}}$, as it does for test statistics independent under the null. In that case, it is possible to define a sequentially

rejective procedure based on the critical value function

$$\alpha_H(\mathcal{R}) = 1 - (1-\alpha)^{1/|\mathcal{H}\setminus\mathcal{R}|}$$

for the raw p-values $\{p_H\}_{H\in\mathcal{H}}$ based on the test statistics $\{S_H\}_{H\in\mathcal{H}}$. This is the step-down Šidák procedure [Holland and DiPinzio Copenhaver (1987)]. Its familywise error control can be proven from Theorem 1 using Šidák's inequality.

5.2. *Resampling-based multiple testing.* A completely different approach to avoiding the conservativeness associated with the Bonferroni–Shaffer inequality is to use resampling techniques to let the multiple testing procedure estimate or accommodate the actual dependence structure between the test statistics.

Well-known resampling-based multiple testing procedures use the fact that the single-step and monotonicity conditions can both be kept by taking $c_H(\mathcal{R})$ as the maximum over $M \in \mathbb{M}$ of the $1-\alpha$ quantile of the distribution of $\max_{H\in\mathcal{T}(M)} S_H$ [Romano and Wolf (2005)]. This quantile is usually unknown, but it may be estimated by resampling methods, provided we are willing to make additional assumptions. Westfall and Young (1993) assume *subset pivotality*, which asserts that for every $M \in \mathbb{M}$, there is some $N \in \bigcap\mathcal{H}$ such that the distribution of $\max_{H\in\mathcal{T}(M)} S_H$ is identical under P_M and P_N. Under this assumption, resampling of $\{S_H\}_{H\in\mathcal{H}\setminus\mathcal{R}}$ under the complete null hypothesis, using permutations or the bootstrap, can give consistent estimates of the desired quantiles. The subset pivotality condition has been the subject of some discussion [Dudoit, Van der Laan and Pollard (2004), Westfall and Troendle (2008)] and several authors have given alternative assumptions that allow estimation of the quantiles of interest [Romano and Wolf (2005), Dudoit and Van der Laan (2008)]. Whatever the underlying assumptions, consistent estimation of the quantiles of $\max_{H\in\mathcal{T}(M)} S_H$ only guarantees control of the familywise error in an asymptotic sense. Asymptotic control of the familywise error is beyond the scope of this article.

We focus instead on resampling-based methods with exact familywise error control, putting these into the framework of the sequential rejection principle. Following Romano and Wolf (2005), we may obtain exact control by generalizing the treatment of permutation testing in Lehmann and Romano [(2005), Chapter 15] to a multiple testing procedure. This method does not explicitly strive to estimate the quantiles of the distribution of $\max_{H\in\mathcal{T}(M)} S_H$.

To define a resampling-based sequentially rejective multiple testing procedure with exact familywise error control, we choose a set $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_r\}$ of r functions that we shall refer to as "null-invariant transformations," or *null-invariants*, each of which is a bijection from the outcome space Ω onto itself. For everything to be well defined, we must assume that the null-invariants map every measurable set onto a measurable set, but we will not concern ourselves with such technicalities here. Using the null-invariants, we can define transformed test statistics $S_H \circ \pi_i$ for every $H \in \mathcal{H}$ and $i \in \{1, \ldots, r\}$. The name *null-invariants* for the transformations

$\boldsymbol{\pi}$ comes from assumption (17) below, which says that transformation of S_H by π_i does not change the distribution of S_H if H is a true null hypothesis.

For the sake of concreteness, we give some motivating examples of null-invariant transformations which fulfill the universal null-invariance condition (17). Let $\stackrel{\mathrm{d}}{=}$ denote equality in distribution. As a first example, in a one-sample situation, suppose that for n i.i.d. subjects, we have sampled a p-dimensional vector $\mathbf{X} = \{X_i\}_{i=1}^p$, symmetrically distributed around a vector $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^p$, that is,

$$\mathbf{X} - \boldsymbol{\theta} \stackrel{\mathrm{d}}{=} \boldsymbol{\theta} - \mathbf{X}.$$

If we want to test $H_i : \theta_i = 0$ for $i = 1, \ldots, p$ with Student T- or Wilcoxon statistics, then all 2^n transformations that map the measured $\mathbf{X}$ to $-\mathbf{X}$ for a subset of the n subjects are null-invariant transformations. Secondly, in a two-sample situation, suppose that we have an i.i.d. sample of size n from a p-dimensional vector $\mathbf{X} = \{X_i\}_{i=1}^p$ and an independent i.i.d. sample of size m from a p-dimensional vector $\mathbf{Y} = \{Y_i\}_{i=1}^p$, and that

$$\mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{Y} + \boldsymbol{\theta}$$

for some vector $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^p$. If we want to test $H_i : \theta_i = 0$ for $i = 1, \ldots, p$ with Student T- or Mann–Whitney statistics, the usual permutations of the group labels are all null-invariant transformations.

The sequentially rejective procedure based on $\boldsymbol{\pi}$ will be defined as follows. Let $s = r - [\alpha r]$, where $[\alpha r]$ is the largest integer that is at most equal to αr. For any test statistic S, define the random variable $(S \circ \boldsymbol{\pi})_{(s)}$ as the sth smallest value among $S \circ \boldsymbol{\pi} = \{S \circ \pi_i\}_{i=1}^r$. It is convenient to define the sequentially rejective multiple testing procedure based on the null-invariants $\boldsymbol{\pi}$ using the open rejection set variant (10) of the general procedure. The critical value function is given by

$$k_H(\mathcal{R}) = \Big(\max_{J \in \mathcal{H} \setminus \mathcal{R}} S_J \circ \boldsymbol{\pi} \Big)_{(s)}. \tag{16}$$

Note that, in contrast with all procedures described above, the critical values $k_H(\mathcal{R})$ are random variables. The notation for the critical values in (16) is suggestive of the algorithm for permutation-based multiple testing of Ge, Dudoit and Speed (2003).

The familywise error control of the procedure based on (16) can be proven with the open rejection set version of the sequential rejection principle of Theorem 1, although not without additional assumptions. The monotonicity of the critical values for every outcome $\omega \in \Omega$ is immediate from the definition. To prove the single-step condition, we use Theorem 2, adapted from Theorem 15.2.1 of Lehmann and Romano (2005), which considers testing of a single hypothesis. The proof of the theorem is given in Appendix B.

THEOREM 2. *Suppose that the transformations* $\{\pi_1, \ldots, \pi_r\}$ *form a group in the algebraic sense. Also, suppose that for every* $M \in \mathbb{M}$ *and for every* $i \in \{1, \ldots, r\}$,

$$\{S_H \circ \boldsymbol{\pi}\}_{H \in \mathcal{T}(M)} \stackrel{\mathrm{d}}{=} \{S_H \circ \boldsymbol{\pi} \circ \pi_i\}_{H \in \mathcal{T}(M)}, \tag{17}$$

where $\stackrel{\mathrm{d}}{=}$ *denotes equality in (joint) distribution. Then, for every* $M \in \mathbb{M}$,

$$\mathrm{P}_M\Big(\bigcup_{H \in \mathcal{T}(M)} \{S_H > k_H(\mathcal{H} \setminus \mathcal{T}(M))\} \Big) \leq \alpha. \tag{18}$$

The condition of Theorem 2 that the transformations $\boldsymbol{\pi}$ form an algebraic group is not very stringent. The typical null-invariant transformations, such as permutations, that are frequently used in hypothesis testing usually meet this requirement. Instead of the complete group $\{\pi_1, \ldots, \pi_r\}$, we may also take a random sample (with or without replacement) from the group. It is easy to verify in the proof of Theorem 2 that, in that case, the result (18) of the theorem holds in expectation over the sampling distribution.

The other condition, (17), which we call the *universal null-invariance condition*, is more crucial. This condition requires that the joint distribution of the test statistics of the true null hypotheses and their transformations by $\boldsymbol{\pi}$ is not altered by another application of a transformation in $\boldsymbol{\pi}$. This motivates the naming of the transformations as "null-invariants." The condition is a generalization of the *randomization hypothesis* for single hypothesis tests [Lehmann and Romano (2005), page 633], which says that under the null hypothesis, the distribution of the data is not affected by the transformations in $\boldsymbol{\pi}$. In many situations, a practical way to check the condition (17) is to check the randomization hypothesis for the subset of the data that is used for the calculation of $\{S_H\}_{H \in \mathcal{T}(M)}$.

The crucial "universal" part of the universal null-invariance condition is that the set of null-invariants $\{\pi_1, \ldots, \pi_r\}$ is not allowed to depend on H, M or $\mathcal{R}$: the same set of transformations must be null-invariant for the joint distribution of all true null hypotheses in every model M and for every step of the procedure.

5.3. *Step-up methods*. Sequential rejection is immediately associated with step-down methods, and several of the methods we have so far considered (Holm, Šidák, resampling-based multiple testing) are of the traditional step-down category. However, the sequential rejection principle is not limited to applications within this class of methods, but may also be used to good effect in combination with methods in the step-up category. Step-up methods are usually presented as methods that sequentially accept hypotheses, rather than sequentially rejecting them. We present an alternative, sequentially rejective view of step-up methods, as follows. Suppose that for every $\mathcal{R} \subset \mathcal{H}$, we choose a sequence of ordered critical values

$$\alpha_1(\mathcal{R}) \geq \cdots \geq \alpha_{|\mathcal{H} \setminus \mathcal{R}|}(\mathcal{R}). \tag{19}$$

Now, we can define a sequentially rejective procedure by setting

$$\mathcal{N}(\mathcal{R}) = \bigcup \{\mathcal{K} \subseteq \mathcal{H} \setminus \mathcal{R} : p_H \leq \alpha_{|\mathcal{H}\setminus(\mathcal{R}\cup\mathcal{K})|+1}(\mathcal{R}) \text{ for every } H \in \mathcal{K}\}. \tag{20}$$

This function says that after having rejected a collection of hypotheses $\mathcal{R}$, we proceed in the next step of the procedure to reject all hypotheses in $\mathcal{H} \setminus \mathcal{R}$ with p-value smaller than $\alpha_k(\mathcal{R})$ whenever there are at least $|\mathcal{H} \setminus \mathcal{R}| - k + 1$ of those, and it does this for $k = 1, \ldots, |\mathcal{H} \setminus \mathcal{R}|$ simultaneously. Equivalently, in terms of ordered p-values, for $k = 1, \ldots, |\mathcal{H} \setminus \mathcal{R}|$, it rejects the hypothesis with the kth largest p-value if that p-value is smaller than $\alpha_k(\mathcal{R})$ and rejects every null hypothesis with a p-value smaller than that of any rejected hypothesis.

By Theorem 1, the sequential rejection procedure based on (20) controls the familywise error if it fulfills the monotonicity and single-step conditions. We have summarized this result in the following corollary.

COROLLARY 1. *If, whenever $\mathcal{R} \subseteq \mathcal{S}$, for $i = 1, \ldots, |\mathcal{H} \setminus \mathcal{S}|$, we have*

$$\alpha_i(\mathcal{S}) \geq \alpha_i(\mathcal{R}) \tag{21}$$

and for every $M \in \mathbb{M}$, whenever $\mathcal{R} = \mathcal{F}(M)$, we have

$$\mathrm{P}_M\Bigg(\bigcup_{\mathcal{K} \subseteq \mathcal{T}(M)} \{p_H \leq \alpha_{|\mathcal{T}(M)\setminus\mathcal{K}|+1}(\mathcal{F}(M)) \text{ for every } H \in \mathcal{K}\} \Bigg) \leq \alpha, \tag{22}$$

then the sequentially rejective procedure based on (20) satisfies

$$\mathrm{P}_M\big(\mathcal{R}_\infty \subseteq \mathcal{F}(M)\big) \geq 1 - \alpha.$$

Procedures based on Corollary 1 can be seen as having an inner and an outer loop. The inner loop performs familiar step-up testing; the outer loop recalibrates the critical values of the step-up procedure based on rejections in the inner loop.

PROOF OF COROLLARY 1. We prove the corollary by checking the conditions of Theorem 1. The single step condition (3) is immediate from (22). We proceed to check monotonicity (2). Choose some $\mathcal{R} \subseteq \mathcal{S} \subset \mathcal{H}$. Monotonicity holds trivially if $\mathcal{N}(\mathcal{R}) = \varnothing$, so we may suppose that $\mathcal{N}(\mathcal{R}) \neq \varnothing$. Let $H \in \mathcal{N}(\mathcal{R})$: we have to show that H belongs either to $\mathcal{N}(\mathcal{S})$ or to $\mathcal{S}$. By definition of $\mathcal{N}(\mathcal{R})$, there is some $\mathcal{K} \subseteq \mathcal{H} \setminus \mathcal{R}$ such that $H \in \mathcal{K}$ and $p_J \leq \alpha_{|\mathcal{H}\setminus(\mathcal{R}\cup\mathcal{K})|+1}(\mathcal{R})$ for every $J \in \mathcal{K}$. By (19) and (21), we have

$$\alpha_{|\mathcal{H}\setminus(\mathcal{R}\cup\mathcal{K})|+1}(\mathcal{R}) \leq \alpha_{|\mathcal{H}\setminus(\mathcal{R}\cup\mathcal{K})|+1}(\mathcal{S}) \leq \alpha_{|\mathcal{H}\setminus(\mathcal{S}\cup\mathcal{K})|+1}(\mathcal{S}).$$

We have either $H \in \mathcal{S}$ or $H \notin \mathcal{S}$. When $H \notin \mathcal{S}$, we have $H \in \mathcal{K} \setminus \mathcal{S} = \tilde{\mathcal{K}}$. Then $\tilde{\mathcal{K}} \subseteq \mathcal{H} \setminus \mathcal{S}$ is such that $p_J \leq \alpha_{|\mathcal{H}\setminus(\mathcal{S}\cup\mathcal{K})|+1}(\mathcal{S}) = \alpha_{|\mathcal{H}\setminus(\mathcal{S}\cup\tilde{\mathcal{K}})|+1}(\mathcal{S})$ for every $J \in \tilde{\mathcal{K}}$, thus $H \in \mathcal{N}(\mathcal{S})$. $\square$

The simplest nonsequential application of Corollary 1 is the method of Hochberg (1988). This method assumes that the inequality of Simes (1986) holds for the collection of true null hypotheses $\mathcal{T}(M)$ so that

$$\mathrm{P}_M\Bigg(\bigcup_{\mathcal{K}\subseteq\mathcal{T}(M)}\Bigg\{\bigcap_{H\in\mathcal{K}}\Bigg\{p_H \le \frac{\alpha\cdot|\mathcal{K}|}{|\mathcal{T}(M)|}\Bigg\}\Bigg\}\Bigg) \le \alpha.$$

This inequality holds for independent test, but also under some types of dependence [Sarkar (1998)]. In Hochberg's method, $\alpha_i(\mathcal{R}) = \alpha/i$ for every $\mathcal{R}$. Monotonicity is straightforward for this method, and the single-step condition (22) follows immediately from Simes' inequality because

$$\frac{\alpha}{|\mathcal{T}(M)\setminus\mathcal{K}|+1} \le \frac{\alpha\cdot|\mathcal{K}|}{|\mathcal{T}(M)|}$$

if $\mathcal{K}\subseteq\mathcal{T}(M)$.

The value of the embedding of Hochberg's method into the sequential rejection framework is most obvious when we consider logically related hypotheses. Hommel (1988) already remarked that if it is known a priori that $|\mathcal{T}(M)| \le k < |\mathcal{H}|$, then the critical values of Hochberg's method can be relaxed to $\alpha_i(\mathcal{R}) = \alpha/\min(i,k)$. This can be easily seen from the condition (22) by realizing that this condition does not involve $\alpha_i(\mathcal{R})$ for $i > |\mathcal{T}(M)|$, so this value can be chosen freely. Such a relaxation of the critical values is particularly useful if the step-up procedure is embedded in a further sequentially rejective procedure, for example, in the case of three hypotheses, one that first tests a global null hypothesis $H_1\cap H_2\cap H_3$ at level α before testing H_1, H_2 and H_3 in a step-up fashion. By the sequential rejection principle, such a test procedure may proceed at the second stage, assuming that the global null hypothesis is false.

Truly sequential results may be obtained in other situations with restricted combinations [Hochberg and Rom (1995)] if we let the critical values of the step-up procedure depend on the set of previous rejections. We can define a step-up analogy to Shaffer's S2 method (14), defining

$$\alpha_i(\mathcal{R}) = \frac{\alpha}{\min(i,\max\{|\mathcal{T}(M)|:\mathcal{T}(M)\cap\mathcal{R}=\varnothing\})}. \tag{23}$$

Strong control for this method follows from Corollary 1. Monotonicity for this method is trivial and the single-step condition still follows immediately from Simes' inequality.

We give two simple examples with restricted combinations in which this method is more powerful than the regular method of Hochberg. First, consider the case of testing all pairwise comparisons and take the situation with three hypotheses $H_{12}:\mu_1=\mu_2$, $H_{23}:\mu_2=\mu_3$ and $H_{13}:\mu_1=\mu_3$ as an example. In this case, $|\mathcal{T}(M)|$ can only take the values 0, 1 or 3. The test statistics may conform to Simes inequality, for example, if the data for each null hypothesis come from independent studies. Hochberg's procedure would reject if all three hypotheses have p-values

at most α, if any two hypotheses have p-values at most $\alpha/2$ or if any single hypothesis has p-value at most $\alpha/3$. The sequentially rejective step-up procedure defined in (23) would, if Hochberg's procedure would have made only a single rejection, additionally reject one of the remaining hypotheses if either of them had a p-value of at most α. Second, consider testing the three hypotheses $H_1 : \mu_1 \leq 0$, $H_2 : \mu_2 \leq 0$ and $H_3 : \mu_1 + \mu_2 \leq 0$. If the respective test statistics T_1 and T_2 for H_1 and H_2 are independent and normally distributed, and we use $T_3 = T_1 + T_2$ as test statistic for H_3, then the test statistics conform to the conditions of Sarkar (1998) so that Simes' inequality may be used. Note that falsehood of H_3 implies falsehood of at least one of H_1 and H_2. Therefore, if the p-value of H_3 would be below $\alpha/3$ and the p-value of one of H_1 or H_2 would be between $\alpha/2$ and α, but the p-value of the other would be above α, then the sequential method based on (23) would reject two hypotheses, whereas Hochberg's procedure would reject only one.

6. Gatekeeping and graph-based testing. Multiple testing methods may also be used in a situation in which hypotheses are not exchangeable, but where interest in one hypothesis is conditional on the rejection of other hypotheses. This is an area of extensive recent interest, both in clinical trials and in genomics research. In this section, we review gatekeeping and graph-based testing procedures, and demonstrate how the sequential rejection principle may be applied to uniformly improve upon existing methods.

6.1. *Gatekeeping*. Gatekeeping strategies [see Dmitrienko and Tamhane (2007) for an overview] are popular in clinical trials, in which often multiple primary, secondary and possibly tertiary endpoints are considered. In a gatekeeping strategy, the null hypotheses in $\mathcal{H}$ are divided into k families, $\mathcal{G}_1, \ldots, \mathcal{G}_k$, each $\mathcal{G}_i \subset \mathcal{H}$. Hypotheses in a family $\mathcal{G}_{i+1}$ are not tested before at least one hypothesis in the family $\mathcal{G}_i$ has been rejected.

Gatekeeping strategies are sequential in a very natural way [Dmitrienko et al. (2006)] and they are easily fitted into the framework of the general sequentially rejective procedure. We illustrate this for the basic unweighted serial [Westfall and Krishen (2001)] and parallel [Dmitrienko, Offen and Westfall (2003)] gatekeeping strategies for two families, $\mathcal{G}_1$ and $\mathcal{G}_2$.

The standard serial gatekeeping procedure uses Holm in the first family and Holm in the second family, testing the second family only after the first has been completely rejected. It can be defined as a sequentially rejective procedure with the critical value function

$$\alpha_H(\mathcal{R}) = \begin{cases} \alpha/|\mathcal{G}_1 \setminus \mathcal{R}|, & \text{if } H \in \mathcal{G}_1, \\ \alpha/|\mathcal{G}_2 \setminus \mathcal{R}|, & \text{if } H \in \mathcal{G}_2 \text{ and } \mathcal{G}_1 \subseteq \mathcal{R}, \\ 0, & \text{otherwise.} \end{cases}$$

Both the monotonicity and single-step conditions are trivially checked for this procedure.

The usual parallel gatekeeping procedure for two families $\mathcal{G}_1$ and $\mathcal{G}_2$ uses Bonferroni in the first family and Holm in the second. It starts testing the second family whenever at least one hypothesis in the first family has been rejected, but tests the second family at a reduced level if not all hypotheses in $\mathcal{G}_1$ have been rejected. This procedure can be defined with the critical value function

$$\alpha_H(\mathcal{R}) = \begin{cases} \dfrac{\alpha}{|\mathcal{G}_1|}, & \text{if } H \in \mathcal{G}_1, \\ \dfrac{\alpha \cdot |\mathcal{R} \cap \mathcal{G}_1|}{|\mathcal{G}_2 \setminus \mathcal{R}| \cdot |\mathcal{G}_1|}, & \text{if } H \in \mathcal{G}_2. \end{cases}$$

Monotonicity is again trivial. The single-step condition follows from the Bonferroni inequality (13), writing

$$\sum_{H \in \mathcal{H} \setminus \mathcal{R}} \alpha_H(\mathcal{R}) = \sum_{H \in \mathcal{G}_1 \setminus \mathcal{R}} \frac{\alpha}{|\mathcal{G}_1|} + \sum_{H \in \mathcal{G}_2 \setminus \mathcal{R}} \frac{\alpha \cdot |\mathcal{R} \cap \mathcal{G}_1|}{|\mathcal{G}_2 \setminus \mathcal{R}| \cdot |\mathcal{G}_1|} \le \alpha.$$

It is clear from this equation that there is the potential for a gain in power for the procedure in the situation where $\mathcal{G}_2 \subset \mathcal{R}$ and $\mathcal{G}_1 \not\subseteq \mathcal{R}$ because, in that case, the inequality on the right-hand side is a strict inequality. We may set $\alpha_H(\mathcal{R}) = \alpha/|\mathcal{G}_1 \setminus \mathcal{R}|$ for $H \in \mathcal{G}_1$ if $\mathcal{G}_2 \subset \mathcal{R}$ without losing the single-step condition. This has also been noted by Guilbaud (2007).

Versions of the gatekeeping procedure for more than two families, as well as weighted versions, are easily formulated within the sequential rejection framework and the conditions of Theorem 1 are easy to check. The same holds for the many recent extensions and variants of gatekeeping [Dmitrienko et al. (2007), Edwards and Madsen (2007), Guilbaud (2007), Dmitrienko, Tamhane and Wiens (2008)]. Earlier generalizations of the class of gatekeeping procedures, such as that of Hommel, Bretz and Maurer (2007), did not include the case of logically related hypotheses, such as are present, for example, in the procedure of Edwards and Madsen (2007).

6.2. *Graph-based procedures.* Our main motivation for the development of the sequential rejection principle has been our interest in the development of multiple testing procedures for graph-structured hypotheses. Multiple testing in graphs is a subject of great interest, both for applications in clinical trials and in genomics. Specific procedures for controlling the familywise error for graph-structured hypotheses have been proposed by several authors. Examples include the fallback procedure [Wiens and Dmitrienko (2005)], which redistributes the alpha allocated to rejected hypotheses to neighboring hypotheses, the method of Meinshausen (2008), which sequentially tests hypotheses ordered in a hierarchical clustering graph, the focus level method [Goeman and Mansmann (2008)], which combines Holm's procedure with closed testing for hypotheses in a partially closed directed acyclic graph, and the method of Rosenbaum (2008), which sequentially tests ordered hypotheses. All of these methods can be formulated as special cases of the

sequentially rejective multiple testing procedure (12) that control the familywise error with Theorem 1 and the Bonferroni–Shaffer inequality (13).

Several authors [Dmitrienko et al. (2007), Hommel, Bretz and Maurer (2007), Bretz et al. (2009), Burman, Sonesson and Guilbaud (2009)] have proposed general methods for recycling the alpha in graph-structured hypothesis testing, using very general graph structures. These methods can be seen as special cases of the sequential rejection principle, all basing their single-step condition on the weaker right-hand side inequality of (13). In particular, we mention the powerful graphical approaches of Bretz et al. (2009) and Burman, Sonesson and Guilbaud (2009), which are very easy to interpret and communicate, and are flexible enough to cover diverse methods such as gatekeeping, fixed sequence and fallback procedures. The authors of these papers structure the tests in gatekeeping procedures in a directed graph with weighted edges. An initial distribution of alpha is chosen and, once a hypothesis is rejected, the alpha allocated to the rejected hypothesis is redistributed according to the graph. The graphical visualization of the testing procedure increases the understanding of how a testing strategy works and is a useful tool for developing, as well as communicating, procedures. However, these methods cannot make use of logical relationships between hypotheses and, as such, do not incorporate graph-based methods which exploit such relationships, such as those of Edwards and Madsen (2007), Goeman and Mansmann (2008) and Meinshausen (2008).

6.3. *Testing in trees.* To illustrate the ease with which multiple testing procedures in graphs can be formulated and improved, we consider the case of the tree-based method of Meinshausen (2008). Every node in the tree corresponds to a null hypothesis to be tested. We assume that logical relationships exist between the hypotheses in the tree, in the sense that each parent hypothesis is the intersection of its child hypotheses: if $\text{children}(H) \neq \varnothing$, we have

$$H = \bigcap \text{children}(H).$$

Tree-structured hypotheses of this type may arise if a general research question is repeatedly split up into more specific sub-questions.

Meinshausen (2008) proposed a simple test procedure for tree structures and a more advanced one which exploits the logical relationships between the hypotheses in the manner of Shaffer (1986). We shall discuss both methods in turn and show how they can be improved using the sequential rejection principle.

The simple method would start testing the hypothesis at the top of the tree of Figure 1 at level α and, after rejection of that hypothesis, would continue testing both child nodes at level $\alpha/2$. If one of these child nodes gets rejected, its child nodes are then tested at level $\alpha/4$. The procedure continues until no further rejection is achieved. For general trees, this procedure is easily represented in the

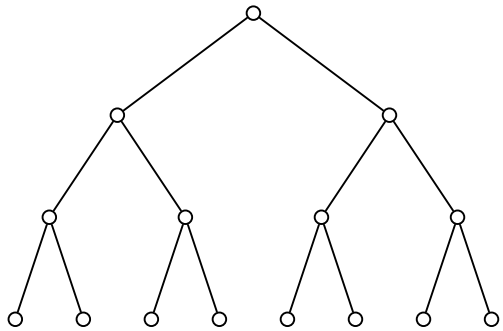

FIG. 1. *A symmetric binary tree of four levels.*

sequential rejection framework by the critical value function

$$\alpha_H(\mathcal{R}) = \begin{cases} \dfrac{\alpha \cdot L_H}{L}, & \text{if ancestors}(H) \subseteq \mathcal{R}, \\ 0, & \text{otherwise,} \end{cases}$$

where L_H is the number of descendant leaves of a hypothesis H and $L = |\mathcal{L}|$ is the total number of leaves $\mathcal{L}$ in the graph. Call a hypothesis H "active" if it has $\alpha_H(\mathcal{R}) > 0$ and is not rejected.

Control of the familywise error can easily be checked by the sequential rejection principle. Monotonicity of critical values is immediate from the definition. To check the single-step condition, note that we only need to consider control for those rejected sets $\mathcal{R}$ which are equal to $\mathcal{F}(M)$ for some $M \in \mathbb{M}$. Due to the logical relationships between the hypotheses, every $\mathcal{F}(M)$ is a subtree and the active hypotheses are the children of the leaves of this subtree. The set of active leaves of the original tree and the sets of descendant leaves below each active hypothesis are, therefore, all disjoint and the union of these sets contains exactly the $L'(\mathcal{R}) = |\mathcal{L} \setminus \mathcal{R}|$ unrejected leaves, so

$$\sum_{H \in \mathcal{H} \setminus \mathcal{R}} \alpha_H(\mathcal{R}) \leq \frac{\alpha \cdot L'(\mathcal{R})}{L} \leq \alpha.$$

This proves the single-step condition for Meinshausen's basic procedure.

From the inequality above, we can immediately see that we can set uniformly sharper critical values without loss of the single-step condition by setting

$$\alpha_H(\mathcal{R}) = \begin{cases} \dfrac{\alpha \cdot L_H}{L'(\mathcal{R})}, & \text{if ancestors}(H) \subseteq \mathcal{R}, \\ 0, & \text{otherwise,} \end{cases} \tag{24}$$

using the number $L'(\mathcal{R})$ of unrejected leaves, rather than the number of original leaves, in the denominator. This improvement is analogous to the improvement from the procedure of Bonferroni to the procedure of Holm.

The two procedures outlined above do not yet make effective use of the logical relationships between the null hypotheses in the graph. One way, proposed by Meinshausen (2008), to make use of those, is to use what he calls the Shaffer improvement. To keep notation simple for this improvement, consider only the case of a symmetric binary tree, which is a tree with a single root, in which every node has zero or two child nodes, and in which the subtrees formed by the descendants of child nodes of the same parent are identical (see Figure 1). For such a tree, Meinshausen proposed to use

$$\alpha_H(\mathcal{R}) = \begin{cases} \dfrac{\alpha \cdot L_H}{L}, & \text{if } H \notin \mathcal{L} \text{ and ancestors}(H) \subseteq \mathcal{R}, \\ \dfrac{2\alpha \cdot L_H}{L}, & \text{if } H \in \mathcal{L} \text{ and ancestors}(H) \subseteq \mathcal{R}, \\ 0, & \text{otherwise.} \end{cases}$$

The critical value function is identical to the first critical value function for all hypotheses that are not leaves, but multiplies all critical values of leaf node hypotheses by 2.

Control of the familywise error for this hypothesis follows from the sequential rejection principle in much the same way as above. To see why the factor 2 can be applied, note that when checking the single-step condition, we may assume that all rejections in $\mathcal{R}$ are correct rejections. In particular, once we have rejected a parent of a leaf node, because that hypothesis is the intersection of its two child hypotheses, we may assume that at least one of its children is false. Therefore, in the single-step condition, when calculating a bound for $\sum_{H \in \mathcal{T}(M)} \alpha_H(\mathcal{R})$, only one out of each pair of leaf nodes with common parent contributes to the sum.

It is convenient to rewrite the critical value function in terms of these pairs. Let P_H be the number of leaf node pairs that either include H or are descendants of H so that $P_H = L_H/2$ if H is not a leaf and $P_H = L_H = 1$ if H is a leaf. Let $P = L/2$ be the total number of leaf node pairs. We can then write

$$\alpha_H(\mathcal{R}) = \begin{cases} \dfrac{\alpha \cdot P_H}{P}, & \text{if ancestors}(H) \subseteq \mathcal{R}, \\ 0, & \text{otherwise.} \end{cases}$$

Consider the set of true null hypotheses that are active. Note that, by the same reasoning as above, each leaf node pair has at most one member or ancestor in that set and leaf node pairs which have been completely rejected by the procedure have no member or ancestor in that set. Therefore,

$$\sum_{H \in \mathcal{T}(M)} \alpha_H(\mathcal{R}) \le \frac{\alpha \cdot P'(\mathcal{R})}{P} \le \alpha,$$

where $P'(\mathcal{R})$ is the number of leaf node pairs that have not yet been completely rejected. This proves the single-step condition for Meinshausen's method with Shaffer's adjustment.

Again, we see that it is possible to set uniformly sharper critical values without loss of the single-step condition, setting

$$\alpha_H(\mathcal{R}) = \begin{cases} \dfrac{\alpha \cdot P_H}{P'(\mathcal{R})}, & \text{if ancestors}(H) \subseteq \mathcal{R}, \\ 0, & \text{otherwise}, \end{cases} \tag{25}$$

which provides a uniform improvement.

A second way to make use of logical relationships in Meinshausen's procedure is to remark that the procedure (24) is inadmissible according to the criterion (5) and may be improved on the basis of that criterion. This improvement is for general trees. We note that the single-step condition only needs to be shown for sets $\mathcal{R} \in \Phi$ and that $\mathcal{R} \in \Phi$ implies that for every $H \in \mathcal{R}$, there is always at least one leaf $K \in \text{offspring}(H)$ for which $K \in \mathcal{R}$. Therefore, define

$$\mathcal{D} = \{H \in \mathcal{R} : \text{offspring}(H) \cap \mathcal{R} = \varnothing\},$$

the leaf nodes of the rejected subgraph, and define $L''(\mathcal{R}) = L - |\mathcal{D}|$. Noting that $L''(\mathcal{R}) = L'(\mathcal{R})$ if $\mathcal{R} \in \Phi$ and that $L''(\mathcal{R}) \leq L'(\mathcal{S})$ for every $\mathcal{R} \subset \mathcal{S} \in \Phi$ if $\mathcal{R} \notin \Phi$, we see that (24) can be changed to

$$\alpha_H(\mathcal{R}) = \begin{cases} \dfrac{\alpha \cdot L_H}{L''(\mathcal{R})}, & \text{if ancestors}(H) \subseteq \mathcal{R}, \\ 0, & \text{otherwise}, \end{cases} \tag{26}$$

without losing familywise error control. This is a uniform improvement over (24) because $L''(\mathcal{R}) > L'(\mathcal{R})$ for all $\mathcal{R} \notin \Phi$. It is easy to check that (26) conforms to the condition (5).

It is interesting to note that the two improvements (25) and (26) of Meinshausen's method do not dominate each other. This suggests that many extensions, variants and alternative improvements are possible, but these are beyond the scope of this paper.

We also remark that the variant of Meinshausen's procedure without Shaffer's improvement is a special case of the methods of Burman, Sonesson and Guilbaud (2009) and Bretz et al. (2009). The improvement (24) might have been obtained in an easy way using the approaches of these authors and is also valid in the absence of logical relationships between hypotheses. The methods (25) and (26) that exploit logical relationships, however, are not contained in the frameworks of Bretz et al. (2009) and Burman, Sonesson and Guilbaud (2009), and require the use of the sequential rejection principle.

7. Multiplicity-adjusted *p*-values. Often in multiple testing situations, interest is not just in rejection and nonrejection of hypotheses at a pre-specified level α, but also in reporting multiplicity-adjusted p-values. Such multiplicity-adjusted p-values are defined for each null hypothesis as the smallest α-level that allows rejection of that hypothesis. In the general sequentially rejective procedure, they

can easily be found using the following algorithm, described earlier by Goeman and Mansmann (2008) for the specific case of the focus level procedure.

Suppose the critical value function $\mathbf{c}$ depends on a parameter α in such a way that (1) the sequentially rejective procedure based on $\mathbf{c}_\alpha$ controls the familywise error at most at α, and that (2) for all H and $\mathcal{R}$, $c_{H,\alpha_1}(\mathcal{R}) \geq c_{H,\alpha_2}(\mathcal{R})$ if $\alpha_1 \leq \alpha_2$, that is, critical values are nonincreasing in α. Multiplicity-adjusted p-values can then be calculated in the following way.

Initialize $\alpha_0 = 0$ and $\mathcal{R}^0_\infty = \varnothing$. Iterate for $i = 1, 2, \ldots$

1. Set α_i to the smallest α for which $S_H \geq c_{H,\alpha}(\mathcal{R}^{i-1}_\infty)$ for any $H \in \mathcal{H} \setminus \mathcal{R}^{i-1}_\infty$.
2. Follow the sequentially rejective procedure with the critical value function $\mathbf{c}_{\alpha_i}$, starting from $\mathcal{R}^i_0 = \mathcal{R}^{i-1}_\infty$, to find $\mathcal{R}^i_\infty$.
3. Set the multiplicity-adjusted p-values of all $H \in \mathcal{R}^i_\infty \setminus \mathcal{R}^{i-1}_\infty$ to α_i.

The procedure can be stopped when either $\mathcal{R}^i_\infty = \mathcal{H}$ or when $\alpha_i \geq 1$. If the latter happens, all $H \in \mathcal{H} \setminus \mathcal{R}^{i-1}_\infty$ can be given multiplicity-adjusted p-value 1.

In step 2 of this algorithm, the sequentially rejective procedure for the next higher value of α starts from the final rejected set of the previous value of α. This is what makes the algorithm relatively efficient. It is interesting to note that this "warm start" is allowed as another consequence of the monotonicity condition (7): if that condition holds, then the sequentially rejective procedure that starts at $\mathcal{R}^i_0 = \mathcal{R}^{i-1}_\infty$ converges to the same final rejected set as the sequentially rejective procedure that starts at $\mathcal{R}^i_0 = \varnothing$.

8. Discussion. The sequential rejection principle is a fundamental property of familywise error control which has been implicity exploited in many important methods. The sequential rejection principle links Holm's procedure to Bonferroni's. It presents both the closed testing procedure and the partitioning principle as consequences of Shaffer's procedure. It ties the tests in different families of a gatekeeping procedure together and it connects the step-down version of resampling-based multiple testing to the single-step version. The procedure is not limited in its application to step-down methods, but can also effectively be used in the context of step-up methods, as we have demonstrated for Hochberg's method in the case of logically related null hypotheses.

This paper has made the sequential rejection principle explicit. It shows how many well-known methods can be constructed as special cases of a general sequentially rejective multiple testing procedure, which is a monotone sequence of single-step procedures with a limited form of weak familywise error control. This general procedure is interesting from a theoretical point of view, showing a close relatedness between seemingly different multiple testing procedures. The general procedure encompasses a great number of well-known sequentially rejective familywise error-controlling procedures and even some that have never been viewed as sequentially rejective before.

The relationship between the sequential rejection principle and the partitioning principle deserves some attention. Even though we have shown that the partitioning principle may be derived as a special case of the sequential rejection principle, we do not claim that sequential rejection is a more powerful or more fundamental principle than partitioning. Rather, the sequential rejection principle presents an alternative perspective on multiple testing, which is flexible enough to include both closed testing and partitioning as special cases, but which does not always require construction of the full partitioning or closure of the hypotheses of interest.

The most important aspect of the sequential rejection principle, however, is its practical usefulness. This ranges from simple applications, such as quickly answering the question whether any multiple testing correction is needed for simultaneous post hoc testing of $H : \mu_1 = 0$ and $J : \mu_2 = 0$ after $K : \mu_1 = \mu_2$ has been rejected, to the construction of multiple testing procedures for complicated graphs. Recently, there has been considerable interest in the latter application, both in the field of clinical trials with multiple endpoints and in the field of genomics. The sequential rejection principle can be a valuable tool in this area since the general sequentially rejective procedure lends itself very easily to graph-based testing, with conditions for strong control of the familywise error for the procedure that are intuitive and easy to check. The sequential rejection principle improves upon earlier proposals for general graph-based multiple testing procedures because it is capable of incorporating logical relationships between null hypotheses and because it is not restricted to Bonferroni-based control at each single step.

APPENDIX A: RELAXING THE MONOTONICITY CONDITION? A COUNTEREXAMPLE

In this section, we show that the relaxed version (8) of the monotonicity condition (7) is not sufficient for familywise error control. We do this by first constructing a sequentially rejective procedure that conforms to (9) and which controls the familywise error in each single step at level α, but which does not conform to (7). Next, we construct a data generating distribution for which this procedure has a familywise error greater than α. The example is highly artificial, but it serves as an interesting counterexample to the possibility of relaxation of the monotonicity criterion.

The sequential procedure is of gatekeeping type, with four hypotheses: J, J', K and K'. The hypotheses J and K are primary, and the hypotheses J' and K' are secondary, being tested only after at least one of J and K has been rejected. Suppose that we have test statistics U_J, $U_{J'}$, U_K and $U_{K'}$, corresponding to the four hypotheses. Suppose, also, that the general model $\mathbb{M}$ says that, for $H \in \{J, K, J', K'\}$, each U_H is marginally uniform $\mathcal{U}(0, 1)$ if H is true, and $\mathcal{U}(0, b_H)$ with $b_H < 1$ if H is false. The test statistics are therefore very much like p-values and, as a consequence, we would reject each H for small values of U_H, as in the notation of Section 3. To construct the sequentially rejective procedure,

TABLE 1
Critical value function $\boldsymbol{\alpha}(\cdot)$ of the sequentially rejective procedure

	Previously rejected hypotheses			
$\boldsymbol{\alpha}$	$\boldsymbol{\varnothing}$	$\boldsymbol{\{J\}}$	$\boldsymbol{\{K\}}$	$\boldsymbol{\{J, K\}}$
α_J	ε	—	ε	—
α_K	ε	ε	—	—
$\alpha_{J'}$	—	$\alpha - \varepsilon$	—	$\alpha/2$
$\alpha_{K'}$	—	—	$\alpha - \varepsilon$	$\alpha/2$

choose some $0 < \alpha \leq 1/2$ and some $0 < \varepsilon < \alpha/2$. The critical value function $\boldsymbol{\alpha}(\cdot)$ of the procedure is summarized in Table 1 for the rejection sets relevant to the first two steps of the procedure.

The single-step condition of this procedure is easily checked, as the column sums of the table are bounded by α. It is also immediately clear that the procedure based on the critical value function of Table 1 does not satisfy the monotonicity condition (7) since

$$\alpha_{J'}(\{J, K\}) = \tfrac{1}{2}\alpha < \alpha - \varepsilon = \alpha_{J'}(\{J\}). \tag{27}$$

However, the procedure does satisfy the relaxed monotonicity condition (8) since $\mathcal{R}_1 = \{J\}$ can never be followed by $\mathcal{R}_2 = \{J, K\}$, so (27) is not relevant for that condition.

We now give an example of a distribution for which the procedure based on the critical value function of Table 1 does not control the familywise error. Suppose that, under the true model, U_J, $U_{J'}$, U_K and $U_{K'}$ all depend on a single uniform $\mathcal{U}(0, 1)$ variable U, in such a way that

$$\begin{aligned} U_J &= tU, \\ U_K &= t(1 - U), \\ U_{J'} &= U, \\ U_{K'} &= 1 - U \end{aligned}$$

for some $2\varepsilon \leq t \leq \varepsilon/\alpha$. Note that J' and K' are true null hypotheses, whereas J and K are false. For this distribution, $\{U \leq \alpha - \varepsilon\}$ implies rejection of J, but not K, in step 1, followed by rejection of J' in step 2, while, at the same time, $\{U \geq 1 - \alpha + \varepsilon\}$ implies rejection of K, but not J, in step 1, followed by rejection of K' in step 2. The total probability of making a false rejection is therefore

$$\text{FWER} \geq \mathrm{P}(\{U \leq \alpha - \varepsilon\} \cup \{U \geq 1 - \alpha + \varepsilon\}) = 2\alpha - 2\varepsilon > \alpha$$

and we conclude that the procedure does not control the familywise error.

The procedure based on Table 1 can go wrong because the critical value function allows the first step of the sequentially rejective procedure to preselect the null hypothesis that is most likely to give a false rejection in the second step. The monotonicity requirement (7) prevents this, but the relaxed monotonicity requirement (8) does not.

APPENDIX B: PROOF OF THEOREM 2

Choose any $M \in \mathbb{M}$ and let $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}(M)$. Let the random variables $\delta_1, \ldots, \delta_r$ be defined as

$$\delta_i = \begin{cases} 1, & \text{if } \max_{H \in \mathcal{T}(M)} S_H \circ \pi_i > \left(\max_{J \in \mathcal{T}(M)} S_J \circ \boldsymbol{\pi} \circ \pi_i \right)_{(s)}, \\ 0, & \text{otherwise.} \end{cases}$$

By condition (17), for all i,

$$\mathrm{E}_M(\delta_i) = \mathrm{P}_M\left(\bigcup_{H \in \mathcal{T}(M)} \{S_H > k_H(\mathcal{F})\} \right), \tag{28}$$

where E_M denotes expectation with respect to the measure P_M.

Because $\{\pi_1, \ldots, \pi_r\}$ form a group in the algebraic sense, it follows that for every i,

$$\{\pi_i \circ \pi_1, \ldots, \pi_i \circ \pi_r\} = \{\pi_1, \ldots, \pi_r\}.$$

Therefore, for every i,

$$\left(\max_{J \in \mathcal{T}(M)} S_J \circ \boldsymbol{\pi} \right)_{(s)} = \left(\max_{J \in \mathcal{T}(M)} S_J \circ \pi_i \circ \boldsymbol{\pi} \right)_{(s)}.$$

Consequently,

$$\sum_{i=1}^{r} \delta_i = \#\left\{ i : \max_{H \in \mathcal{T}(M)} S_H \circ \pi_i > \left(\max_{J \in \mathcal{T}(M)} S_J \circ \boldsymbol{\pi} \right)_{(s)} \right\} \le r - s \le r\alpha$$

for all $\omega \in \Omega$. Combining this with (28), we have

$$\mathrm{P}_M\left(\bigcup_{H \in \mathcal{T}(M)} \{S_H > k_H(\mathcal{F})\} \right) = r^{-1} \sum_{i=1}^{r} \mathrm{E}_M(\delta_i) = \mathrm{E}_M\left(r^{-1} \sum_{i=1}^{r} \delta_i \right) \le \alpha.$$

Acknowledgments. We are grateful to the Associate Editor and the anonymous referees for several suggestions which have helped to shorten the proofs and to make the results of this paper much more elegant.

REFERENCES

BRETZ, F., MAURER, W., BRANNATH, W. and POSCH, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Stat. Med.* **28** 586–604.

BURMAN, C.-F., SONESSON, C. and GUILBAUD, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Stat. Med.* **28** 739–761.

CALIAN, V., LI, D. M. and HSU, J. C. (2008). Partitioning to uncover conditions for permutation tests to control multiple testing error rates. *Biometrical J.* **50** 756–766.

DMITRIENKO, A., OFFEN, W. W. and WESTFALL, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Stat. Med.* **22** 2387–2400.

DMITRIENKO, A. and TAMHANE, A. C. (2007). Gatekekeeping procedures with clinical trial applications. *Pharmaceutical Statistics* **6** 171–180.

DMITRIENKO, A., TAMHANE, A. C., WANG, X. and CHEN, X. (2006). Stepwise gatekeeping procedures in clinical trial applications. *Biometrical J.* **48** 984–991. MR2307673

DMITRIENKO, A., TAMHANE, A. C. and WIENS, B. L. (2008). General multistage gatekeeping procedures. *Biometrical J.* **50** 667–677.

DMITRIENKO, A., WIENS, B. L., TAMHANE, A. C. and WANG, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat. Med.* **26** 2465–2478. MR2364399

DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer, New York. MR2373771

DUDOIT, S., VAN DER LAAN, M. J. and POLLARD, K. S. (2004). Multiple testing part I: Single-step procedures for control of general type I error rates. *Stat. Appl. Genet. Mol. Biol.* **3** Article 13. MR2101462

EDWARDS, D. and MADSEN, J. (2007). Constructing multiple test procedures for partially ordered hypotheses sets. *Stat. Med.* **26** 5116–5124. MR2412460

FINNER, H. and STRASSBURGER, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *Ann. Statist.* **30** 1194–1213. MR1926174

GE, Y., DUDOIT, S. and SPEED, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test* **12** 1–77. MR1993286

GOEMAN, J. J. and MANSMANN, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24** 537–544.

GUILBAUD, O. (2007). Bonferroni parallel gatekeeping—Transparant generalizations, adjusted p-values, and short direct proofs. *Biometrical J.* **49** 917–927. MR2416452

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802. MR0995126

HOCHBERG, Y. and ROM, D. (1995). Extensions of multiple testing procedures based on Simes' test. *J. Statist. Plann. Inference* **48** 141–152. MR1366786

HOLLAND, B. S. and DIPINZIO COPENHAVER, M. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* **43** 417–423. MR0897410

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70. MR0538597

HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75** 383–386.

HOMMEL, G. and BERNHARD, G. (1999). Bonferroni procedures for logically related hypotheses. *J. Statist. Plann. Inference* **82** 119–128. MR1736436

HOMMEL, G., BRETZ, F. and MAURER, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Stat. Med.* **26** 4063–4073. MR2405792

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York. MR2135927

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056

MEINSHAUSEN, N. (2008). Hierarchical testing of variable importance. *Biometrika* **95** 265–278. MR2521583

ROMANO, J. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypotheses testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821

ROMANO, J. and WOLF, M. (2010). Balanced control of generalized error rates. *Ann. Statist.* **38** 598–633. MR2590052

ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **25** 248–252. MR2409727

SARKAR, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: A proof of the Simes conjecture. *Ann. Statist.* **26** 494–504. MR1626047

SHAFFER, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81** 826–831.

ŠIDÁK, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62** 626–633. MR0216666

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872

STEFANSSON, G., KIM, W. and HSU, J. C. (1988). On confidence sets in multiple comparisons. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **2** 89–104. Springer, New York. MR0927125

WESTFALL, P. H. and KRISHEN, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J. Statist. Plann. Inference* **99** 25–40. MR1858708

WESTFALL, P. H. and TROENDLE, J. F. (2008). Multiple testing with minimal assumptions. *Biometrical J.* **50** 745–755.

WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

WIENS, B. L. and DMITRIENKO, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *J. Biopharm. Statist.* **15** 929–942. MR2210804

DEPARTMENT OF MEDICAL STATISTICS
AND BIOINFORMATICS
LEIDEN UNIVERSITY MEDICAL CENTER
P.O. BOX 9600
2300 RC LEIDEN
THE NETHERLANDS
E-MAIL: j.j.goeman@lumc.nl

DEPARTMENT OF CHEMICAL
PROCESS ENGINEERING
UNIVERSITY OF PADOVA
VIA F. MARZOLO 9
35131 PADOVA
ITALY
E-MAIL: solari@stat.unipd.it