Original Article

An Extension of the Holm Procedure Based on Partitioning Principle for Adaptive Treatment Selection Designs with Structured Hypotheses

Toshifumi Sugitani
Department of Biomedical Statistics,
Osaka University Graduate School of Medicine
e-mail:sugitani@medstat.med.osaka-u.ac.jp

Multiple comparisons are attracting increasing attention in the evaluation of statistical evidence in clinical trials including at least one or any combination of (i) multiple hypotheses, (ii) repeated hypotheses testing at interim analyses, and (iii) mid-course design adaptations. In this paper, we discuss an efficient and sensible multiple testing procedure for two-stage adaptive treatment selection designs including structured hypotheses. Specifically, we extend the Holm procedure for serial gatekeeping structured hypotheses in adaptive clinical trials. The proposed approach is based on the idea of combining partition testing with the inverse normal combination test. A clinical trial example is used to illustrate the implementation of the proposed procedure.

Key words: Gatekeeping, Inverse normal combination test, Multiplicity, Partitioning principle, Two-stage adaptive design.

1. Introduction

In the last decade, clinical trial designs have become increasingly complex, to improve quality, efficiency, and effectiveness of drug development processes. Such complex designs may include at least one or any combination of the following elements: multiple hypotheses (e.g., hierarchically ordered hypotheses or co-primary hypotheses), repeated hypothesis testing at interim analyses (e.g., early stopping of a trial), and mid-course design adaptations (e.g., treatment selection or sample size reestimation).

Typical examples include the Evaluation of the Safety and Cardioprotective Effects of Eniporide in Acute Myocardial Infarction (ESCAMI) study (Zeymer et al., 2001), which was an international, prospective, randomized, double-blind, placebo-controlled phase II dose finding study. The primary endpoint of the study was infarct size determined by the cumulative release of α-hydroxybutyrate dehydrogenase (α-HBDH), and the key secondary endpoints were creatine kinase (CK) and CK-isoenzyme (CK-MB), troponin T, and troponin I. Here, the key secondary
endpoints were considered equally important from clinical perspectives, and were of interest only if the primary endpoint shows statistically significant efficacy improvement over placebo. For the design and analysis, the ESCAMI study employed the two-stage adaptive design proposed by Bauer and Köhne (1994), where pre-planned adaptations at interim were to select a subset of doses to be carried forward into the second stage and to determine the number of patients to be recruited for the second stage. However, the Bauer and Köhne’s method cannot be used for the test of the primary and four secondary endpoints while accounting for their clinical hierarchy as well as the strong control of the familywise error rate (FWER). Here, the strong control of the FWER indicates that the probability of erroneously rejecting at least one true null hypothesis is controlled under any configuration of true and false null hypotheses.

Complex clinical trial designs present statisticians a challenging task of developing new multiple test procedures to handle newly arising multiplicity problems efficiently. Recently, many novel multiple test procedures were introduced in the literature; for example, see Tamhane et al. (2012), Maurer and Bretz (2013a,b), and Sugitani et al. (2013). Among such proposals, extensions of the Holm procedure (Holm, 1979) are well considered (Dmitrienko et al., 2008; Li, 2011; Ye et al., 2013) and are often used in practice, as the Holm procedure is very simple to understand and easy to implement even for non-statisticians. However, such extensions of the Holm procedure cannot handle structured hypotheses in the context of adaptive designs.

For a confirmatory adaptive trial to be statistically valid, a minimum requirement is the control of the FWER as well as prospective specification of adaptation criteria. In this spirit, many novel multiple test procedures have been introduced so far over the last two decades. In the situation where multiple hypotheses are of interest, broadly, two classes of closed testing procedures (Marcus et al., 1976) have been introduced so far in the literature, that is, (i) adaptive combination test by Bauer and Kieser (1999), Hommel (2001), and Bretz et al. (2006), and (ii) partial conditional error rate approach by Posch et al. (2011). As these procedures can flexibly handle the multiple hypotheses in adaptive clinical trials, they are frequently used in practice and discussed in the literature. However, in the situation where the multiple hypotheses are hierarchically structured a priori based on relative clinical importance, they cannot be easily handled using the closed testing procedures like adaptive combination test or partial conditional error rate approach. More specifically, with adaptive combination test or partial conditional error rate approach, one cannot easily have the same weighing scheme as the multiple test procedures such as graphical approach (Bretz et al., 2009; Maurer and Bretz, 2013a,b) or gatekeeping procedure (Dmitrienko and Tamhane, 2011; 2013).

Instead, in this paper, we consider using partition testing (Hsu and Berger, 1999; Finner and Strassburger, 2002) to introduce an adaptive multiple test procedure that has a similar weighting scheme to the serial gatekeeping procedure (Westfall and Krishen, 2001). More specifically, motivated by the ESCAMI study, we propose via partition testing an extension of the Holm
procedure for adaptive treatment selection designs that involve serial gatekeeping structured hypotheses, i.e., the hypotheses that are tested if all hypotheses in more important families are rejected. Our proposal can be seen as an extension of the partition testing by Huang and Hsu (2007), using the inverse normal combination test (Lehmacher and Wassmer, 1999).

The rest of the paper is organized as follows. In Section 2, we briefly outline the partition testing by Huang and Hsu (2007). In Section 3, we discuss its extension to adaptive treatment selection designs with serial gatekeeping structured hypotheses. In Section 4, we illustrate the proposed procedure with a real-data example from the ESCAMI study. Finally, we summarize the findings as concluding remarks in Section 5.

2. Partitioning principle and Holm procedure

Partitioning principle is a general principle for constructing multiple test procedures that control the FWER in the strong sense, by partitioning hypotheses into mutually disjoint hypotheses and then testing the respective disjoint hypotheses at the entire level $\alpha$ (Hsu and Berger 1999; Finner and Strassburger, 2002). On the basis of the partitioning principle, Huang and Hsu (2007) provided an alternative derivation of the Holm procedure, as described below.

Assume that we are interested in testing a family of hypotheses $H_i, i \in I = \{1, \ldots, v\}$, and $p$-value for $H_i$ is denoted by $p_i$. To derive the Holm procedure, Huang and Hsu (2007) used natural partition (Finner and Strassburger, 2002) described as

$$\Theta(J) = \begin{cases} 
(\cap_{i \in J} H_i) \cap (\cap_{i \in I \setminus J} K_i) & \text{for } J \subset I, J \neq \emptyset \\
\cap_{i \in I} H_i & \text{for } I = J
\end{cases}$$

where $K_i$ denotes an alternative hypothesis for $H_i$. The natural partitioned hypotheses expressed in equation (1) are mutually disjoint and hence can be tested at the entire level $\alpha$ without inflating the FWER, according to partitioning principle. Huang and Hsu (2007) showed that if rejection rules for $\Theta(J)$’s are determined such that

$$p_i < \alpha/|J|, \text{ for all } i \in J \subseteq I,$$

the resulting test procedure reduces to the following Holm procedure:

**Step 1:** If $p_{(1)} < \alpha/v$, reject $H_{(1)}$, and go to the next step; otherwise, accept the hypotheses $H_{(1)}, \ldots, H_{(v)}$.

**Step 2:** If $p_{(2)} < \alpha/(v - 1)$, reject $H_{(2)}$, and go to the next step; otherwise, accept the hypotheses $H_{(2)}, \ldots, H_{(v)}$.

Likewise, test other hypotheses.

**Step v:** If $p_{(v)} < \alpha$, reject $H_{(v)}$ and stop; otherwise, accept $H_{(v)}$.

where $| \cdot |$ denotes the number of elements, and $p_{(i)}$ an ordered $p$-value such that $p_{(1)} < \cdots < p_{(v)}$.
with the corresponding hypothesis $H_{(i)}$.

3. Proposed procedure

Assume that we compare $k$ treatments with control in a two-stage adaptive design in which some of the promising treatments may be selected at an interim without formal hypothesis testing. Further, assume that each treatment–control comparison involves $v$ endpoints divided into $m$ ($m \leq v$) hierarchical families according to their relative clinical importance. In this case, we are interested in testing hypotheses $H_i$, $i \in I = I_1 \cup \cdots \cup I_k$, where $I_s = \{(s-1)v + 1, \ldots, sv\}$, indicating a comparison of treatment $s$, $s \in A = \{1, \ldots, k\}$, and control based on $v$ endpoints. The index set $I_s$ is further divided as $I_s = I_{s,1} \cup \cdots \cup I_{s,m}$, where $I_{s,t} = \{(s-1)v + \sum_{x=1}^{t-1} v_x + 1, \ldots, (s-1)v + \sum_{x=1}^{t-1} v_x + vt\}$, $t = 1, \ldots, m$, indicating that the $v$ endpoints are divided into $m$ hierarchical families of hypotheses, such that $\sum_{t=1}^{m} v_t = v$.

For simplicity, consider a situation where $k = 2$ treatments are compared to control using two primary and two secondary endpoints (i.e., $v = 4$, $m = 2$, $v_1 = 2$, and $v_2 = 2$). In this case, we are interested in testing hypotheses $H_i$, $i \in I = I_1 \cup I_2$ with $I_1 = I_{1,1} \cup I_{1,2}$ and $I_2 = I_{2,1} \cup I_{2,2}$, where $I_{1,1} = \{1, 2\}$, $I_{1,2} = \{3, 4\}$, $I_{2,1} = \{5, 6\}$ and $I_{2,2} = \{7, 8\}$. Consider partitioning $\Theta_1 = \cup_{i \in I_1} H_i$ into mutually disjoint hypotheses. In the first step, $\Theta_1$ is partitioned into $\Theta_{s,t}(J_{s,t})$'s, $J_{s,t} \subseteq I_{s,t}$, as follows:

$$
\Theta_{1,1}(\{1, 2\}) = H_1 \cap H_2,
\Theta_{1,1}(\{1\}) = H_1 \cap K_2,
\Theta_{1,1}(\{2\}) = K_1 \cap H_2,
\Theta_{1,2}(\{3, 4\}) = H_3 \cap H_4,
\Theta_{1,2}(\{3\}) = H_3 \cap K_4,
\Theta_{1,2}(\{4\}) = K_3 \cap H_4.
$$

In the second step, $\Theta_{s,t}(J_{s,t})$'s are further partitioned into $\Theta^*_{s,t}(J_{s,t})$'s, $J_{s,t} \subseteq I_{s,t}$, as follows:

$$
\Theta^*_{1,1}(\{1, 2\}) = H_1 \cap H_2,
\Theta^*_{1,1}(\{1\}) = H_1 \cap K_2,
\Theta^*_{1,1}(\{2\}) = K_1 \cap H_2,
\Theta^*_{1,2}(\{3, 4\}) = (K_1 \cap K_2) \cap (H_3 \cap H_4),
\Theta^*_{1,2}(\{3\}) = (K_1 \cap K_2) \cap (H_3 \cap K_4),
\Theta^*_{1,2}(\{4\}) = (K_1 \cap K_2) \cap (K_3 \cap H_4).
$$

In the same manner, $\Theta_2$ is also partitioned into mutually disjoint hypotheses $\Theta^*_{s,t}$'s. Note that $\Theta^*_{1,t}$'s and $\Theta^*_{2,t}$'s are not mutually disjoint. Figure 1 illustrates these partitioned hypotheses graphically to clarify the underlying partitioning strategy. Generalization of this partition algo-
Fig. 1. Graphical illustration of partition algorithm. To control the FWER at level $\alpha$, the (weighted) Bonferroni adjustment is used, that is, local significance levels $\alpha_1$ and $\alpha_2$ are allocated for treatments 1 and 2, respectively, such that $\alpha_1 + \alpha_2 = \alpha$.

rithm is provided in Appendix.

For testing $\Theta^*_{s,t}(J_{s,t})$’s, we consider using the inverse normal combination test (Lehmacher and Wassmer, 1999) described as

$$C(p_{i,1}, p_{i,2}) = 1 - \Phi\left(\frac{\sqrt{n_1/(n_1 + n_2)}\Phi^{-1}(1 - p_{i,1}) + \sqrt{n_2/(n_1 + n_2)}\Phi^{-1}(1 - p_{i,2})}{\Phi^{-1}(1 - p_{i,1}) + \Phi^{-1}(1 - p_{i,2})}\right),$$

(3)

where $\Phi^{-1}$ denotes the inverse of cumulative probability of the standard normal distribution; $p_{i,j}$, the one-sided $p$-value for $H_i$, $i \in I$ in stage $j = 1, 2$; and $n_j$, the pre-planned sample size per group in stage $j$. Here, a requirement for the control of the FWER is that the stagewise $p$-values meet the regularity conditions $\Pr_{H_i}(p_{i,1} \leq \alpha) \leq \alpha$ and $\Pr_{H_i}(p_{i,2} \leq \alpha | p_{i,1}) \leq \alpha$ for all $0 \leq \alpha \leq 1$ (Brannath et al., 2002). This means that the distribution of $p_{i,1}$ and the conditional distribution of $p_{i,2}$ given $p_{i,1}$ are stochastically larger than or equal to the uniform distribution on $[0, 1]$ under null hypothesis $H_i$. Brannath et al. (2009) further showed that these regularity conditions still hold asymptotically for time-to-event data, where independent increment of Fisher’s information does not generally hold.

The reason why we use the inverse normal combination method rather than other $p$-value combination methods (e.g., Fisher’s product method; Bauer and Köhne, 1994) is that the inverse normal method enables one to have a very simple rejection boundary, i.e., for controlling the FWER at level $\alpha \in (0, 1)$, we can have a simple rejection boundary of $C(p_{i,1}, p_{i,2}) < \alpha$. This simple rejection boundary further enables one to derive a new multiple test procedure that has the same rejection rule as the serial gatekeeping Holm procedure, as described below.

As mentioned previously, the partitioned hypotheses $\Theta^*_{s,t}(J_{s,t})$ are only locally disjoint, so that multiplicity adjustment for multiple treatment–control comparisons is needed to keep the FWER at level $\alpha$. In this paper, we use the Bonferroni method to adjust for such multiplicity. Namely, we allocate a local significance level $\alpha_s$, $s \in \{1, 2\}$, to each treatment–control comparison, such that $\alpha_1 + \alpha_2 = \alpha$. This Bonferroni adjustment strategy is sometimes termed “single stage
Bonferroni method” (for example, see Bretz et al., 2006) and controls the FWER in the strong sense. This adjustment strategy leads to the following rejection rules for $\Theta^*_s,t(J_s,t)$’s:

$$C(p_{i,1},p_{i,2}) < \alpha_s/|J_s,t|, \text{ for all } i \in J_s,t \subseteq I_{s,t}. \quad (4)$$

which are summarized in Table 1. Comparing the rejection rule (4) with (2), one can see how the latter is modified to derive the following extension of the Holm procedure.

Assume further that treatment 1 is discontinued at interim. If we order the observed combined \(p\)-values for treatment 2 as \(C(p_{5,1},p_{5,2}) < C(p_{6,1},p_{6,2})\) in \(I_{2,1}\), and \(C(p_{7,1},p_{7,2}) < C(p_{8,1},p_{8,2})\) in \(I_{2,2}\), then we have the following sequential rejective procedure:

**Step 1:** If \(C(p_{5,1},p_{5,2}) < \alpha_2/2\), reject \(H_5\), and go to the next step; otherwise, accept the hypotheses \(H_5, \ldots, H_8\).

**Step 2:** If \(C(p_{6,1},p_{6,2}) < \alpha_2\), reject \(H_6\), and go to the next step; otherwise, accept the hypotheses \(H_6, H_7, \text{ and } H_8\).

**Step 3:** If \(C(p_{7,1},p_{7,2}) < \alpha_2/2\), reject \(H_7\), and go to the next step; otherwise, accept the hypotheses \(H_7\) and \(H_8\).

**Step 4:** If \(C(p_{8,1},p_{8,2}) < \alpha_2\), reject \(H_8\) and stop; otherwise accept \(H_8\).

It can be seen that the proposed procedure has the same rejection rule as the serial gatekeeping Holm procedure within each treatment–control comparison, with the only distinction that the inverse normal combination test is used as test statistics. Hence, if only treatment 1 is selected, our proposed procedure implements serial gatekeeping Holm procedure at level \(\alpha_1\). Similarly, if both treatments are selected, our proposed procedure implements the serial gatekeeping Holm procedures for treatment 1-vs-control and treatment 2-vs-control comparisons at levels \(\alpha_1\) and \(\alpha_2\), respectively. Note that the proposed procedure allows sample size reassessment without compromising the FWER, as it uses the inverse normal combination test as test statistics. Formal proof is provided in Appendix to show that our proposed procedure strongly controls the FWER.

Finally, we note that the single stage Bonferroni method is different from the post-hoc strategy where the entire significance level \(\alpha\) is divided by the number of selected treatments after the interim analysis and then the split local significance levels are allocated to the selected treatments. The former strongly controls the FWER regardless of how treatment(s) is selected at interim (i.e., formal or informal), because it allocates local significance levels \textit{a priori} to individual hypotheses. On the other hand, the latter is affected by treatment selection and can inflate the FWER due to the selection bias. Wang et al. (2010) called this FWER “learning free type I error rate” and demonstrated via simulation studies that the learning free type I error rate indeed exceeded the nominal significance level.
| Partition hypothesis | Rejection rule |
|----------------------|----------------|
| $\Theta_{1,1}^*(\{1,2\})$: $H_1 \cap H_2$ | $C(p_{1,1},p_{1,2}) < \alpha_1/2$ or $C(p_{2,1},p_{2,2}) < \alpha_1/2$ |
| $\Theta_{1,1}^*(\{1\})$: $H_1 \cap K_2$ | $C(p_{1,1},p_{1,2}) < \alpha_1$ |
| $\Theta_{1,1}^*(\{2\})$: $K_1 \cap H_2$ | $C(p_{2,1},p_{2,2}) < \alpha_1$ |
| $\Theta_{2,1}^*(\{3,4\})$: $(K_1 \cap K_2) \cap (H_3 \cap H_4)$ | $C(p_{3,1},p_{3,2}) < \alpha_1/2$ or $C(p_{4,1},p_{4,2}) < \alpha_1/2$ |
| $\Theta_{2,1}^*(\{3\})$: $(K_1 \cap K_2) \cap (H_3 \cap K_4)$ | $C(p_{3,1},p_{3,2}) < \alpha_1$ |
| $\Theta_{2,1}^*(\{4\})$: $(K_1 \cap K_2) \cap (K_3 \cap H_4)$ | $C(p_{4,1},p_{4,2}) < \alpha_1$ |
| $\Theta_{2,1}^*(\{5,6\})$: $H_5 \cap H_6$ | $C(p_{5,1},p_{5,2}) < \alpha_2/2$ or $C(p_{6,1},p_{6,2}) < \alpha_2/2$ |
| $\Theta_{2,1}^*(\{5\})$: $H_5 \cap K_6$ | $C(p_{5,1},p_{5,2}) < \alpha_2$ |
| $\Theta_{2,1}^*(\{6\})$: $K_5 \cap H_6$ | $C(p_{6,1},p_{6,2}) < \alpha_2$ |
| $\Theta_{2,2}^*(\{7,8\})$: $(K_5 \cap K_6) \cap (H_7 \cap H_8)$ | $C(p_{7,1},p_{7,2}) < \alpha_2/2$ or $C(p_{8,1},p_{8,2}) < \alpha_2/2$ |
| $\Theta_{2,2}^*(\{7\})$: $(K_5 \cap K_6) \cap (H_7 \cap K_8)$ | $C(p_{7,1},p_{7,2}) < \alpha_2$ |
| $\Theta_{2,2}^*(\{8\})$: $(K_5 \cap K_6) \cap (K_7 \cap H_8)$ | $C(p_{8,1},p_{8,2}) < \alpha_2$ |
4. Illustration with the ESCAMI study

To illustrate the proposed procedure, we use the ESCAMI study mentioned in Section 1. The study employed a two-stage adaptive design, where pre-planned adaptations at interim were to select a subset of doses to be carried forward into the second stage and to determine the number of patients to be recruited for the second stage. In the study, a sample size of 100 patients per group was fixed for the first stage, and 316 patients per group was planned for the second stage. At the interim analysis, based on the first stage data of 433 patients, two dose levels, 100 mg and 150 mg eniporide, were selected for the second stage. In addition, the sample size was reestimated on the basis of the conditional power argument, resulting in a total of 978 patients for the second stage. At the final analysis, the stagewise p-values were combined using the Bauer and Köhne method (1994). However, the combined results did not show any beneficial effects of both selected dose levels on the primary or secondary endpoints. For more details, see Zeymer et al. (2001).

A major concern about the design and analysis of the ESCAMI study is that the key secondary endpoints, which were coupled with the primary endpoint at the interim dose-selection as well as the final analysis, were not considered in the strong control of the FWER. Therefore, the key secondary endpoints could not serve as evidence to help interpret the result of the primary endpoint, even if the secondary endpoints showed any significant statistical difference between the eniporide doses and placebo in addition to those shown by the primary endpoint.

For illustrative purpose, we will re-analyze the data focusing on 50 mg, 100 mg, and 150 mg eniporide doses and placebo, which correspond to $s = 1$ (50 mg vs placebo), 2 (100 mg vs placebo), 3 (150 mg vs placebo), respectively. The primary endpoint ($t = 1$) is $\alpha$-HBDH and the secondary endpoints ($t = 2$) are CK and CK-MB. One-sided p-values from the student’s t-tests are summarized in Table 2. In the proposed procedure, we set the pre-specified stagewise sample sizes according to pre-planned sample size per group in stages 1 and 2, that is, approximately, $3n_1 = n_2$. Furthermore, for each dose–placebo comparison, we set the (one-sided) local significance level such that $\alpha_s = \frac{1}{3} \cdot 0.025 = 0.008$, $s \in A = \{1, 2, 3\}$, in order to reflect the equal importance among the dose levels. Combined p-values from the inverse normal combination test in equation (3) are summarized in Table 3. In our proposed procedure, the primary endpoint, $\alpha$-HBDH, is tested at level $\alpha_s = 0.008$ for both 100 mg–placebo and 150 mg–placebo comparisons. Since $C(p(4),1,p(4),2) = 0.569 > 0.008$ and $C(p(7),1,p(7),2) = 0.105 > 0.008$, no dose is declared significantly different from placebo in the primary endpoint and hence no key secondary endpoints are considered for further hypothesis testing.

5. Concluding remarks

Clinical trials designed with multiple hypotheses and mid-course design adaptations are recognized as a modern scientific tool for accelerating medical product development programs by facilitating efficient decision making. In fact, such complex clinical trials are increasingly being
Table 2. Raw p-values of the ESCAMI study data.

| Stage       | Family         | Endpoint       | 50 mg vs placebo | 100 mg vs placebo | 150 mg vs placebo |
|-------------|----------------|----------------|------------------|------------------|-------------------|
| Primary (t = 1) | α-HBDH AUC     | $p_{1,1} = 0.599$ | $p_{4,1} = 0.136$ | $p_{7,1} = 0.003$ |
| Secondary (t = 2) | CK AUC         | $p_{2,1} = 0.434$ | $p_{5,1} = 0.040$ | $p_{8,1} = 0.002$ |
|             | CK-MB AUC      | $p_{3,1} = 0.199$ | $p_{6,1} = 0.166$ | $p_{9,1} = 0.007$ |
| Primary (t = 1) | α-HBDH AUC     | NA             | $p_{4,2} = 0.798$ | $p_{7,2} = 0.555$ |
| Secondary (t = 2) | CK AUC         | NA             | $p_{5,2} = 0.917$ | $p_{8,2} = 0.850$ |
|             | CK-MB AUC      | NA             | $p_{6,2} = 0.780$ | $p_{9,2} = 0.574$ |

P-values are one-sided and calculated based on student’s t-tests.

Table 3. Combined p-values from the inverse normal combination test.

| Family       | Endpoint       | 50 mg vs placebo | 100 mg vs placebo | 150 mg vs placebo |
|--------------|----------------|------------------|------------------|-------------------|
| Primary (t = 1) | α-HBDH AUC     | NA               | $C(p_{(4)},1,P_{(4)},2) = 0.569$ | $C(p_{(7)},1,P_{(7)},2) = 0.105$ |
| Secondary (t = 2) | CK AUC         | NA               | $C(p_{(6)},1,P_{(6)},2) = 0.627$ | $C(p_{(9)},1,P_{(9)},2) = 0.294$ |
|             | CK-MB AUC      | NA               | $C(p_{(5)},1,P_{(5)},2) = 0.573$ | $C(p_{(8)},1,P_{(8)},2) = 0.143$ |

conducted, but this necessitates the use of more sophisticated statistical methods for addressing the inherent multiplicity issues.

In this paper, we discussed an extension of the Holm procedure that is tailored to adaptive treatment selection designs with serial gatekeeping study hypotheses, based on partition idea in Huang and Hsu (2007). We discussed a general algorithm to produce mutually disjoint partitioned hypotheses for serial gatekeeping structured hypotheses in adaptive settings. The proposed procedure is simple and easy to use in practice, like the original Holm procedure, and is more powerful for hypotheses of more clinical importance in such situations. The proposed procedure can reassess a sample size at interim without compromising the FWER, as it employs the inverse-normal combination test for the final analysis. This idea of combining a multiple test procedure with inverse normal combination test would be very fruitful in many other applications in which no other multiple test procedure is applicable thus far. Possible applications include adaptive group-sequential settings where structured hypotheses are tested in a group-sequential design with mid-course design adaptations. Indeed, for such an application, a new multiple test procedure is discussed elsewhere (Sugitani et al., submitted) using the Bonferroni-based graphical approaches (Bretz et al., 2009; Maurer and Bretz, 2013a,b).

Acknowledgements

The author thanks two anonymous referees, who helped improve the presentation of the paper.
REFERENCES

Bauer, P., Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 18, 1833–1848.

Bauer, P., Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* 50, 1029–1041.

Brannath, W., Posch, M., Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* 97, 236–244.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 28, 1445–1463.

Bretz, F., Schmidli, H., Koenig, F., Racine, A., Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts (with discussion). *Biometrical Journal* 48, 623–634.

Dmitrienko, A., Tamhane, A. C. (2011). Mixtures of multiple testing procedures for gatekeeping applications in clinical trial applications. *Statistics in Medicine* 30, 1473–1488.

Finner, H., Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *The Annals of Statistics* 30, 1194–1213.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

Hommel, G. (2001). Adaptive modifications of hypothesis after an interim analysis. *Biometrical Journal* 43, 581–589.

Hsu, J. C., Berger, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. *Journal of the American Statistical Association* 94, 468–482.

Huang, Y., Hsu, J. C. (2007). Hochberg’s step-up method: cutting corners off Holm’s step-down method. *Biometrika* 94, 965–975.

Lehmacher, W., Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* 55, 1286–1290.
Li, J. (2011). A simple and flexible Holm gatekeeping procedure. *Biometrical Journal* **53**, 797–809.

Marcus, R., Peritz, E., Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Maurer, W., Bretz, F. (2013a). Memory and other properties of multiple test procedures generated by entangled graphs. *Statistics in Medicine* **32**, 1739–1753.

Maurer, W., Bretz, F. (2013b). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research* **5**, 311–320.

Posch, M., Maurer, W., Bretz, F. (2011). Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim. *Pharmaceutical Statistics* **10**, 96–104.

Sugitani, T., Bretz, F., Maurer, W. A simple and flexible graphical approach for adaptive group-sequential clinical trials, submitted.

Sugitani, T., Hamasaki, T., Hamada, C. (2013). Partition testing in confirmatory adaptive designs with structured objectives. *Biometrical Journal* **55**, 341–359.

Tamhane, A. C., Wu, Y., Mehta, C. R. (2012). Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (II): sample size re-estimation. *Statistics in Medicine* **31**, 2041–2054.

Wang, S. J., Hung, H. M. J., O’Neill, R. T. (2010). Impacts on type I error rate with inappropriate use of learn and confirm in confirmatory adaptive design trials. *Biometrical Journal* **52**, 798–810.

Ye, Y., Li, A., Liu, L., Yao, B. (2013). A group sequential Holm procedure with multiple primary endpoints. *Statistics in Medicine* **32**, 1112–1124.

Zeymer, U., Suryapranata, H., Monassier, J. P., Opolski, G., Davies, J., Rasmanis, G., Linssen, G., Tebbe, U., Schroder, R., Tiemann, R., Machnig, T., Neuhaus, K. L. (2001). The Na+/H+ exchange inhibitor Eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. *Journal of the American College of Cardiology* **38**, 1644–1651.
Appendix

A. General partition algorithm

In general, mutually disjoint hypotheses are obtained through the following two steps: In the first step, hypotheses are partitioned based on natural partition (Finner and Strassburger, 2002), as follows:

$$\Theta_{s,t}(J_{s,t}) = \begin{cases} \{ \cap_{i \in J_{s,t}} H_i \cap \{ \cap_{i \in I_{s,t} \setminus J_{s,t}} K_i \} \} & \text{for } J_{s,t} \subset I_{s,t}, J_{s,t} \neq \emptyset, \\ \cap_{i \in I_{s,t}} H_i & \text{for } J_{s,t} = I_{s,t} \end{cases} (A.1)$$

In the second step, accounting for the relative importance among multiple families of hypotheses, these partitioned hypotheses are further partitioned as

$$\Theta_{s,t}^*(J_{s,t}) = \Theta_{s,t}(J_{s,t}) \cap \Theta_{s,t-1}^c \text{ for } J_{s,t} \subseteq I_{s,t}, J_{s,t} \neq \emptyset, (A.2)$$

where $$\Theta_{s,t-1}^c = \cap_{i \in I_{s,t} \cup \cdots \cup I_{s,t-1}} K_i$$, and we define $$\Theta_{s,0}^c$$ as the whole parameter space $$\Omega$$. This two step partition algorithm can be seen as a serial gatekeeping extension of natural partition and hence provides mutually disjoint hypotheses $$\Theta_{s,t}^*(J_{s,t})$$'s for $$\Theta_s = \cup_{i \in J_{s,t}} H_i$$. Here, it should be noted that $$\Theta_s$$’s, $$s \in A$$, are not mutually disjoint, so that some multiplicity adjustments are need among $$\Theta_s$$’s. For example, this can be achieved by using the Bonferroni method as described below.

B. General decision path

Our proposed procedure uses the mutually disjoint hypotheses $$\Theta_{s,t}^*(J_{s,t})$$’s to derive an adaptive extension of Holm procedure. As noted above, however, $$\Theta_s$$’s are not mutually disjoint, so that we use the Bonferroni method to control the FWER in the strong sense. Namely, for each $$\Theta_s$$ (i.e., each treatment–control comparison), a local significance level $$\alpha_s$$ is assigned such that $$\sum_{s \in A} \alpha_s = \alpha$$. For the rejection rule of $$\Theta_{s,t}^*(J_{s,t})$$’s, we consider the following one, which can be seen as an extension of rejection rule (2) by Huang and Hsu (2007):

$$C(p_{i,1}, p_{i,2}) < \alpha_s / |J_{s,t}|, \text{ for all } i \in J_{s,t} \subseteq I_{s,t}, \tag{B.1}$$

where $$C(p_{i,1}, p_{i,2})$$ is a combined p-value from the inverse normal combination method (Lehmacher and Wassmer, 1999). Assume further that some treatments are discontinued at an interim analysis. Let $$D, D \subseteq A$$, denote the discontinued set of the treatments and $$B = A \setminus D$$, the continued set of the treatments. In our proposed procedure, hypotheses in $$I_{s,t}, s \in B, t = 1, \ldots, m$$, are ordered as $$H_{(s-1)v + \sum_{x=1}^{t-1} v_x} \succ \cdots \succ H_{(s-1)v + \sum_{x=1}^{t-1} v_x + v_t}$$, according to their observed combined p-values $$C(p_{(s-1)v + \sum_{x=1}^{t-1} v_x + v_t}, 1) < \cdots < C(p_{(s-1)v + \sum_{x=1}^{t-1} v_x}, 1)$$.

Then, based on the results by Huang and Hsu (2007), we obtain the following sequential rejective procedure for each $$\Theta_s$$:

**Step 1:** If $$C(p_{(s-1)v+1,1}, p_{(s-1)v+1,2}) < \alpha_s / v_1$$, reject $$H_{(s-1)v+1}$$, and go to the next step; otherwise, accept all hypotheses in $$I_{s,1}, \ldots, I_{s,m}$$.  

Jpn J Biomet Vol. 34, No. 2, 2013
Step 2: If $C(p((s_1-1)v+2),1,p((s_1-1)v+1),2) < \alpha_s/(v_1 - 1)$, reject $H((s_1-1)v+2)$, and go to the next step; otherwise, accept the remaining hypotheses in $I_{s,1},\ldots,I_{s,m}$.

Step $v_1$: If $C(p((s_1-1)v+v_1),1,p((s_1-1)v+v_1),2) < \alpha_s$, reject $H((s_1-1)v+v_1)$, and go to the next step; otherwise, accept the remaining hypotheses in $I_{s,1},\ldots,I_{s,m}$.

Step $v_1 + 1$: If $C(p((s_1-1)v+v_1+1),1,p((s_1-1)v+v_1+1),2) < \alpha_s/v_2$, reject $H((s_1-1)v+v_1+1)$, and go to the next step; otherwise, accept the remaining hypotheses in $I_{s,2},\ldots,I_{s,m}$.

Step $v_1 + v_2$: If $C(p((s_1-1)v+v_1+v_2),1,p((s_1-1)v+v_1+v_2),2) < \alpha_s$, reject $H((s_1-1)v+v_1+v_2)$, and go to the next step; otherwise, accept the remaining hypotheses in $I_{s,2},\ldots,I_{s,m}$.

Likewise, test the remaining hypotheses in $I_{s,3},\ldots,I_{s,m}$.

Step $v$: If $C(p(sv_1),1,p(sv_1),2) < \alpha_s$, reject $H(sv_1)$ and stop; otherwise, accept $H(sv_1)$.

As the FWER is strongly controlled at nominal level $\alpha_s$ for each $\Theta$ such that $\sum_{s \in A} \alpha_s = \alpha$, it can be easily seen from the Bonferroni inequality that our proposed procedure strongly controls the FWER at nominal level $\alpha$ for all $H_i$’s of interest.