Research Article

Statistics in Medicine

Received 19 December 2015, Accepted 19 April 2016 Published online 25 May 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6985

Graphical approaches using a Bonferroni mixture of weighted Simes tests

Kaifeng Lu[*†]

Graphical approaches to multiple testing procedures are very flexible and easy to communicate with non-statisticians. The availability of the R package gMCP further propelled the application of graphical approaches in randomized clinical trials. Bretz *et al.* (*Biometrical Journal* 2011; 53:894–913) introduced a class of nonparametric testing procedures based on a Bonferroni mixture of weighted Simes tests for intersection hypotheses. Such approaches are extremely useful when the conditions for the Simes test are known to hold for hypotheses within certain subsets but may not hold for hypotheses across subsets. We describe the calculation of adjusted *p*-values for such approaches, which is currently not available in the gMCP package. We also optimize the generation of the weights for each intersection hypothesis in the closure of a graph-based multiple testing procedure, which can dramatically reduce the computing time for simulation-based power calculations. We show the validity of the Simes test for comparing several treatments with a control, performing noninferiority and superiority tests, or testing the treatment effect in an overall and a subpopulation for the normal, binary, count, and time-to-event data. The proposed method is illustrated using an example for designing a confirmatory clinical trial. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: adjusted *p*-value; closed testing procedure; intersection hypothesis; power calculation

1. Introduction

There are many sources of multiplicity in the analysis of randomized clinical trials. Strong control of the family-wise error rate is a prerequisite for confirmatory trials to ensure reproducibility of significant findings. Methodology research for addressing multiplicity issues has seen rapid developments over the past 20 years (see the references in [1, 2]). Graphical approaches introduced in [3, 4] have given rise to intuitive, versatile, and powerful multiple testing procedures for the design and analysis of adequate and well-controlled trials. The gMCP package in R has further simplified the application and led to the increasing popularity of graphical approaches.

A class of nonparametric testing procedures based on a Bonferroni mixture of weighted Simes tests was introduced on page 907 of [4]. Such procedures can be extremely useful when one would like to use weighted Simes tests to test intersection hypotheses in the closure of a multiple testing procedure to increase the power over weighted Bonferroni tests, but the conditions underlying the Simes test cannot be assumed to hold across all elementary hypotheses except for certain subsets of hypotheses. For example, a subset of elementary hypotheses may correspond to the comparisons of several doses of an investigational product to a common control for a given endpoint, and different subsets are formed by considering different endpoints.

Equation (8) on page 907 of [4] describes the rejection rule for a Bonferroni mixture of weighted Simes tests for intersection hypotheses. However, the procedure for obtaining the adjusted *p*-values for elementary hypotheses for such approaches was not provided. In addition, such approaches are currently not implemented in the gMCP package, which may account for the lack of application of such approaches in practice.

Algorithm 1 on page 896 of [4] outlines the procedure for obtaining the weights for each elementary hypothesis within an intersection hypothesis, and the algorithm has to be repeated for each intersection

Allergan Plc, Jersey City, NJ 07311, USA
[*]*Correspondence to: Kaifeng Lu, Allergan Plc, Jersey City, NJ 07311, USA.*
[†]*E-mail: Kaifeng.lu@allergan.com*

hypothesis to generate the complete set of weights for the full closure. This can be very time-consuming when the total number of elementary hypotheses is relatively large.

We will describe the calculation of adjusted p-values for the class of nonparametric testing procedures based on a Bonferroni mixture of weighted Simes tests for the intersection hypotheses, thus filling a gap in the gMCP package. We will also optimize the generation of the weights for each intersection hypothesis in the closure of a graph-based multiple testing procedure, which can dramatically reduce the computing time for simulation-based power calculations.

2. Adjusted *p*-values for elementary hypotheses

Consider the problem of testing m elementary hypotheses $H_1, \ldots, H_m$. Let $I = \{1, \ldots, m\}$ denote the associated index set. For each $J \subseteq I$, assume a collection of weights $w_j(J)$ such that $0 \leqslant w_j(J) \leqslant 1$ and $\sum_{j \in J} w_j(J) \leqslant 1$. Let I_h, $h = 1, \ldots, l \leqslant m$, be a partition of I (i.e., $I = \cup_{h=1}^{l} I_h$ and $I_h \cap I_i = \emptyset$ for $h \neq i$) such that for each subset of hypotheses H_i, $i \in I_h$, conditions underlying the Simes test (e.g., positive regression dependence) hold for the respective test statistics. Equation (8) of [4] indicates that we can reject H_J, $J \subseteq I$, if for some j and h with $j \in J_h = I_h \cap J$,

$$p_j \leqslant \alpha \sum_{k \in J_{h,j}} w_k(J) \tag{1}$$

where $J_{h,j} = \{k \in J_h : p_k \leqslant p_j\}$.

The rejection rule (1) can be adapted to calculate the p-value for the intersection hypothesis H_J. Specifically, let $p_{(1),J_h} \leqslant p_{(2),J_h} \leqslant \cdots \leqslant p_{(m_h),J_h}$ denote the ordered p-values for the m_h elementary hypotheses in H_{J_h}, and let these elementary hypotheses have weights (in J) equal to $w_{(1),J_h}(J)$, $w_{(2),J_h}(J)$, $\ldots$, $w_{(m_h),J_h}(J)$, respectively. The p-value for the intersection hypothesis H_J based on the Bonferroni mixture of weighted Simes tests within each H_{J_h} is defined as

$$p_J = \min_{1 \leqslant h \leqslant l} \min_{1 \leqslant j \leqslant m_h} p_{(j),J_h} / \sum_{i=1}^{j} w_{(i),J_h}(J) \tag{2}$$

The equivalence of $p_J \leqslant \alpha$ and the rejection rule (1) is easy to verify.

By the closure principle, the adjusted p-value for each hypothesis H_i can be obtained by computing the p-value for each intersection hypothesis H_J with $i \in J$ and then taking the maximum over them, that is,

$$\tilde{p}_i = \max_{J : i \in J} p_J \tag{3}$$

For example, consider the weighting strategy from example 1 in [4] but with $p_1 = 0.01$, $p_2 = 0.005$, $p_3 = 0.015$, and $p_4 = 0.022$ (also see page 907 of [4]). Let $I_1 = \{1, 2\}$, $I_2 = \{3, 4\}$. It can be verified that the p-values for the 15 intersection hypotheses in table 1 in [4] are equal to 0.01, 0.01, 0.01, 0.01, 0.02, 0.01, 0.02, 0.01, 0.01, 0.01, 0.005, 0.005, 0.022, 0.015, and 0.022, respectively. The adjusted p-values for the elementary hypotheses are given by $\tilde{p}_1 = 0.02$, $\tilde{p}_2 = 0.01$, $\tilde{p}_3 = 0.022$, and $\tilde{p}_4 = 0.022$, which are identical to the adjusted p-values for the elementary hypotheses when weighted Simes tests are used for the intersection hypotheses. In comparison, the adjusted p-values for the elementary hypotheses when weighted Bonferroni tests are used for the intersection hypotheses are $\tilde{p}_1 = 0.02$, $\tilde{p}_2 = 0.01$, $\tilde{p}_3 = 0.03$, and $\tilde{p}_4 = 0.03$. For this example, at $\alpha = 0.025$, the closed weighted Simes test and the closed Bonferroni mixture of weighted Simes tests would reject all four hypotheses, two more than with the closed weighted Bonferroni test.

The R code for performing the two closed test procedures is as follows:

```
library(gMCP)
pvalues <- c(0.01,0.005,0.015,0.022)
w <- c(0.5,0.5,0,0)
g <- matrix(c(0,0,1,0,0,0,0,1,0,1,0,0,1,0,0,0), nrow=4, ncol=4,
byrow=T)
graph <- matrix2graph(g, w)
gMCP(graph, pvalues, alpha=0.025) # weighted Bonferroni
gMCP(graph, pvalues, test="Simes", alpha=0.025) # weighted Simes
```

The adjusted p-values for the elementary hypotheses based on the closed Bonferroni mixture of weighted Simes tests can be obtained as follows:

```
subsets <- list(c(1,2), c(3,4))
fadjpsim(wgtmat, pvalues, subsets)
# Bonferroni mixture of weighted Simes
```

where the R function `fadjpsim` is given in Appendix A and the weight matrix for the intersection hypotheses `wgtmat` is given in table 1 of [4]. Adjusted p-values for the closed weighted Bonferroni procedure can be obtained by setting `subsets <- list(1,2,3,4)`. Similarly, adjusted p-values for the closed weighted Simes procedure can be obtained by setting `subsets <- list(c(1,2,3,4))`. The function `fadjpsim` takes a matrix of raw p-values for the elementary hypotheses as the input for `pvalues`, and hence, it facilitates simulation-based power calculations

3. Weights of intersection hypotheses

The extended graphical approaches introduced in [4] dissociate the underlying weighting strategy from the employed test procedure. Such an approach enhances transparency by first deriving suitable weighting strategies that reflect the given study objectives and subsequently applying appropriate test procedures that do not necessarily have to be based on Bonferroni's inequality. When the adopted test procedure is not consonant (e.g., weighted Simes), a shortcut is not always possible, in which case, we need to compute the weights of all intersection hypotheses.

Consider a graph defined by a vector of weights initially assigned to each elementary hypothesis, $\{w_i(I) : i = 1, \ldots, m\}$, and a transition matrix that specifies how much of the local weight assigned to the hypothesis corresponding to its row index is passed on to the hypothesis corresponding to its column index, $\mathbf{G} = (g_{ij})$, where $0 \leqslant g_{ij} \leqslant 1$, $g_{ii} = 0$ and $\sum_{j=1}^{m} g_{ij} \leqslant 1$ for all $i,j \in I$. For a given index set $J \subseteq I$, algorithm 1 on page 896 of [4] specifies how the graph is updated once a vertex $j \in J^c$ is removed, which, in turn, determines the weights $w_j(J)$, $j \in J$, by removing all vertices in J^c one at a time. This algorithm is then repeated for each $J \subseteq I$ to generate the $m(2^{m-1} - 1)$ weights (excluding the initial m ones) for the full closure. Using this method, the operation of removing a vertex and updating the graph has to be performe

$$\sum_{j=1}^{m-1} \binom{m}{j} (m - j) = m(2^{m-1} - 1) \tag{4}$$

times. This follows because there are a total of $\binom{m}{j}$ index sets $J \subseteq I$ with j elements, and for each of such index sets, one needs to remove $m - j$ vertices from J^c, for $j = 1, \ldots, m - 1$. Thus, the computation can be heavy for relatively large m. This algorithm is implemented via the `generateWeights` function in the gMCP package.

By ordering the intersection hypotheses in a particular way, one can reduce the number of operations of removing a vertex and updating the graph. We start with the global null hypothesis, H_I, for which, the weight vector and the transition matrix are specified by the initial graph and no removal of a vertex is needed. Let $S_m = \{H_I\}$. Next, we consider the set of intersection hypotheses with $m - 1$ elements, $S_{m-1} = \{H_J : |J| = m - 1\}$. There are $\binom{m}{m-1}$ such intersection hypotheses, and the weight vector and the transition matrix for each such intersection hypothesis can be obtained in one step by removing the single vertex from its complement index set and updating the graph associated with H_I. Keep the set of weight vectors and transition matrices for S_{m-1} in the computer memory. Now, we consider the set of intersection hypotheses with $m - 2$ elements, $S_{m-2} = \{H_J : |J| = m - 2\}$. There are $\binom{m}{m-2}$ such intersection hypotheses, and the weight vector and the transition matrix for each such intersection hypothesis, for example, $H_J \in S_{m-2}$, can be obtained in one step by removing a vertex $j \in J^c$ and updating the graph associated with $H_{J\cup\{j\}} \in S_{m-1}$. Keep the set of weight vectors and transition matrices for S_{m-2} in the computer memory. We proceed in this fashion until we have obtained the weight vector and the transition matrix for each of the $\binom{m}{1}$ intersection hypotheses with only one element, that is, the set of elementary hypotheses, $S_1 = \{H_j : j = 1, \ldots, m\}$. Using this method, the operation of removing a vertex and updating the graph only needs to be performed

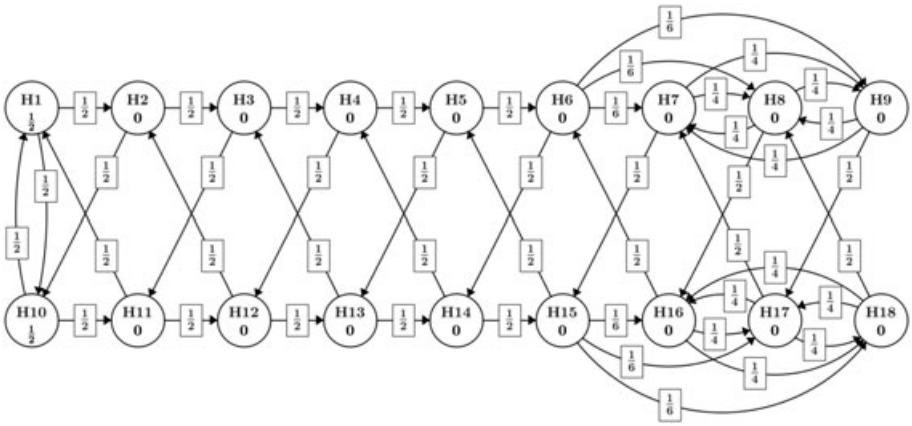

Figure 1. Weighting strategy for multiple comparisons across two doses of an experimental drug and the primary and secondary endpoints. H1 to H9 denote the hypotheses for the two co-primary and seven secondary endpoints associated with the high dose versus placebo comparisons, and H10 to H18 denote the corresponding hypotheses associated with the low dose versus placebo comparisons.

$$\sum_{j=1}^{m-1} \binom{m}{j} = 2^m - 2 \tag{5}$$

times. Comparing (5) with (4), the modified method only requires $2/m$ of the number of operations needed for the original method, which can represent a significant saving for relatively large m. The modified method can be implemented using the R function `fwgtmat` given in Appendix B.

To illustrate the potential time savings of the modified method, consider a multiple attack migraine study that compares two doses of an experimental drug to placebo with respect to two co-primary efficacy endpoints and seven secondary efficacy endpoints. The co-primary efficacy endpoints are pain freedom at 2 h post-dose for the first attack and pain freedom consistency across multiple attacks. The secondary efficacy endpoints include sustained pain freedom from 2 to 24 h post-dose for the first attack, pain relief at 2 h post-dose for the first attack, sustained pain relief from 2 to 24 h post-dose for the first attack, pain relief consistency across multiple attacks, absence of photophobia at 2 h post-dose for the first attack, absence of phonophobia at 2 h post-dose for the first attack, and absence of nausea at 2 h post-dose for the first attack. Figure 1 depicts the weighting strategy for a graphical approach for multiple comparisons. The ordering of the endpoints accounts for both clinical relevance and statistical power of each endpoint. For example, pain freedom is considered to be a more clinically relevant target than pain relief, and the symptom endpoints have less statistical power than the pain endpoints for migraine studies. The recycling of weights among the three secondary efficacy endpoints for migraine-associated symptoms helps to significantly increase the power for nausea (which is the most difficult endpoint to detect the treatment effect) while incurring minimal sacrifice on the power for the other two symptom endpoints. Additional power gain is achieved by allowing the recycling of weights between the two doses. The fraction of weights to shift from one dose to the other dose can be fine tuned to optimize the power performance. When executed on a Windows 7 computer with Intel Core i5 CPU, it took about 8.9 min for the `generateWeights` function in the gMCP package to compute the weight matrix for the intersection hypotheses, while it took about 46 s for the `fwgtmat` function to obtain the same weight matrix. The time savings are substantial especially if one needs to compare the performance of different weighting strategies using large-scale simulations. Finally, the co-primary endpoints will serve as the gatekeeper for the secondary endpoints within each dose of the experimental drug so that the null hypotheses associated with the secondary endpoints cannot be rejected unless the null hypotheses associated with the co-primary endpoints have been rejected for the same dose.

4. Validity of Simes test in specific settings

To apply the closure principle to graphical approaches using a Bonferroni mixture of weighted Simes tests, we need to establish the validity of the Simes test for pre-specified subsets of null hypotheses. Sarkar

and Chang [5] proved the Simes inequality for a class of positively dependent multivariate distributions that are conditionally independent and identically distributed with a distribution that is stochastically increasing in the value of the conditioning variable. This class contains many continuous multivariate distributions commonly encountered in multiple hypothesis testing situations such as the equicorrelated multivariate normal with zero means, unit variances, and a nonnegative correlation. Sarkar [6] showed that the Simes inequality holds for multivariate totally positive of order 2 random variables with common marginals. The class of multivariate totally positive of order 2 distributions is larger than the one considered in Sarkar and Chang [5]. As indicated in Sarkar [7], the proof of theorem 3.1 in Sarkar [6] also establishes the Simes inequality for a slightly larger class of distributions satisfying the positive dependence through stochastic ordering condition, which is essentially the same as the positive regression dependence on subset condition. In particular, the Simes inequality holds for multivariate normal with nonnegative correlations. Theorem 3.1 in Sarkar [7] established the Simes inequality for standard multivariate t with nonnegative correlations under certain sign restrictions. To see that equations (1.2) and (1.3) in Sarkar [7] imply $P(P_{(j)} \geqslant j\alpha/m, j = 1, \ldots, m) \geqslant 1 - \alpha$ for X's having a common marginal distribution F, where P_j denotes the random p-value corresponding to a left-tailed or right-tailed test based on test statistics X_j, we note that P_j is defined as $F(X_j)$ for a left-tailed test and as $1 - F(X_j)$ for a right-tailed test and take $a_j = F^{-1}(j\alpha/m)$ and $b_j = F^{-1}(1 - (m - j + 1)\alpha/m)$ for $j = 1, \ldots, m$ in equations (1.2) and (1.3), respectively. For standard multivariate t with nonnegative correlations, F is the marginal univariate t and $a_m = F^{-1}(\alpha) < 0$ and $b_1 = F^{-1}(1 - \alpha) > 0$ for $\alpha < 0.5$.

We will use these results to establish the validity of the Simes test when we compare several treatments with a control, perform noninferiority and superiority tests, or test the treatment effect in an overall and a subpopulation on the same endpoint.

4.1. Many-to-one comparisons

Normal data. Suppose there are $m + 1$ independent $N(\mu_j, \sigma^2)$ populations with unknown μ_j and σ^2, $j = 0, 1, \ldots, m$. Consider the problem of testing $H_j : \mu_j - \mu_0 \leqslant 0$ (or $\geqslant 0$) against $\bar{H}_j : \mu_j - \mu_0 > 0$ (or < 0) for $j = 1, \ldots, m$, based on independent samples of size n_j from the jth population for $j = 0, 1, \ldots, m$. The estimated mean difference between treatment j and the control is $\hat{\beta}_j = \bar{y}_j - \bar{y}_0, j = 1, \ldots, m$. The estimated covariance matrix for the estimated mean differences is given by

$$V_{\hat{\beta}} = \hat{\sigma}^2 \left(\text{diag} \left(\left\{ \frac{1}{n_j} \right\} \right) + \frac{1}{n_0} J_m \right)$$

where $\hat{\sigma}^2$ denotes the usual pooled variance estimate, diag(a) a diagonal matrix with the jth diagonal element equal to the jth element of vector a, and J_m an $m \times m$ matrix with all elements equal to one. As pointed out in [5], in establishing the conservativeness of the Simes test for testing any of the two types of null hypotheses, it may be assumed without any loss of generality that the null hypotheses are $H_j^* : \mu_j = \mu_0$, for $j = 1, \ldots, m$, which are to be tested against the same left-tailed or right-tailed alternatives. The joint distribution of the Wald test statistics $\left\{ \hat{\beta}_j / V_{\hat{\beta}_j}^{1/2} \right\}$ when $\mu_j = \mu_0$ for all j is standard multivariate t with the associated multivariate normal having positive correlations.

Binary data. Suppose there are $m + 1$ independent Bern(π_j) populations with unknown success probabilities π_j, $j = 0, 1, \ldots, m$. Consider the problem of testing $H_j : \pi_j - \pi_0 \leqslant 0$ (or $\geqslant 0$) against $\bar{H}_j : \pi_j - \pi_0 > 0$ (or < 0) for $j = 1, \ldots, m$, based on independent samples of size n_j from the jth population for $j = 0, 1, \ldots, m$. The estimated log odds ratio for treatment j versus the control is $\hat{\beta}_j = \log(p_j/(1 - p_j)) - \log(p_0/(1 - p_0)), j = 1, \ldots, m$, where p_j is the estimated proportion of successes in treatment j. The estimated covariance matrix for the estimated log odds ratios is given by

$$V_{\hat{\beta}} = \text{diag} \left(\left\{ \frac{1}{n_j p_j (1 - p_j)} \right\} \right) + \frac{1}{n_0 p_0 (1 - p_0)} J_m$$

The joint distribution of the Wald test statistics $\left\{ \hat{\beta}_j / V_{\hat{\beta}_j}^{1/2} \right\}$ when $\pi_j = \pi_0$ for all j is asymptotically multivariate normal with zero means, unit variances, and positive correlations.

Count data. Suppose there are $m+1$ independent Negbin(λ_j, κ) populations with unknown rate parameters $\lambda_j, j = 0, 1, \ldots, m$, and dispersion parameter κ. The density function for a subject in the jth population with exposure t_i is

$$p(y_i|\lambda_j,\kappa,t_i)=\frac{\Gamma(y_i+1/\kappa)}{y_i!\Gamma(1/\kappa)}\frac{(\kappa\lambda_j t_i)^{y_i}}{(1+\kappa\lambda_j t_i)^{y_i+1/\kappa}}$$

Consider the problem of testing $H_j:\lambda_j-\lambda_0\leqslant 0$ (or $\geqslant 0$) against $\bar{H}_j:\lambda_j-\lambda_0>0$ (or <0) for $j=1,\ldots,m$, based on independent samples of size n_j from the jth population for $j=0,1,\ldots,m$. The estimated log rate ratio for treatment j versus the control is $\hat{\beta}_j=\log\left(\hat{\lambda}_j\right)-\log\left(\hat{\lambda}_0\right), j=1,\ldots,m$, where $\hat{\lambda}_j$ denotes the estimated rate parameter for treatment j. For the special case where all subjects in treatment j have the same exposure t_j, $\hat{\lambda}_j=\sum_{i=1}^{n}y_i x_{ij}/(n_j t_j)$, where $x_{ij}=1$ if subject i is on treatment j and 0 otherwise. The estimated covariance matrix for the estimated log rate ratios is given by

$$V_{\hat{\beta}}=\mathrm{diag}\left(\left\{\frac{1+\hat{\kappa}\hat{\mu}_j}{n_j\hat{\mu}_j}\right\}\right)+\frac{1+\hat{\kappa}\hat{\mu}_0}{n_0\hat{\mu}_0}J_m$$

where $\hat{\mu}_j=\hat{\lambda}_j t_j$ for $j=0,1,\ldots,m$. The joint distribution of the Wald test statistics $\left\{\hat{\beta}_j/V_{\hat{\beta}_j}^{1/2}\right\}$ when $\lambda_j=\lambda_0$ for all j is asymptotically multivariate normal with zero means, unit variances, and positive correlations. Similar result is established in Appendix C for the general case when the treatment exposure may vary across subjects within a treatment.

Time-to-event data. Suppose there are $m+1$ independent populations for failure times with survival functions $S_j(t)$, $j=0,1,\ldots,m$, and assume a proportional hazards model, $S_j(t)=S_0(t)^{\exp(\beta_j)}$ for $j=1,\ldots,m$, where β_j denotes the log hazard ratio for treatment j versus the control. Consider the problem of testing $H_j:\beta_j\leqslant 0$ (or $\geqslant 0$) against $\bar{H}_j:\beta_j>0$ (or <0) for $j=1,\ldots,m$, based on independent samples of size n_j from the jth population for $j=0,1,\ldots,m$, subject to random censoring. For the special case of exponential failure time distributions with $S_j(t)=e^{-\lambda_j t}$, $\beta_j=\log(\lambda_j)-\log(\lambda_0)$, and $\hat{\lambda}_j=D_j/T_j$, where D_j and T_j denote the total number of events and the total exposure for treatment j, respectively. The estimated covariance matrix for the estimated log hazard ratios is given by

$$V_{\hat{\beta}}=\mathrm{diag}\left(\left\{\frac{1}{D_j}\right\}\right)+\frac{1}{D_0}J_m$$

The joint distribution of the Wald test statistics $\left\{\hat{\beta}_j/V_{\hat{\beta}_j}^{1/2}\right\}$ when $\beta_j=0$ for all j is asymptotically multivariate normal with zero means, unit variances, and positive correlations. Similar result is established in Appendix D for general failure time distributions. Of note, the log-rank test statistics are asymptotically multivariate normal with zero means, unit variances, and negative correlations, and hence the Simes test may not be valid for the log-rank test statistics.

4.2. *Combined noninferiority and superiority testing*

Consider the noninferiority hypothesis $H_1:\beta\leqslant-\delta$, where $\delta>0$ is the pre-specified noninferiority margin. Let $H_2:\beta\leqslant 0$ be the corresponding superiority hypothesis and $H_3,\ldots,H_m$ other null hypotheses included in the multiplicity scheme. Because $H_1\cap\bar{H}_2=\emptyset$, the hypotheses $H_1,\ldots,H_m$ do not satisfy the free combination condition. The reduced closure tree consists of intersection hypotheses in the full closure tree except those containing $H_1\cap H_2$. We will show that applying the Bonferroni mixture of weighted Simes tests to the full closure tree is equivalent to applying the procedure to the reduced closure tree, assuming that the node H_2 has zero weight and has only one incoming edge in the initial graph and that edge is initiated from the node H_1. The node H_1 can have outgoing edges other than the one pointing to H_2, and the node H_2 can have its own outgoing edges in the initial graph.

Consider an intersection hypothesis H_J not in the reduced closure tree. By definition, $J\supseteq\{1,2\}$. We claim that $w_2(J)=0$. By algorithm 1 on page 896 of [4], we need to remove the nodes in J^c one by one to obtain $w_2(J)$. First, remove a node $j\in J^c$ from the initial graph on I. Because there is no directed edge from H_j to H_2 in the initial graph, by algorithm 1, $w_2(I\setminus\{j\})=0$ in the updated graph. For all $\ell\in I\setminus\{j,1,2\}$, because $g_{\ell 2}=g_{j2}=0$ in the initial graph, by algorithm 1, $g_{\ell 2}(I\setminus\{j\})=0$ in the updated graph. Because $g_{12}>0$ in the initial graph, we have $g_{1j}<1$ so that $g_{1j}g_{j1}<1$, and because $g_{j2}=0$, by Algorithm 1, $g_{12}(I\setminus\{j\})=g_{12}/(1-g_{1j}g_{j1})>0$ in the updated graph. Therefore, the node H_2 retains the assumed properties in the updated graph as in the initial graph. By proceeding with the removal of other nodes in J^c one at a time, and by deduction, $w_2(J)=0$ as claimed.

Now, we will show that $p_J \leqslant p_{J\setminus\{2\}}$, where p_J is the p-value defined in (2) for the intersection hypothesis H_J based on the Bonferroni mixture of weighted Simes tests within each H_{J_h}. Without loss of generality, let $J_1 = J \cap I_1 = \{1, 2, \ldots, m_1\}$. The term in (2) corresponding to J_1 can be written as

$$\min_{1\leqslant j\leqslant m_1} \frac{p_j}{\sum_{1\leqslant i\leqslant m_1} w_i(J) I(p_i \leqslant p_j)}$$

which is less than or equal to

$$\min_{1\leqslant j\leqslant m_1, j\neq 2} \frac{p_j}{\sum_{1\leqslant i\leqslant m_1} w_i(J) I(p_i \leqslant p_j)} = \min_{1\leqslant j\leqslant m_1, j\neq 2} \frac{p_j}{\sum_{1\leqslant i\leqslant m_1, i\neq 2} w_i(J) I(p_i \leqslant p_j)}$$

where the aforementioned equation holds because $w_2(J) = 0$. On the other hand, the right-hand side of the aforementioned equation is exactly the term in $p_{J\setminus\{2\}}$ corresponding to $(J \setminus \{2\}) \cap I_1 = J_1 \setminus \{2\}$. Because the other terms in p_J match the corresponding terms in $p_{J\setminus\{2\}}$, we have $p_J \leqslant p_{J\setminus\{2\}}$.

Therefore, by (3), the adjusted p-value for any elementary hypothesis based on the full closure tree is identical to that based on the reduced closure tree, assuming that the node H_2 has zero weight and has only one incoming edge in the initial graph and that edge is initiated from the node H_1. When this assumption is violated, the adjusted p-value for any elementary hypothesis based on the full closure tree is larger than or equal to that based on the reduced closure tree and hence is more conservative. It follows that we can include the combined noninferiority and superiority testing in the Simes test framework.

4.3. Testing in an overall and a subpopulation

Consider stratified randomization with stratum 1 representing the subpopulation of interest and stratum 2 the complement subpopulation. Let $\hat{\beta}_i$ denote the estimated treatment effect (e.g., mean difference, log odds ratio, log rate ratio, or log hazard ratio) in stratum $i = 1, 2$, and $\hat{\beta} = w\hat{\beta}_1 + (1-w)\hat{\beta}_2$ the estimated treatment effect for the overall population, where w denotes the weight for the subpopulation of interest. We have $C_{\hat{\beta},\hat{\beta}_1} = wV_{\hat{\beta}_1}$. If $V_{\hat{\beta}_i}$ has only nonnegative off-diagonal elements for $i = 1, 2$, then $V_{\hat{\beta}} = w^2 V_{\hat{\beta}_1} + (1-w)^2 V_{\hat{\beta}_2}$ and $V_{(\hat{\beta}', \hat{\beta}_1')'}$ have only nonnegative off-diagonal elements, in which case, the joint distribution of the Wald test statistics for testing $\beta = 0$ and $\beta_1 = 0$ is asymptotically multivariate normal with zero means, unit variances, and nonnegative correlations. The asymptotic multivariate normal distribution can be replaced with standard multivariate t for normal samples.

In summary, for normally distributed data, because the raw p-values for the elementary hypotheses are usually based on the t distribution, we rely on the standard multivariate t with the associated multivariate normal having nonnegative correlations to justify the use of the Simes test. For binary, count, or time-to-event data, because the raw p values for the elementary hypotheses are usually based on the asymptotic normal distribution, we rely on the asymptotic multivariate normal with nonnegative correlations to justify the use of the Simes test.

5. Example

We will illustrate the application of the closed Bonferroni mixture of weighted Simes tests to the construction of weighting strategies and the power calculations for a graph-based multiple testing procedure. Specifically, we will consider the multiplicity problem discussed in [8], where there are a total of six null hypotheses of interest as follows:

- H_1: high dose is inferior to active control for the primary endpoint;
- H_2: low dose is inferior to active control for the primary endpoint;
- H_3: high dose is not superior to active control for the primary endpoint;
- H_4: high dose is not superior to active control for the secondary endpoint;
- H_5: low dose is not superior to active control for the primary endpoint; and
- H_6: low dose is not superior to active control for the secondary endpoint.

In addition, assume that we will adopt the graphical approach with weighting strategy depicted in figure 4 of [8], which is redrawn in Figure 2 with a tuning parameter γ in place of the infinitesimal probability ϵ. A similar graph was used in Hung and Wang [9] in the context of testing the noninferiority and superiority hypotheses for a primary and a secondary endpoint.

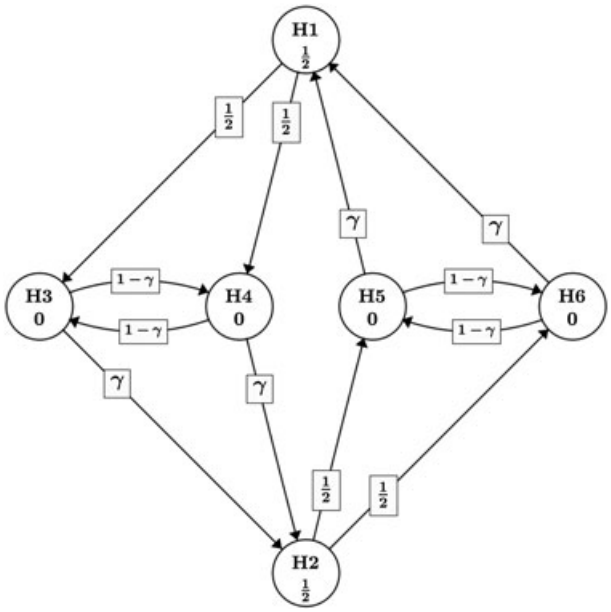

Figure 2. Weighting strategy for multiple comparisons across two doses of an experimental drug and the primary and secondary endpoints. H1 and H3 denote the noninferiority and superiority hypotheses for the primary endpoint on the high dose, H4 denotes the superiority hypothesis for the secondary endpoint on the high dose, and H2, H5, and H6 denote the corresponding hypotheses on the low dose. All comparisons are against the active control.

To simulate data, assume that the primary and secondary endpoints for a subject in treatment i ($i = 0, 1, 2$) have a bivariate normal distribution with means μ_{i1} and μ_{i2}, variances σ_1^2 and σ_2^2, and correlation ρ. Here, $i = 0$ denotes the active control, $i = 1$ the high dose, and $i = 2$ the low dose. Without loss of generality, assume that the variances are equal to one, that is, $\sigma_1^2 = \sigma_2^2 = 1$. Let $\rho = 0.4$. For the primary endpoint, let $\mu_{01} = 0.40$, $\mu_{11} = 0.63$, $\mu_{21} = 0.45$, and the noninferiority margin $\Delta_1 = 0.20$. For the secondary endpoint, let $\mu_{02} = 0.10$, $\mu_{12} = 0.45$, and $\mu_{22} = 0.40$. Furthermore, we assume 1:1:1 randomization with $n_0 = n_1 = n_2 = 220$ subjects per treatment arm. It is easy to verify that the marginal power for the six elementary hypotheses are 0.99, 0.74, 0.67, 0.96, 0.08, and 0.88, respectively. Similar to the strategy used in [10], instead of simulating subject level data, we can directly simulate treatment level data. Specifically, first we generate a 2×2 random matrix $W = \{W_{ij}\}$ from a Wishart distribution with degrees of freedom $n - 3$ and scale matrix $Q = (1 - \rho)I_2 + \rho J_2$, where $n = n_0 + n_1 + n_2$ denotes the total sample size, I_k an identity matrix with k rows and columns, and J_k a $k \times k$ matrix with all elements equal to 1. Then, we generate a 2×1 random vector $u_i = (u_{i1}, u_{i2})'$ from a bivariate normal distribution with mean vector $(\mu_{i1}, \mu_{i2})'$ and covariance matrix Q/n_i, for $i = 0, 1, 2$. Finally, we construct the test statistics to test the six elementary hypotheses,

$$T_1 = \frac{u_{11} - u_{01} + \Delta_1}{\hat{\sigma}_1 \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}$$

$$T_2 = \frac{u_{21} - u_{01} + \Delta_1}{\hat{\sigma}_1 \sqrt{\frac{1}{n_2} + \frac{1}{n_0}}}$$

$$T_3 = \frac{u_{11} - u_{01}}{\hat{\sigma}_1 \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}$$

$$T_4 = \frac{u_{12} - u_{02}}{\hat{\sigma}_2 \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}$$

$$T_5 = \frac{u_{21} - u_{01}}{\hat{\sigma}_1 \sqrt{\frac{1}{n_2} + \frac{1}{n_0}}}$$

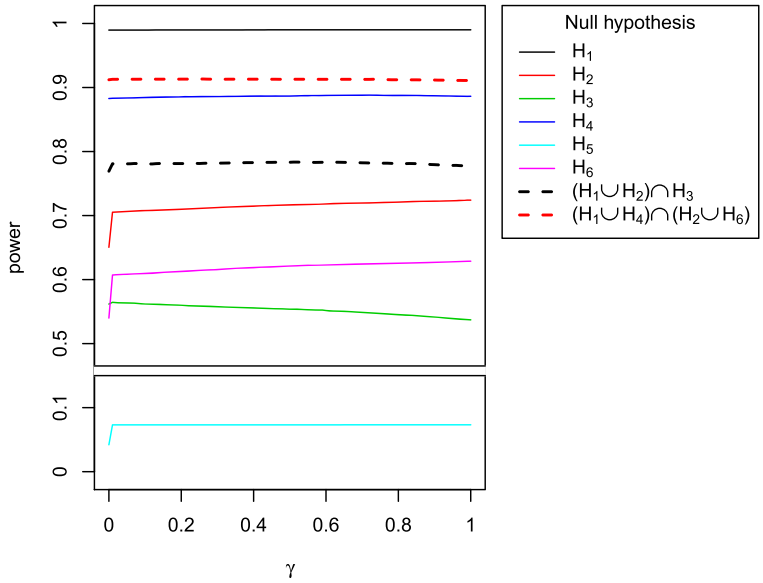

Figure 3. Simulated adjusted power for elementary hypotheses and two objective functions based on 10,000 replications using a Bonferroni mixture of weighted Simes tests with weighting strategy in Figure 2 and with subsets $I_1 = \{1, 2, 3, 5\}$ and $I_2 = \{4, 6\}$.

$$T_6 = \frac{u_{22} - u_{02}}{\hat{\sigma}_2 \sqrt{\frac{1}{n_2} + \frac{1}{n_0}}}$$

where $\hat{\sigma}_1^2 = W_{11}/(n-3)$ and $\hat{\sigma}_2^2 = W_{22}/(n-3)$.

Using the results from Section 4, the Simes test can be applied to hypotheses within each endpoint. Using the notation from Section 2, $I = \{1, 2, 3, 4, 5, 6\} = I_1 \cup I_2$, where $I_1 = \{1, 2, 3, 5\}$, $I_2 = \{4, 6\}$. We will use the following two objective functions to determine the optimal values of the tuning parameter γ:

- Probability of demonstrating either noninferiority on both doses or superiority on the high dose for the primary endpoint. The corresponding null hypothesis is $(H_1 \cup H_2) \cap H_3$.
- Probability of demonstrating noninferiority on primary endpoint and superiority on secondary endpoint for either the high or low dose. The corresponding null hypothesis is $(H_1 \cup H_4) \cap (H_2 \cup H_6)$.

Similar to [11, 12], we will perform simulations to estimate the power for the elementary hypotheses and the two objective functions. Here, we assume that the Bonferroni mixture of weighted Simes tests with weighting strategy in Figure 2 and with subsets $I_1 = \{1, 2, 3, 5\}$ and $I_2 = \{4, 6\}$ will be used as multiplicity adjustment.

Figure 3 depicts the curves of adjusted power for the elementary hypotheses and the two objective functions based on 10,000 replications for each value of $\gamma = 0(0.01)1$. The adjusted power for H_2 and H_6 was increased by more than 5% by increasing γ from 0 to 0.01. An additional 2% gain in the adjusted power for H_2 and H_6 can be achieved by increasing γ from 0.01 to 1 but at the expense of about 3% loss in the adjusted power for H_3. The power curves for the other elementary hypotheses and the two objective functions were relatively flat with respect to the value of γ. The value of the average of the two objective functions ranged from 0.844 to 0.848 for $\gamma \in [0.01, 1]$, with the maximum value attained around $\gamma = 0.5$, but the maximizer did not make much practical difference given the narrow range of the objective function values.

Because the power for H_5 is very low, it might be sensible to drop H_5 from the family of hypotheses of interest in hopes that the power for the other hypotheses can be improved. Figure 4 depicts a revised weighting strategy with H_5 removed. Table I compares the adjusted power for the two weighting strategies. The removal of H_5 resulted in the largest increase in the adjusted power for H_3, followed by the adjusted power for H_6 and then H_4. There was virtual no impact on the adjusted power for H_1 and H_2. Overall, the results support the removal of H_5 from the multiplicity scheme. For comparison, we also included the closed weighted Bonferroni tests. For the weighting strategy in Figure 4, the use of the closed Bonferroni mixture of weighted Simes tests led to approximately 1% increase in the adjusted power for

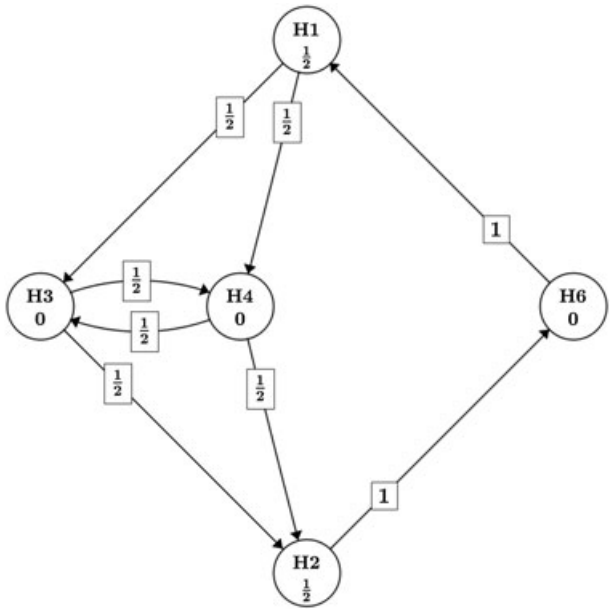

Figure 4. Revised weighting strategy for multiple comparisons across two doses of an experimental drug and the primary and secondary endpoints. The hypotheses are defined in Figure 2.

Table I. Simulated adjusted power for elementary hypotheses and two objective functions based on 10,000 replications.

	Power for rejecting the null hypothesis			
	Figure 2 with $\gamma = 0.5$		Figure 4	
Null hypothesis	Bonferroni	BonSimes	Bonferroni	BonSimes
H_1	0.990	0.990	0.990	0.991
H_2	0.706	0.717	0.706	0.717
H_3	0.550	0.554	0.612	0.621
H_4	0.885	0.887	0.897	0.900
H_5	0.071	0.073	—	—
H_6	0.612	0.621	0.647	0.657
$(H_1 \cup H_2) \cap H_3$	0.773	0.783	0.773	0.784
$(H_1 \cup H_4) \cap (H_2 \cup H_6)$	0.910	0.913	0.917	0.920

Bonferroni denotes the closed weighted Bonferroni tests, and BonSimes denotes the closed Bonferroni mixture of weighted Simes tests. The hypotheses are defined in Figure 2.

hypotheses H_2, H_3, and H_6, and the first objective function. This improvement in power is equivalent to what could be achieved using the closed weighted Bonferroni procedure by adding about six subjects per group. When the number of elementary hypotheses is small, the closed weighted Bonferroni procedure is not very conservative.

6. Discussion

Graphical approaches to multiple testing procedures have become a well-established tool in designing confirmatory clinical trials. Graphical approaches based on weighted Bonferroni tests are very flexible and do not require any assumption about the joint distribution of the test statistics or any specific dependence structure among them. Power can be improved by using weighted Simes tests when the test statistics satisfy certain positive dependence conditions such as positive regression dependence, which holds for multivariate normal with nonnegative correlations. Using the results from [7], we were able to confirm the validity of the Simes test (at least asymptotically) for the analysis of normal, binary, count, or

time-to-event data when comparing several treatments to a control, performing noninferiority and superiority tests, or testing the treatment effect in an overall and a subpopulation on the same endpoint. Example 3 in [4] presents a multiplicity problem involving a combination of these factors. However, it is often difficult to verify the conditions for the Simes test if the intersection hypothesis involves different endpoints. In these cases, graphical approaches based on a Bonferroni mixture of weighted Simes tests can be very useful.

The R package gMCP has a user-friendly interface, which greatly simplifies the specification and implementation of graphical approaches based on weighted Bonferroni or weighted Simes tests. However, graphical approaches based on a Bonferroni mixture of weighted Simes tests are currently not available in the gMCP package. We have defined adjusted *p*-values for graphical approaches based on a Bonferroni mixture of weighted Simes tests and provided a R function for the calculation. In addition, we have described an improved algorithm for generating the weight matrix for intersection hypotheses in the closure of a graph-based multiple testing procedure and provided an accompanying R function. We recommend the separation of the task of generating the weight matrix and the task of calculating adjusted *p*-values for graphical approaches based on weighted Simes tests or a Bonferroni mixture of weighted Simes tests. This can dramatically reduce the computing time for simulation-based power calculations.

We have illustrated our proposed method using an example for designing a confirmatory clinical trial involving two doses of an experimental drug compared with an active control, noninferiority and superiority tests, and primary and secondary endpoints. We recommend the use of simulation studies to select the values of tuning parameters to optimize power functions of interest. We also recommend the exclusion of hypotheses which have very low power to be rejected from the set of hypotheses considered for multiplicity control. This can dramatically increase the power for other hypotheses which have a more realistic chance to be rejected.

Instead of weighted Simes tests, weighted parametric tests (e.g., Dunnett type max-*t* or min-*p* tests) may also be used for prespecified subsets of hypotheses if the exact or asymptotic multivariate distributions of the test statistics are available. Chapter 3 of [2] describes multiple comparison procedures in general parametric models, and section 3.2 of [4] illustrates graphical approaches using a Bonferroni mixture of weighted parametric tests.

Appendix A: R function for obtaining adjusted *p*-values for a closed Bonferroni mixture of weighted Simes tests

```
fadjpsim <- function(wgtmat, pvalues, subsets) {
    ntests = nrow(wgtmat)
    m = ncol(wgtmat)
    pvalues = matrix(pvalues, ncol=m) # ensure pvalues is a matrix
    r = nrow(pvalues)
    nfamil = length(subsets)
    famil = matrix(0, nfamil, m)
    for (i in 1:nfamil) famil[i,subsets[[i]]] = 1
    sim = 1:r
    pinter = matrix(0, r, ntests)
    incid = matrix(0, ntests, m)
    for (i in 1:ntests) {
        number = ntests - i + 1
        cc = floor(number/2^(m - (1:m))) %% 2 # binary representation
        famil0 = famil * (matrix(1,nfamil,1) %*% cc)
        nhyps0 = rowSums(famil0)
        sub = (nhyps0 > 0)
        nfamil1 = sum(sub)
        famil1 = famil0[sub,]
        nhyps1 = nhyps0[sub]
        hyps1 = ((which(t(famil1)>0) - 1) %% m) + 1 # active
        hypotheses
        k = length(hyps1)
        famil2 = famil1 * (1:nfamil1)
```

```
        fam = t(famil2)[t(famil2) > 0] # family membership
        for (f in 1:nfamil1) { # for cusum of weights within each
        family
            h = nhyps1[f]
            col = matrix(1,h,1) %*% (1:h)
            u = (t(col) >= col) + 0
            if (f==1) {cum = u} else {
                cum = rbind(cbind(cum, matrix(0, nrow(cum), h)),
                            cbind(matrix(0, h, ncol(cum)), u))
            }
        }
        w = wgtmat[i, hyps1]
        p = pvalues[, hyps1]
        a = cbind(rep(sim,each=k), rep(fam,r), as.vector(t(p)),
        rep(w,r))
        a = matrix(a[order(a[,1], a[,2], a[,3]),], ncol=4)
        cw = as.vector(cum%*%matrix(a[, 4], nrow=k))
        pos = (cw > 0)
        qpos = a[pos, 3] / cw[pos]
        q = rep(max(qpos) + 1, r*k)
        q[pos] = qpos # no contribution from hypotheses with zero
        weights
        x = apply(matrix(q, nrow=k, ncol=r), 2, min)
        pinter[, i] = x
        incid[i, ] = cc
    }
    padj = matrix(0, r, m)
    for (j in 1:m) {
        ind = matrix(rep(incid[,j], each=r), nrow=r)
        padj[,j] = apply(pinter*ind, 1, max)
     }
    padj[padj > 1] = 1
    padj
}
```

Appendix B: R function for calculating the weight matrix for the intersection hypotheses

```
fwgtmat <- function(w, g) {
    m = length(w)
    iset = 1:m
    ntests = 2^m - 1
    col = matrix(1,m,1) %*% (1:m)
    row = t(col)
    wgtmat = matrix(0, ntests, m)
    gtrmat = matrix(0, ntests, m*m)
    for (i in 1:ntests) {
        number = ntests - i + 1
        cc = floor(number/2^(m - (1:m))) %% 2 # binary representation
        jset = which(cc > 0)
        jcset = setdiff(iset, jset)
        if (length(jcset) >= 1) {
            j = jcset[1]
            jpset = union(jset, j) # a set with 1 more element than jset
            ip = 2^m - sum(2^(m - jpset)) # index of the superset
            w = wgtmat[ip, ]
```

```
            g = matrix(gtrmat[ip, ], m, m, byrow=T)
            w[jset] = w[jset] + w[j]*g[j,jset]
            w[jcset] = 0
            mat = as.matrix(g[,j]) %*% t(as.matrix(g[j,]))
            dia = diag(mat) %*% matrix(1, 1, m)
            cc = t(as.matrix(cc)) # check conditions
            con = ((t(cc) %*% cc) * (row != col) * (dia < 1) > 0)
            g[con] = (g[con] + mat[con])/(1 - dia[con])
            g[!con] = 0
        }
        wgtmat[i,] = w
        gtrmat[i,] = as.vector(t(g))
    }
    wgtmat
}
```

Appendix C: Covariance matrix for estimated log rate ratios for negative binomial model

The mean structure for the negative binomial model can be written as

$$\log(\mu_i) = \log(t_i) + \alpha + x_{i1}\beta_1 + \cdots + x_{im}\beta_m,\ i = 1, \ldots, n$$

where t_i denotes the exposure for subject i (hence $\log(t_i)$ is the offset term) and $x_{ij} = 1$ if subject i is on treatment j and 0 otherwise. The covariance matrix for the maximum likelihood estimator $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_m\right)'$ is given by the inverse of the Fisher information matrix, $V_{\hat{\theta}} = I\left(\hat{\theta}\right)^{-1}$, where

$$I(\theta) = \sum_{i=1}^{n} w_i x_i x_i'$$

Here, $x_i = (1, x_{i1}, \ldots, x_{im})'$ is the ith row of the design matrix, and $w_i = \mu_i/(1 + \kappa\mu_i)$ is the weight for subject i. Because $x_{ij}x_{ik} = I(j = k)$, we have

$$I(\theta) = \begin{pmatrix} \sum w_i & \sum w_i x_{i1} & \sum w_i x_{i2} & \cdots & \sum w_i x_{im} \\ \sum w_i x_{i1} & \sum w_i x_{i1} & 0 & \cdots & 0 \\ \sum w_i x_{i2} & 0 & \sum w_i x_{i2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum w_i x_{im} & 0 & 0 & \cdots & \sum w_i x_{im} \end{pmatrix}$$

Using the formula for inverting partitioned matrices, we can show that

$$V_{\hat{\beta}} = \operatorname{diag}\left(\left\{\left(\sum_{i=1}^{n} \hat{w}_i x_{ij}\right)^{-1}\right\}\right) + \left(\sum_{i=1}^{m} \hat{w}_i x_{i0}\right)^{-1} J_m$$

Therefore, the off-diagonal elements of $V_{\hat{\beta}}$ are positive.

Appendix D: Covariance matrix for estimated log hazard ratios for Cox proportional hazards model

Let $t_1 < \cdots < t_K$ denote the uncensored failure times. Let n_{jk} denote the number of subjects in treatment j at risk at time t_k, where $j = 0, 1, \ldots, m$, $k = 1, \ldots, K$. Let the covariate vector consist of dummy variables for the treatment indicators for treatment 1 to treatment m, respectively. Using the results from section

4.2.2 of [13], the asymptotic covariance matrix for the estimated log hazard ratios is equal to the inverse of the observed information matrix, $V_{\hat{\beta}} = I\left(\hat{\beta}\right)^{-1}$, where

$$I(\beta) = \sum_{k=1}^{K} A(\beta, t_k)$$

$A(\beta, t_k)$ is an $m \times m$ matrix with the (r, s)th element equal to

$$a_{rs}(\beta, t_k) = p_r(\beta, t_k)(I(r = s) - p_s(\beta, t_k))$$

and

$$p_r(\beta, t_k) = \frac{n_{rk} \exp(\beta_r)}{\sum_{j=0}^{m} n_{jk} \exp(\beta_j)}$$

for $r, s = 1, \ldots, m$, $k = 1, \ldots, K$.

Let $A_k = A\left(\hat{\beta}, t_k\right)$ and $p_{jk} = p_j\left(\hat{\beta}, t_k\right)$. Furthermore, let $p_k = (p_{1k}, \ldots, p_{mk})'$ denote the vector of estimated probabilities for the experimental treatments at time t_k. Then, $A_k = \text{diag}(p_k) - p_k p_k'$, and $V_{\hat{\beta}} = \left(\sum_{k=1}^{K} A_k\right)^{-1}$. Define

$$B_{k+1} = \text{diag}(p_1 + \cdots + p_K) - \sum_{\ell=1}^{k} p_\ell p_\ell'$$

for $k = 1, \ldots, K$. Let $B_1 = \text{diag}(p_1 + \cdots + p_K)$. Because $B_{k+1} = B_k - p_k p_k'$, using the results from [14], we have

$$B_{k+1}^{-1} = B_k^{-1} + v_k B_k^{-1} p_k p_k' B_k^{-1} \tag{D.1}$$

where $v_k = \left(1 - p_k' B_k^{-1} p_k\right)^{-1}$, $k = 1, \ldots, K$. Because

$$\begin{aligned} B_k - \text{diag}(p_k) &= \text{diag}\left(\sum_{\ell \neq k} p_\ell\right) - \sum_{\ell=1}^{k-1} p_\ell p_\ell' \\ &= \sum_{\ell=1}^{k-1} \left\{\text{diag}(p_\ell) - p_\ell p_\ell'\right\} + \text{diag}\left(\sum_{\ell=k+1}^{K} p_\ell\right) \end{aligned}$$

and $\text{diag}(p_\ell) - p_\ell p_\ell'$ is positive definite for each ℓ, we have $B_k > \text{diag}(p_k)$. It follows that $p_k' B_k^{-1} p_k < p_k' \{\text{diag}(p_k)\}^{-1} p_k = \sum_{j=1}^{m} p_{jk}$, $1 - p_k' B_k^{-1} p_k > p_{0k}$, and $v_k > 0$ for $k = 1, \ldots, K$.

Suppose B_k^{-1} has positive diagonal elements and nonnegative off-diagonal elements, then, by (D.1) and the positivity of v_k, B_{k+1}^{-1} also has positive diagonal elements and nonnegative off-diagonal elements. Because $B_1^{-1} = \text{diag}((p_{j1} + \cdots + p_{jK})^{-1})$ has positive diagonal elements and nonnegative off-diagonal elements, by deduction, B_{K+1}^{-1} has positive diagonal elements and nonnegative off-diagonal elements. Finally, because $B_{K+1}^{-1} = V_{\hat{\beta}}$, $V_{\hat{\beta}}$ has only nonnegative off-diagonal elements.

Acknowledgements

The author thanks two reviewers for their very helpful and constructive comments that led to significant improvement of the paper.

References

1. Dmitrienko A, Tamhane AC, Bretz F. *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/CRC: Boca Raton, FL, 2010.
2. Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. Chapman & Hall/CRC: Boca Raton, FL, 2010.
3. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:586–604.
4. Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parameteric tests. *Biometrical Journal* 2011; **53**:894–913.

5. Sarkar SK, Chang CK. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**:1601–1608.
6. Sarkar SK. Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *Annals of Statistics* 1998; **26**:494–504.
7. Sarkar SK. On the Simes inequality and its generalization. Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen. *IMS Collections* 2008; **1**:231–242.
8. Wang B, Ting N. An application of graphical approach to construct multiple testing procedures in a hypothetical phase III design. *Frontiers in Public Health* 2014; **1**:Article 75.
9. Hung HMJ, Wang SJ. Challenges to multiple testing in clinical trials. *Biometrical Journal* 2010; **52**:747–756.
10. Lu K. Sample size calculations with multiplicity adjustment for longitudinal clinical trials with missing data. *Statistics in Medicine* 2012; **31**:19–28.
11. Bretz F, Maurer W, Hommel G. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* 2011; **30**:1489–1501.
12. Dmitrienko A, Millen BA, Brechenmacher T, Paux G. Development of gatekeeping strategies in confirmatory clinical trials. *Biometrical Journal* 2011; **53**:875–893.
13. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Date* (2nd edn). Wiley: Hoboken, NJ, 2002.
14. Miller KS. On the inverse of the sum of matrices. *Mathematics Magazine* 1981; **54**:67–72.