Chain procedures: A class of flexible closed testing procedures with clinical trial applications

Brian A. Millen (Eli Lilly and Company)
Alex Dmitrienko (Eli Lilly and Company)

December 9, 2010

Abstract

We define a class of multiple testing procedures for testing a family of hypotheses based on a pre-specified or data-driven testing sequence. These procedures, termed chain procedures, are characterized by independent sets of parameters which govern the initial allocation of the overall α level among the null hypotheses of interest and the process for iteratively reallocating available (or unspent) α among the remaining eligible null hypotheses. As a result, chain procedures are more flexible than popular stepwise procedures such the Holm or fallback procedures. While presenting the broad class of chain procedures, this paper focuses on the development of parametric chain procedures for problems with a known joint distribution of the hypothesis test statistics. Chain procedures are closed testing procedures and thus control the familywise error rate in the strong sense. Further, we discuss optimal selection of parameters of chain procedures based on clinically relevant application-specific criteria. Finally, we illustrate application of the chain testing method using a clinical trial example aimed at the development of a tailored therapy.

1 Inroduction

With clinical trial designs and their associated null hypotheses becoming increasingly complex, multiple testing procedures which provide strong control of the familywise error rate (FWER) are necessary to ensure valid interpretation of trial results. A large number of such testing procedures are available. Reviews of such methods may be found in Hochberg and Tamhane (1987), Hsu (1996) and Dmitrienko et al. (2009). Common testing methods include stepwise testing procedures that rely on a data-driven ordering of the null hypotheses of interest, e.g., Holm procedure (Holm,

1979) and Hochberg procedure (Hochberg, 1988), and procedures that test the null hypotheses in a pre-specified order, e.g., the fixed-sequence procedure (Westfall and Krishen, 2001) and fallback procedure (Wiens, 2003; Wiens and Dmitrienko, 2005). Despite the large number of available procedures, methods that are powerful and flexible enough to be readily customized for particular applications are needed.

In this paper we present a new class of multiple testing procedures known as chain procedures which serve as an extension of traditional stepwise procedures. Unlike other stepwise procedures, chain procedures employ two independent sets of parameters for the two key characteristics of any stepwise procedure. The first one is the α allocation rule which specifies the initial weights of the null hypotheses of interest to define their relative importance. The second characteristic is the α propagation rule which determines the process of allocating α levels to the remaining eligible null hypotheses after the procedure rejects a null hypothesis. As will become evident, it is this feature that makes chain procedures unique and makes them broadly flexible for an array of applications. The term *chain procedure* is used to highlight the similarity between the testing algorithm and a chain (discrete random process), e.g., a Markov chain. In this analogy, the individual null hypotheses tested by a chain procedure may be thought of as states and the hypothesis testing process as the evolution of the system.

The paper is organized as follows. Section 2 presents and reviews simple Bonferroni-based chain procedures. This discussion is provided for completeness and is primarily used to motivate the parametric chain procedures introduced in Sections 3 and 4. Section 5 discusses optimality criteria based on various definitions of power in multiple testing problems and defines optimization algorithms that are relevant to the application of *all* chain procedures. Sections 6 presents an example application of parametric chain procedures using a clinical trial aimed at the development of a targeted therapy. The paper closes with a discussion in Section 7, which provides unifying context for the application of chain procedures in clinical trials.

2 Bonferroni-based chain testing procedures

This section provides motivation for parametric chain procedures defined in Sections 3 and 4 and begins with a discussion of p-value-based chain procedures derived from the Bonferroni procedure.

To fix ideas, consider a clinical trial with three null hypotheses of no treatment effect denoted by H_i, $i = 1, 2, 3$. Assume that the testing sequence is pre-specified, i.e., the null hypothesis H_1 is tested first, followed by H_2, and H_3 is tested after H_2. The raw p-values for the null hypotheses are denoted by p_i, $i = 1, 2, 3$. Examples of multiple testing procedures with a pre-specified testing sequence include the fixed-

sequence and fallback procedures.

In fallback testing, an α allocation rule apportions the overall α level, e.g., $\alpha = 0.05$, among the hypotheses in the family. The initial weights of the null hypotheses are denoted by w_i, $i = 1, 2, 3$ ($w_1 + w_2 + w_3 = 1$ and $w_i \geq 0$, $i = 1, 2, 3$). In addition, there is an α propagation rule which states that, when a null hypothesis is rejected, the fraction of α used for its local test is carried over to the next null hypothesis in the sequence. For example, the fallback procedure tests H_1 at αw_1 and, if H_1 is rejected, the significance level for H_2 is given by $\alpha(w_2 + w_1)$. Note that the α propagation rule is unnecessarily restrictive since the significance level for H_3 remains unchanged after H_1 is rejected.

Chain procedures in problems with a pre-specified testing sequence are based on sequentially rejective algorithms similar to that used in the fallback procedure but offer considerably more flexibility in terms of the choice of α propagation rules. As was stated in the Introduction, the key idea behind chain procedures is to separate the α propagation rule from the α allocation rule by introducing another set of parameters that determine how the weights of remaining null hypotheses are updated at each step of the testing algorithm. A natural extension of the fallback procedure described above is a chain procedure with the α allocation and α propagation rules defined in Figure 1. The α allocation rule is formulated in terms of the hypothesis weights w_1, w_2 and w_3. Further, the α propagation rule is based on the so-called *transition parameters* g_{12}, g_{13} and g_{23} ($g_{12}+g_{13} = 1$, $g_{12} \geq 0$, $g_{13} \geq 0$ and $g_{23} = 1$). Specifically, if H_1 is rejected, the significance levels for testing H_2 and H_3 are set to $\alpha(w_2+w_1g_{12})$ and $\alpha(w_3 + w_1g_{13})$, respectively, where g_{12} identifies the proportion of α carried forward from H_1 to H_2 and g_{13} identifies the proportion of α carried forward from H_1 to H_3. Unlike the fallback procedure, this chain procedure gives researchers more control over individual branches of the decision tree and enables them to build a multiple testing procedure customized to the objectives of this particular clinical trial. Suppose, for example, that it is desirable to test both H_2 and H_3 whenever H_1 is rejected. The fallback procedure would require a positive weight, w_3, assigned to H_3 in order to satisfy this requirement. As a result the α available for allocation to H_1 and H_2 will be reduced, which will lead to power loss for these higher priority tests. However, a chain procedure can be constructed so as to allow testing of H_3 whenever H_1 is rejected without sacrificing any apportionment of α from H_1. This is accomplished via appropriate selection of the transition parameters g_{12} and g_{13}.

In a general problem of testing null hypotheses $H_1, \ldots, H_m$, the α propagation rule for chain procedures with a pre-specified testing sequence is specified by an $m \times m$ transition matrix whose elements are denoted by g_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, m$. Here g_{ij} is the fraction of the unspent α allocated to H_j when H_i is rejected. The general

restrictions on the transition parameters are given by

$$\begin{aligned} &g_{ij} \geq 0,\ i = 1, \ldots, m-1,\ j = i+1, \ldots, m, \\ &g_{ij} = 0,\ i = 1, \ldots, m,\ j = 1, \ldots, i, \\ &\sum_{j=1}^{m} g_{ij} = 1,\ i = 1, \ldots, m. \end{aligned}$$

Note that the transition parameters below the main diagonal are set to 0 and thus the fraction of α available after rejecting H_i, $i = 1, \ldots, m-1$, is distributed among the null hypotheses placed later in the sequence, i.e., H_j, $j = i+1, \ldots, m$. Note that the serial chain procedure simplifies to the fallback procedure if $g_{ij} = 1$ ($j = i+1$, $i = 1, \ldots, m-2$) and $g_{ij} = 0$ ($j > i+1$, $i = 1, \ldots, m-2$). Further, the serial chain procedure simplifies to the fixed-sequence procedure if, in addition to the conditions given in the previous sentence, $w_1 = 1$ and $w_i = 0$ ($i = 2, \ldots, m$).

To emphasize the importance of the pre-specified hypothesis ordering in the implementation of the testing method, chain procedures with pre-specified testing sequences are termed *serial chain procedures*.

Now consider a clinical trial with three null hypotheses and assume that there is no natural ordering among them. An example of multiple testing procedures that can be used in this problem is the Holm procedure. Let p_i^*, $i = 1, 2, 3$, denote the weighted p-values, i.e., $p_i^* = p_i/w_i$, $i = 1, 2, 3$. The Holm procedure starts with the null hypothesis corresponding to the most significant weighted p-value. Suppose that H_1 is tested first and it is rejected. After this, the weights of the remaining null hypotheses are re-computed to account for the rejection of H_1. The updated weights for H_2 and H_3 are given by

$$\frac{w_2}{w_2 + w_3} \text{ and } \frac{w_3}{w_2 + w_3},$$

respectively. In other words, the weight assigned to H_1 is distributed among the remaining null hypotheses and thus the Holm procedure implicitly specifies an α propagation rule. However, the α propagation rule is fixed. Researchers can choose from an infinite number of α allocation rules but the rule for updating the hypothesis weights within the multiple testing procedure follows directly from this choice.

As in the first example, it is easy to construct a more flexible chain procedure by introducing an independent set of parameters to define the α propagation rule, e.g., consider the chain procedure with the α allocation and α propagation rules displayed in Figure 2 (procedure parameters are defined as in Figure 1). The testing algorithm for the chain procedure begins with the most significant weighted p-value based on the initial hypothesis weights w_1, w_2 and w_3. If the corresponding null hypothesis is rejected, the fraction of α assigned to this null hypothesis is transferred to the other

two null hypotheses. For example, if H_1 is rejected, the weights for H_2 and H_3 are set to $w_2 + w_1 g_{12}$ and $w_3 + w_1 g_{13}$, respectively. Thus researchers can easily modify the conditional behavior of the chain procedure after each rejection by changing the values of the transition parameters.

An important difference between chain procedures in problems with pre-specified and data-driven testing sequences is that in the context of ordered null hypotheses the unspent α available after the rejection of each null hypothesis is distributed only among the null hypotheses placed later in the sequence. When the null hypotheses are not naturally ordered, the restrictions on the transition parameters are relaxed to account for the different possible orderings of the realized test procedure:

$$\begin{aligned} &g_{ij} \geq 0,\ i = 1, \ldots, m,\ j = 1, \ldots, m, \\ &g_{ii} = 0,\ i = 1, \ldots, m, \\ &\sum_{j=1}^{m} g_{ij} = 1,\ i = 1, \ldots, m. \end{aligned}$$

Because this testing strategy can be viewed as one which is able to cycle through null hypotheses for any given pre-specified sequence, chain procedures based on data-driven testing sequences are referred to as *cyclical chain procedures*.

Bonferroni-based chain procedures for testing a family with an arbitrary number of null hypotheses were developed by Bretz et al. (2009) and Burman et al. (2009). The general principles outlined in this section can be applied to construct a variety of multiple testing procedures, including *non*-Bonferroni-based p-value-based chain procedures (e.g., Simes-based chain procedures, Simes, 1986), parametric chain procedures that take into account the joint distribution of the hypothesis test statistics and chain procedures derived from a combination of p-value-based and parametric procedures. In this paper we focus on the development of parametric chain procedures. We give a detailed treatment of serial parametric chain procedures in Section 3 and cyclical parametric chain procedures in Section 4.

3 Serial parametric chain procedures

We now consider the problem of testing a family of ordered null hypotheses with a known joint distribution of the hypothesis test statistics, e.g., the test statistics may follow a multivariate normal or t distribution. In this case, chain procedures that incorporate the correlations among the test statistics into the decision rules (i.e., parametric chain procedures) can be defined. These procedures will be more powerful than Bonferroni-based chain procedures introduced in Section 2.

The process of constructing serial parametric chain procedures will be illustrated by considering the problem of testing three ordered null hypotheses H_1, H_2 and H_3

considered in Section 2. Let p_i, $i = 1, 2, 3$, and t_i, $i = 1, 2, 3$, denote the p-values and test statistics associated with the null hypotheses, respectively. The hypothesis weights are denoted by w_i, $i = 1, 2, 3$, and the transition matrix is given by $g = \{g_{ij},\ i = 1, 2, 3,\ j = 1, 2, 3\}$ with $g_{21} = g_{31} = g_{32} = 0$. The resulting α allocation and α propagation rules are depicted in Figure 1.

To better understand the testing algorithm used in serial parametric chain procedures, it is helpful to begin with the formulation of the Bonferroni-based chain procedure for this hypothesis testing problem. As shown in Table 1, this procedure can be formulated as a closed testing procedure. The first column in Table 1 lists all seven (i.e., $2^3 - 1$) non-empty intersection hypotheses in the closed family associated with H_1, H_2 and H_3. The second column defines the local test (rejection rule) for each intersection hypothesis. The local tests for the seven intersection hypotheses are chosen in such a way that the decision rule for H_1 does not depend on the outcome of the tests for H_2 and H_3. Similarly, the decision rule for H_2 is unaffected by the test for H_3. A closed testing procedure rejects a null hypothesis if and only if it rejects all intersection hypotheses which include that null hypothesis. Note that, due to the Bonferroni inequality, the size of each local test does not exceed α and thus the chain procedure controls the FWER at the α level.

It is straightforward to see that the Bonferroni-based chain procedure is governed by the following stepwise algorithm:

Step 1. Test H_1 at level $\alpha_1 = w_1\alpha$.

Step 2. Test H_2 at level $\alpha_2 = w_2\alpha + r_1 g_{12}\alpha_1$, where $r_1 = 1$ if H_1 is rejected and 0 otherwise.

Step 3. Test H_3 at level $\alpha_3 = w_3\alpha + r_1 g_{13}\alpha_1 + r_2 g_{23}\alpha_2$, where $r_2 = 1$ if H_2 is rejected and 0 otherwise.

It is clear from examination of the decision rules in this closed testing representation that the local tests are generally conservative, since they ignore the correlations among the p-values or, equivalently, the hypothesis test statistics associated with each intersection hypothesis. The parametric chain procedure for this hypothesis testing problem is constructed by defining local tests based on the same hypothesis weights and transition parameters as for the Bonferroni-based case but explicitly accounting for the joint distribution of the test statistics. Consider, for example, the intersection hypothesis $H_1 \cap H_2 \cap H_3$ in Table 1. In the p-value-based setting, this intersection hypothesis is rejected if

$$p_1 \leq \alpha w_1 \text{ or } p_2 \leq \alpha w_2 \text{ or } p_3 \leq \alpha w_3.$$

The local test in the parametric setting will be given by

$$w_1 t_1 \geq c \text{ or } w_2 t_2 \geq c \text{ or } w_3 t_3 \geq c,$$

where c is a suitably defined critical value.

The parametric local tests for all individual intersection hypotheses in the closed family are displayed in the second column of Table 2. The parametric chain procedure is defined as a closed testing procedure based on the parametric local tests, i.e., it rejects a null hypothesis if and only if all intersection hypotheses including this particular null hypothesis are rejected. The critical values shown in Table 2 are selected to ensure that the size of each local test equals α and thus the closure principle (Marcus, Peritz and Gabriel, 1976) guarantees that the parametric chain procedure strongly controls the FWER at the α level.

The critical values are computed from the null distribution of t_1, t_2 and t_3. Let T_1, T_2 and T_3 denote test statistics whose joint distribution is equal to that of t_1, t_2 and t_3 under the global null hypothesis. For any vector of hypothesis weights and transition matrix, the five critical values in Table 2 are computed sequentially from the following system of equations (e.g., c_1 is found first followed by c_2, etc):

$$\begin{aligned}
&P(w_1 T_1 \geq c_1 \text{ or } w_2 T_2 \geq c_1 \text{ or } w_3 T_3 \geq c_1) = \alpha,\\
&P(w_1 T_1 \geq c_1 \text{ or } w_2 T_2 \geq c_2) = \alpha,\\
&P(w_1 T_1 \geq c_1 \text{ or } (w_2 g_{23} + w_3) T_3 \geq c_3) = \alpha,\\
&P((w_1 g_{12} + w_2) T_2 \geq c_4 \text{ or } (w_1 g_{13} + w_3) T_3 \geq c_4) = \alpha,\\
&P(T_i \geq c_5) = \alpha,\ i = 1, 2, 3.
\end{aligned}$$

If the test statistics follow a normal or t distribution, the critical values can be derived using the algorithm for computing multivariate normal or t probabilities developed by Genz and Bretz (2009).

While reference to the Bonferroni-based chain procedure is instructive for formulating the parametric chain procedure, it is worth noting that, in problems with unequally weighted null hypotheses, hypothesis weights are interpreted differently by p-values-based and parametric procedures. In a general parametric setting, hypothesis weights are defined in the test statistic space rather than the p-value space, which alters their interpretation due to a non-linear relationship between p-values and test statistics. This is also true for parametric chain procedures. In the p-value-based setting, the transition matrix defines fractions of α carried over from a given hypothesis to the other null hypotheses. Within the parametric framework, the transition matrix still defines the α propagation rules but its elements do not have the direct connotation of a fraction of α allocated to the other null hypotheses after the current null hypothesis is rejected.

A careful review of the rejection rules for the individual intersection hypotheses in Table 2 reveals that the parametric chain procedure is based on the following stepwise testing algorithm.

Step 1. Reject H_1 if $t_1 \geq c_1/w_1$.

Step 2. Reject H_2 if

- $t_2 \geq c_4/(w_1 g_{12} + w_2)$ (H_1 is rejected),
- $t_2 \geq \max(c_1/w_2, c_2/w_2, c_4/(w_1 g_{12} + w_2))$ (H_1 is not rejected).

Step 3. Reject H_3 if

- $t_3 \geq c_5$ (H_1 and H_2 are rejected),
- $t_3 \geq \max(c_4/(w_1 g_{13} + w_3), c_5)$ (H_1 is rejected and H_2 is not rejected),
- $t_3 \geq \max(c_3/(w_2 g_{23} + w_3), c_5)$ (H_1 is not rejected and H_2 is rejected),
- $t_3 \geq \max(c_1/w_3, c_3/(w_2 g_{23} + w_3), c_4/(w_1 g_{13} + w_3), c_5)$ (H_1 and H_2 are not rejected).

While at first glance the algorithm may appear rather complex, it is conceptually similar to the simple Bonferroni-based algorithm. In general, the more null hypotheses are rejected early in the sequence, the easier it is to reject null hypotheses later in the sequence. A smaller critical value is used for H_2 if H_1 is rejected compared to the case when H_1 cannot be rejected. Likewise, it is easier to reject H_3 if at least one of the preceding tests is significant compared to the case where both H_1 and H_2 are not rejected.

To illustrate this stepwise algorithm, assume that $w_1 = w_2 = w_3 = 1/3$ and the transition matrix is given by

$$\begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \end{bmatrix}.$$

Assume again that the three test statistics follow a trivariate normal distribution with a common correlation coefficient $\rho = 0.5$ Using a one-sided $\alpha = 0.025$, the critical values $c_1, \ldots, c_5$ defined in Table 2 are given by 0.78, 0.70, 1.41, 1.11 and 1.96, respectively. The critical value for the first null hypothesis in the sequence, i.e., H_1, is given by $c_1/(1/3) = 2.35$ and thus the cut-off for p_1 is $1 - \Phi(2.35) = 0.0094$. If H_1 is rejected by the the parametric chain procedure, the test for the null hypothesis H_2 is based on the critical value $c_4/(w_1 g_{12} + w_2) = 2.21$ or p-value cut-off $1 - \Phi(2.21) = 0.0135$. Finally, assuming the parametric chain procedure fails to reject H_2, the critical value and p-value cut-off for the last null hypothesis in the sequence,

i.e., H_3, are given by $\max(c_4/(w_1 g_{13} + w_3), c_5) = 2.21$ and $1 - \Phi(2.21) = 0.0135$, respectively.

It is instructive to compare the parametric chain procedure with a fixed testing sequence to a parametric version of fallback procedure proposed by Huque and Alosh (2008) (this procedure will be referred to as the PF procedure). Interestingly, the parametric chain procedure does not simplify to the PF procedure when the fallback transition matrix defined in Section 2 is used, i.e., when $g_{ij} = 1$ ($j = i + 1$, $i = 1, \ldots, m - 2$) and $g_{ij} = 0$ ($j > i + 1$, $i = 1, \ldots, m - 2$). This is due to the fact that the PF procedure relies on a Bonferroni-type α-splitting method to define critical values for the test statistics. By contrast, the parametric chain procedure with the fallback transition matrix, which will be referred to as the modified parametric fallback procedure or MPF procedure, utilizes critical values of multivariate normal or t distributions and thus serves as an extension of the Dunnett procedure. As an example, consider a Phase II clinical trial with a balanced analysis of variance layout

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 0, \ldots, m, \ j = 1, \ldots, n,$$

where y_{ij} denotes the response of the jth patient in the ith treatment group with $i = 0$ denoting the placebo group. Assume that ε_{ij}, $i = 0, \ldots, m$, $j = 1, \ldots, n$, are normally distributed with mean 0 and common standard deviation σ. Further, consider

$$H_i : \ \mu_i - \mu_0 = 0, \quad i = 1, \ldots, m,$$

and suppose that the hypotheses $H_1, \ldots, H_m$ are ordered and equally weighted. In this case, the PF procedure uses the Bonferroni-based p-value cutoff for H_1, i.e., α/m and the MPF procedure uses a less stringent Dunnett-based p-value cutoff. As an illustration, the p-value cutoffs are 0.0083 (PF procedure) and 0.0091 (MPF procedure) for a one-sided $\alpha = 0.025$ when $m = 3$ and $n = 100$. This means that the MPF procedure improves the power for null hypotheses in the beginning of the sequence. It can also be shown that, unlike the PF procedure, the MPF procedure is uniformly more powerful than the Dunnett procedure in the sense that the MPF procedure rejects all null hypotheses rejected by the Dunnett procedure and potentially more null hypotheses. Note that the Dunnett procedure serves as an example of single-step parametric procedures and thus this property parallels a well-known fact that the regular fallback procedure is uniformly more powerful than the widely used single-step p-value-based procedure (Bonferroni procedure).

The chain-based MPF procedure extends another important property of the fallback procedure. Wiens and Dmitrienko (2005) showed that the fallback procedure can be more powerful than the step-down Holm procedure (Holm, 1979). Likewise, it is easy to demonstrate that the MPF procedure can be more powerful than the step-down Dunnett procedure (Naik, 1975; Marcus, Peritz and Gabriel, 1976). Considering

the multiple testing problem defined above, assume that $m = 3$ and $t_1 > t_2 > t_3$. The critical values of the step-down Dunnett procedure, denoted by d_1, d_2 and d_3, are computed from

$$P(\max(T_1, T_2, T_3) \geq d_1) = \alpha, \quad P(\max(T_2, T_3) \geq d_2) = \alpha, \quad P(T_3 \geq d_3) = \alpha,$$

where T_1, T_2 and T_3 follow a central trivariate t distribution with $\nu = 3(n-1)$ d.f. and common correlation coefficient $\rho = 1/2$. The corresponding critical values of the MPF procedure, denoted by c_1, c_2 and c_3, are obtained as follows. First, $c_1 = d_1$ and $c_3 = d_3$. Further, c_2 is found from

$$P(T_1 \geq c_2 \text{ or } T_2 \geq 2c_2) = \alpha$$

and thus $c_2 \leq d_2$. For example, $c_2 = 1.97$ and $d_2 = 2.21$ with a one-sided $\alpha = 0.025$, $m = 3$ and $n = 100$. This implies that the MPF procedure will reject as many, and potentially more, null hypotheses than the step-down Dunnett procedure in this multiple testing problem.

The algorithm for constructing serial parametric chain procedures in problems with three null hypotheses is easy to extend to the problem of testing an arbitrary number of null hypotheses. Consider a general problem of testing the null hypotheses $H_1, \ldots, H_m$. The hypothesis weights are denoted by $w_1, \ldots, w_m$ and the transition matrix is given by $g = \{g_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, m\}$. Using arguments similar to those employed in the Bonferroni case, it is shown in the Appendix how to set up serial parametric chain procedures based on the closure principle. The closed testing representation guarantees that these procedures strongly control the FWER in the general case.

Adjusted p-values for serial parametric chain procedures can be computed using the direct-calculation algorithm proposed in Dmitrienko, Tamhane and Wiens (2008). This algorithm is based on the general definition of adjusted p-values given in Westfall and Young (1993), i.e., the adjusted p-value for H_i, $i = 1, \ldots, m$, is the lowest significance level at which this null hypothesis is rejected by the procedure. Now select a large value of K, e.g., $K = 100,000$, and let $\alpha = k/K$, $0 < k < K$. Using a simple grid search, find the smallest α (corresponding to the smallest k) for which H_i is rejected by the procedure. This value of α approximates the adjusted p-value for H_i.

A serial parametric chain procedure can be constructed in any hypothesis testing problem whenever the joint null distribution of the hypothesis test statistics is known. The parametric chain procedure is uniformly more powerful than the Bonferroni-based chain procedure and should be considered when the situation allows. In the next section we introduce parametric chain procedures for problems wherein there is no assumed ordering of the null hypotheses.

4 Cyclical parametric chain procedures

In this section we consider a parametric formulation of chain procedures when the testing sequence is not pre-specified but will be determined based on the data.

To define cyclical parametric chain procedures, we will again consider the problem of testing three null hypotheses H_1, H_2 and H_3 with the general α allocation and α propagation rules in this problem displayed in Figure 2. Note that in problems with a data-driven testing sequence, the transition parameters below the main diagonal may be positive, e.g., fractions of the significance level used for testing H_3 can be carried forward to H_1 and H_2 after H_3 is rejected.

Table 3 defines the parametric chain procedure in this problem as a closed testing procedure. For any vector of hypothesis weights and transition matrix, the critical values are found from the following system of equations:

$$
\begin{aligned}
&P(w_1T_1 \geq c_{123} \text{ or } w_2T_2 \geq c_{123} \text{ or } w_3T_3 \geq c_{123}) = \alpha,\\
&P((w_1 + w_3g_{31})T_1 \geq c_{12} \text{ or } (w_2 + w_3g_{32})T_2 \geq c_{12}) = \alpha,\\
&P(w_1 + w_2g_{21})T_1 \geq c_{13} \text{ or } (w_3 + w_2g_{23})T_3 \geq c_{13}) = \alpha,\\
&P((w_2 + w_1g_{12})T_2 \geq c_{23} \text{ or } (w_3 + w_1g_{13})T_3 \geq c_{23}) = \alpha,\\
&P(T_i \geq c_i) = \alpha,\ i = 1, 2, 3,
\end{aligned}
$$

where, as in Section 3, T_1, T_2 and T_3 denote test statistics whose joint distribution is equal to that of t_1, t_2 and t_3 under the global null hypothesis.

It is instructive to review the stepwise algorithms for implementation of the Bonferroni-based and parametric chain procedures. This emphasizes the conceptual similarities of the algorithms and provides clarity for the general parametric chain algorithm for the case of an arbitrary number of null hypotheses.

It is shown in Bretz et al. (2009) that the Bonferroni-based chain procedure in this problem is based on the following stepwise algorithm:

Step 1. Define the weighted p-values

$$p^*_{1i} = \frac{p_i}{w_i},\ i = 1, 2, 3,$$

and let i_1 denote the index of the smallest weighted p-value. Reject this null hypothesis if $p_{i_1} \leq w_{i_1}\alpha$ and transfer the significance level used in this test, i.e., $w_{i_1}\alpha$, to the other null hypotheses according to the α propagation rule. For example, if $i_1 = 1$, and $H_{i_1} \equiv H_1$ is rejected in this step, $\alpha w_1 g_{12}$ will be carried over to H_2 and $\alpha w_1 g_{13}$ will be carried over to H_3. If H_{i_1} is rejected, go to Step 2. Otherwise, accept all null hypotheses.

Step 2. Let i_2 and i_3 denote the indices of the two remaining null hypotheses. Define the weighted p-values for these null hypotheses based on the updated weights. In other words, let

$$p^*_{2j} = \frac{p_j}{w_j + w_{i_1} g_{i_1,j}},\ j = i_1, i_2.$$

Note that the weight assigned to either null hypothesis is the sum of the original weight and the weight transferred from the null hypothesis rejected in Step 1. Let i_2 denote the index of the smallest weighted p-value among the remaining null hypotheses. Reject the corresponding null hypothesis H_{i_2}if $p^*_{2,i_2} \leq \alpha$, update the transition matrix and re-compute the weight of the last null hypothesis. The transition matrix is updated to account for the fact that there is only one null hypothesis left. If H_{i_2} is rejected, go to Step 3. Otherwise, accept H_{i_2} and H_{i_3}.

Step 3. Let i_3 denote the index of the remaining null hypothesis. The weight of the last null hypothesis is 1 and thus the rejection rule for the last null hypothesis takes a very simple form. Reject this null hypothesis if $p^*_{3,i_3} = p_{i_3} \leq \alpha$. Otherwise, accept it.

The parametric chain procedure is based on a straightforward stepwise algorithm very similar to the one used in the Bonferroni-based procedure, except the presentation is in the test statistic space rather than the p-value space and the critical values are found appropriately from the joint distribution of the test statistics:

Step 1. Define the weighted test statistics

$$t^*_{1i} = w_i t_i,\ i = 1, 2, 3,$$

and let i_1 denote the index of the largest weighted test statistic. Reject the corresponding null hypothesis if $t^*_{1,i_1} \geq c_{123}$ and transfer the fraction of α used in this test to the other null hypotheses by updating their weights based on the pre-specified α propagation rule. To illustrate, if H_1 is rejected, the weights for H_2 and H_3 are set to $w_2 + w_1 g_{12}$ and $w_3 + w_1 g_{13}$, respectively. If H_{i_1} is rejected, go to Step 2 and accept all null hypotheses otherwise.

Step 2. Let i_2 and i_3 denote the indices of the two remaining null hypotheses. Define the weighted test statistics using the updated hypothesis weights, i.e.,

$$t^*_{2j} = (w_j + w_{i_1} g_{i_1,j}) t_j,\ j = i_2, i_3.$$

Let i_2 denote the index of the largest weighted test statistic. Reject the corresponding null hypothesis if $t^*_{2,i_2} \geq c_{i_2,i_3}$ and transfer the fraction of α used in this test to the remaining null hypothesis. If H_{i_2} is rejected, go to Step 3 and accept H_{i_2} and H_{i_3} otherwise.

Step 3. Let i_3 denote the index of the last null hypothesis. Reject this null hypothesis if $t^*_{3,i_3} = t_{i_3} \geq c_{i_3}$ and accept it otherwise.

It is worth noting that only three of the critical values defined above need to be computed to carry out this stepwise algorithm. Also, the stepwise algorithm can be formulated in terms of p-values by converting the critical values defined above into p-value cut-offs.

To illustrate the stepwise algorithm defined above and also compare cyclical parametric chain procedures to Bonferroni-based ones, we will use the following example. Suppose that the three null hypotheses are equally weighted, i.e., $w_1 = w_2 = w_3 = 1/3$, the transition matrix is symmetric

$$\begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

and testing is performed at a one-sided $\alpha = 0.025$. Further, assume that, under the global null hypothesis, the three test statistics follow a standard trivariate normal distribution with a common correlation coefficient $\rho = 0.5$. The critical value of the parametric chain procedure in Step 1 is computed from the standard trivariate normal distribution, i.e., $c_{123} = 0.78$. Since the three null hypotheses are equally weighted, a common cut-off is used for p_1, p_2 and p_3, i.e., $1 - \Phi(c_{123}/(1/3)) = 0.0094$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. The Bonferroni-based chain procedure uses a more conservative cut-off given by $\alpha/3 = 0.0083$. Suppose that H_1 is rejected by the the parametric chain procedure in Step 1. The p-value cut-offs for testing H_2 and H_3 in Step 2 are obtained from the critical value $c_{23} = 1.11$. This critical value is computed from the bivariate distribution of T_2 and T_3. The cut-offs for p_2 and p_3 are computed as follows:

$$1 - \Phi\left(\frac{c_{23}}{w_2 + w_1 g_{12}}\right) = 1 - \Phi\left(\frac{c_{23}}{w_3 + w_1 g_{13}}\right) = 0.0135.$$

For comparison, the corresponding p-value cut-offs in the Bonferroni-based chain procedure are given by

$$\alpha(w_2 + w_1 g_{12}) = \alpha(w_3 + w_1 g_{13}) = 0.0125.$$

If the parametric chain procedure rejects a null hypothesis in Step 2, the remaining null hypothesis is tested at the full α level, i.e., at 0.025. The Bonferroni-based chain procedure uses the same cut-off for the last null hypothesis. The parametric chain procedure is more powerful than the Bonferroni-based chain procedure in the sense that the former rejects all hypotheses rejected by the latter and potentially more.

In the general case of m null hypotheses, cyclical parametric chain procedures are defined using the algorithm given in the Appendix. The closure principle is used again to demonstrate that general parametric chain procedures preserve the FWER in the strong sense. Adjusted p-values for cyclical parametric chain procedures are computed using the direct-calculation algorithm defined in Section 3.

It is instructive to note that serial parametric chain procedures may be viewed as a special case of the general cyclical framework presented in this section. The incorporation of assumed ordering of null hypotheses provides an important simplification to the formulation of chain procedures and their implementation. For example, in problems with a pre-specified testing sequence null hypotheses placed earlier in the sequence are tested independently of null hypotheses later in the sequence. As a result, the transition matrix does not need to be updated at each step of the testing algorithm. The decision to use a serial chain procedure or a more flexible cyclical chain procedure should be based on the particular application and the ability (or not) to assume ordering of the null hypotheses. As one might expect, cyclical chain procedures may be less powerful than serial chain procedures when the prior hypothesis ordering is appropriately accounted for.

5 Optimal selection of procedure parameters

A natural question, given the broad flexibility of chain procedures and the multitude of problems to which the procedures may be applied, is how to optimally select the values of the procedure parameters: hypothesis weights and transition matrix and, in the case of serial chain procedures, ordering of hypotheses. Such optimization, of course, depends on the metric used to evaluate the procedure. This metric will vary with the application. In this section, we discuss several possible applications, associated metrics, and their influence on the optimal parameters of chain procedures. We begin with more basic metrics such as disjunctive power and conjunctive power (Senn and Bretz, 2007) and then discuss extensions.

For simplicity, assume that the null hypotheses of interest, $H_1, \ldots, H_m$, are all false. Let $t_1, \ldots, t_m$ denote the associated test statistics and $r_1, \ldots, r_m$ denote binary rejection indicators, i.e., $r_i = 1$ when H_i is rejected and 0 otherwise. Further, let ψ_i denote the probability of rejecting H_i using a univariate α-level test under the alternative hypothesis, i.e., $\psi_i = E(r_i)$, $i = 1, \ldots, m$, where $E(\cdot)$ represents the expectation.

In some applications, success is achieved if at least one of the m null hypotheses is rejected. In such cases, the multiple null hypotheses are viewed as equally important and merely serve to provide the researcher with "multiple shots on goal" for success. An appropriate metric in this setting is simple *disjunctive* power, defined as the

probability of rejecting at least one hypothesis, i.e.,

$$P\left(\sum_{i=1}^{m} r_i \geq 1\right).$$

At the opposite extreme, there are applications wherein success is achieved only if all null hypotheses are rejected. Examples include clinical trials in patients with migraine, Alzheimer's disease, HIV, or osteoarthritis (see Offen et al., 2007, for more examples). In these cases, a successful treatment must demonstrate a significant effect on all co-primary endpoints. An appropriate metric in such settings is simple *conjuctive* power, defined as the probability of rejecting all null hypotheses in the family, i.e.,

$$P\left(\sum_{i=1}^{m} r_i = m\right).$$

It is easy to verify that the selection of optimal parameters for p-value-based chain procedures based on conjuctive power is trivial, resulting in a fixed-sequence procedure. For disjunctive power, on the other hand, the choice of weights and transition parameters may be non-trivial.

Between the two extremes of simple conjunctive and disjunctive power lie many appropriate metrics by which one may optimize a multiple testing strategy. When one is interested in rejecting some pre-specified number of null hypotheses, e.g., k null hypotheses ($1 < k < m$), it is reasonable to use *generalized disjunctive power* defined as

$$P\left(\sum_{i=1}^{m} r_i \geq k\right).$$

In other applications, the m null hypotheses may be split into multiple sub-families, and success is defined as rejection of one (or more) null hypotheses in each subfamily. As an example, consider a clinical trial wherein the researcher is interested in demonstrating efficacy according to one of several biomarkers and one of several outcome measures. An appropriate metric in this setting is *subset disjunctive power*, i.e.,

$$P\left(\sum_{i=1}^{j} r_i \geq 1 \text{ and } \sum_{i=j+1}^{m} r_i \geq 1\right).$$

The metrics (and applications) considered thus far assumed that all null hypotheses in the family (or subfamily) were of equal importance. These metrics are easily extended to incorporate weightings representing the relative importance of the null hypotheses. Let v_i represent the importance weight of hypothesis H_i with $0 \leq v_i \leq 1$, $i = 1, \ldots, m$, and $v_1 + \ldots + v_m = 1$. A weighted version of generalized disjunctive power is defined

as

$$P\left(\sum_{i=1}^{m} v_i r_i \geq \frac{k}{m}\right).$$

Weighted subset disjunctive power is defined in a similar fashion.

The final metric we introduce here is weighted proportional power (Benjamini and Hochberg, 1997; Westfall and Krishen, 2001), defined as

$$E\left(\sum_{i=1}^{m} v_i r_i\right) = \sum_{i=1}^{m} v_i \psi_i.$$

This metric provides an importance-weighted average of the rejection probabilities for the null hypotheses and is appropriate when the researcher is interested in maximizing the number of null hypotheses rejected, subject to the pre-specified relative importance of the null hypotheses.

The extended metrics presented above often lead to a nontrivial optimal configuration of parameters for the chain procedure. This property is illustrated in Section 6.

For a given metric (and corresponding application), the following algorithm can be used to find the optimal configuration of parameters for p-value-based and parametric chain procedures:

Step 1. Specify assumed study design parameters (e.g., sample sizes, effect sizes) which define the joint distribution of the test statistics under the alternative hypothesis.

Step 2. Specify appropriate search grids for the parameters of the testing procedure.

- For cyclical chain procedures, one must specify a grid of hypothesis weights and transition parameters.
- For serial chain procedures, one must specify a grid of hypothesis weights and transition parameters, and consider all $m!$ (or some pre-specified subset) possible orderings of the hypothesis tests.

Step 3. For each configuration of these parameters, compute the value of the chosen multiple testing metric. This computation can rely on Monte Carlo simulations, in which case data are simulated from the joint distribution of the test statistics, the outcome (i.e., significance or non-significance) of each test in the sequence is determined from simulated data, and the metric is computed by averaging over a large number of simulation runs. Alternatively, the computation can be performed directly, e.g., using numerical evaluation of appropriate multivariate probabilities.

Step 4. The optimal configuration of procedure parameters is found by maximizing the chosen metric over the search space of possible parameters.

In practice, as the inputs for the algorithm are subject to uncertainty, one may repeat this process for several sets of input parameters and choose the parameter configuration that provides the best results across the selected scenarios. For example, one might use a maximin principle, selecting the parameter configuration which has the best worst-case power among the scenarios. By this approach, if three sets of inputs identify three different optimal chain procedures, labeled A, B, and C, the optimal procedure ultimately selected is the one whose minimum metric value (across the three input scenarios) is highest among the minimum metric values for Procedures A, B, and C. Less conservatively, one might use an average of the assessments across the scenarios (e.g., mean of the chosen power metric) and select the parameter configuration with the maximum average value.

6 Clinical trial example

In this section we will illustrate the serial parametric chain procedure and optimization algorithm introduced using a targeted therapy clinical trial example. Consider a parallel design clinical trial in which the sponsor would like to simultaneously test the efficacy of a treatment in the general patient population, as well as in two pre-specified targeted subsets of the population, by comparison to control. The subgroups may be defined by clinical baseline characteristics of the patient (e.g., severity of illness), phenotypic characteristics of the patient (e.g., gender) or by a genotypic classifier. The three populations will be labeled Group 1 (overall population), Group 2 (first subgroup) and Group 3 (second subgroup). For simplicity, consider a balanced design and let n_i denote the sample size per arm in Group i, $i = 1, 2, 3$, where $n_1 = 300$, $n_2 = 175$ and $n_3 = 60$. The overlap between Groups 2 and 3 (number of patients in each treatment arm common to Groups 2 and 3) is $n_4 = 30$. Assume the trial involves no enrichment for the pre-specified subgroups, i.e., the sample sizes for the subgroups are assumed to be in proportion to their prevalence in the overall population.

Assume the primary efficacy endpoint (change in a clinician-rated scale) is normally distributed. Let δ_i denote the true mean treatment-placebo difference in Group i, $i = 1, 2, 3$, with a positive treatment difference indicating efficacy, and assume a common standard deviation (σ) across all treatment arms and groups. The family of null hypotheses to be tested is expressed as $H_i : \ \delta_i \leq 0$, $i = 1, 2, 3$. The null hypotheses are tested against one-sided alternatives and the FWER is set at $\alpha = 0.025$.

Let t_i denote the two-sample statistic for testing H_i, i.e.,

$$t_i = \frac{\widehat{\delta}_i}{s\sqrt{2/n_i}},\ i = 1, 2, 3,$$

where $\widehat{\delta}_i$ is the maximum likelihood estimate of δ and s is the sample pooled standard deviation. The three test statistics are asymptotically trivariate normal with the means

$$\theta_i = \frac{\delta_i}{\sigma\sqrt{2/n_i}},\ i = 1, 2, 3.$$

and unit variances. The correlations among the test statistics can be found directly as a function of the sample sizes, i.e.,

$$\rho_{12} = \sqrt{\frac{n_2}{n_1}} = 0.764,\quad \rho_{13} = \sqrt{\frac{n_3}{n_1}} = 0.447,\quad \rho_{23} = \frac{n_4}{\sqrt{n_2 n_3}} = 0.293,$$

where $\rho_{ij} = \text{corr}(t_i, t_j)$. Since the joint distribution of the test statistics is known, we apply the parametric chain procedure to test the three null hypotheses.

Note that the above test statistics assume no enrichment (i.e., oversampling) of the subgroups. In the case of population enrichment, we recommend adjusting the test statistics so that no test is overly influenced by the result of a sensitive subgroup. This is accomplished through direct re-weighting of the observed effect sizes from disjoint subsets of the population in the final test statistics. One method for doing this may be taken from Horvitz and Thompson (1952) as recommended by Zhao, Dmitrienko and Tamura (2010).

To optimize the parametric chain procedure for this example, several application-specific considerations must be taken into account. First, the treatment is expected to demonstrate greater effects in the subgroups than in the overall population. However, despite the targeted efficacy information, the treatment is believed to represent a valuable treatment option for patients outside these targeted groups. Thus, it is desirable to register the treatment for potential use in all patients, with specific data available to prescribers (via product label) on the increased effect in the targeted subgroups. Second, if efficacy is not demonstrated in the overall patient population but is demonstrated in one or more targeted subgroups, it would still be desirable to register the treatment for use in the targeted subgroup(s) only. Based on this information, there are several reasonable metrics to consider for optimizing the chain procedure in this clinical trial application. One reasonable metric is generalized disjunctive power based on the probability of demonstrating efficacy in at least two of the three populations. An alternative metric applicable to this problem is proportional power, using

weights which reflect the sponsor's view of the relative value of demonstrating efficacy in the three populations. We will focus on the proportional power metric below. With this metric, we assume the sponsor's prior importance weightings on the three null hypotheses are $v_1 = 0.5$, $v_2 = 0.35$ and $v_3 = 0.15$. These weights are roughly proportional to the sizes of Groups 1, 2 and 3 and, thus, somewhat reflect the market value of a successful outcome in each group.

Assuming the following effect sizes in the overall population and two subgroups

$$\frac{\delta_1}{\sigma} = 0.25, \quad \frac{\delta_2}{\sigma} = 0.35, \quad \frac{\delta_3}{\sigma} = 0.5,$$

the optimal ordering of the three null hypotheses is H_1, H_2 and H_3. The optimal hypothesis weights and transition matrix for the ordered null hypotheses are given by

$$(0.5, 0.4, 0.1), \quad \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

and the optimal weighted proportional power is 83.8%. The α allocation and α propagation rules of the optimal serial parametric chain procedure are depicted in Figure 3. The decision rules on the p-value scale for the optimal procedure as well as the p-value-based and parametric fallback procedures are displayed in Table 4. The computations were performed using an R program which utilized the MVTNORM package (Genz and Bretz, 2009).

Starting with the parametric chain procedure, it is clear that the optimal procedure spends most of α early in the sequence (on Groups 1 and 2) to improve the power of the tests with higher importance weightings. Further, the p-value-based and parametric fallback procedures were performed based on the hypothesis ordering and hypothesis weights prescribed by the optimization algorithm for the serial parametric chain procedure (note that the optimal transition parameters could not be used). As was pointed out in Section 3, a direct comparison of the decision rules between the chain and two fallback procedures is generally complicated since hypothesis weights have slightly different interpretations. It is, however, clear that, in this case, the two fallback procedures put more emphasis on the null hypotheses toward the end of the sequence, i.e., the associated p-value cutoffs are fairly high for the null hypothesis H_2 when H_1 is not rejected and for the null hypothesis H_3 when either one of or both H_1 and H_2 are not rejected. This results in power loss for the null hypotheses associated with Groups 1 and 2 and reduces weighted proportional power.

The last row in Table 4 shows the weighted proportional power for the four procedures based on 100,000 simulations. The optimal chain procedure outperforms the p-value-based and parametric fallback procedures by 3.7 and 2.6 percentage points,

respectively, in terms of the weighted proportional power. It is worth comparing the power gain of the optimal chain procedure over the parametric fallback procedure to the power gain of the parametric fallback procedure over the p-value-based fallback procedure. The latter gain is relatively small (1.1%) and is due to the fact that the parametric fallback procedure utilizes information about the joint distribution of the test statistics in this problem. The optimal chain procedure more than doubles this gain by identifying the best configuration of parameters given the study design assumptions.

This example illustrates the utility of the chain procedure and optimization algorithm (based on appropriate metrics) for a practical clinical trial example involving multiple populations. It is important to point out that the study design and testing strategy in this example are quite efficient compared to a traditional approach of conducting multiple single population trials to separately address the null hypotheses of interest. In this case, decisions about three populations of interest can be made in one trial, rather than three different trials. This translates into savings for the sponsor and, potentially, increased access to approved, targeted therapies for patients. The *a priori* specification of null hypotheses for the key subgroups, with appropriate control on the overall Type I error rate, allows valid inference on multiple groups of patients from the same trial.

Lastly, it is important to assess the sensitivity of the optimal serial parametric chain procedure to the assumptions made in the derivation of optimal weights, transition parameters and hypothesis ordering. The same algorithm has been applied to derive optimal serial parametric chain procedures for three other scenarios defining sets of effect sizes in the three groups. The effect sizes are displayed in Table 5 along with weighted proportional power for the original (labeled Procedure A) and optimal parametric chain procedures derived for each individual scenario. It follows from this table that Procedure A performs very well under all three scenarios. Its power is either equal to or only trivially lower than the power of scenario-specific optimal procedures, e.g., less than 1% on an absolute scale. This suggests that Procedure A is quite robust and should be expected to maintain power at a desirable level in the actual trial.

7 Discussion

In this paper, we have presented a class of closed multiple testing procedures, known as *chain procedures*, for evaluating a set of m null hypotheses. We have introduced parametric versions of the chain procedure for use in problems where the joint distribution of the hypothesis test statistics is known and can be incorporated into the decision rule. As such, we have provided a framework for a broad, flexible class of

mutliple testing procedures. Since chain procedures are closed testing procedures, they control the FWER in the strong sense. In addition, we have provided an algorithm for optimally selecting parameters of chain procedures, along with a discussion of appropriate optimization criteria and power metrics. Lastly, we have illustrated the utility of chain procedures via a clinical trial example in which we applied a parametric serial chain procedure and compared the results obtained by this procedure with those of other stepwise multiple testing procedures employed in clinical trials. We now close with a brief discussion of some properties of chain procedures, with emphasis on their broad applicability for clinical trials.

Chain testing is applicable to many multiplicity problems in clinical trials, including those arising from multiple endpoints, multiple dose-control comparisons, and multiple populations. Application of the metrics presented in Section 5 allows prescription of procedure parameters which optimize the resultant chain testing procedure for the specific application. In the multiple populations clinical trial example earlier in this paper, we noted that the chain procedure could be optimized for rejection of a non-specific subset of null hypotheses (i.e., demonstrating efficacy in at least two populations). In the same application, a sponsor may choose to optimize for rejection of a semi-specific subset of null hypotheses, e.g., demonstrating efficacy in the overall population and at least one of the pre-specified subpopulations, to facilitate broad registration for the drug along with specific labeling for each population. In a trial with multiple endpoints, a sponsor may wish to demonstrate efficacy via both a clinical outcome and a biomarker. With more than one acceptable endpoint for each of these (e.g., death, hospitalization and total cholesterol, blood pressure), a chain procedure which is optimized for the rejection of at least one clinical endpoint and at least one biomarker endpoint (i.e., optimized using a subset disjunctive power metric) may be constructed. Beyond the noted flexibility of the procedures, chain testing is readily adaptable to application-specific constraints defined by the researcher. For example, in some applications, one may desire to test additional populations or endpoints only if at least one of the other tests had resulted in significance. These serial constraints are readily incorporated into chain procedures. Such null hypotheses (corresponding to the additional populations or endpoints) may be assigned zero *a priori* hypothesis weights with their tests governed by appropriate alpha propogation rules. This approach is particularly attractive when the number of hypotheses to be tested is large or when insufficient information is available about an endpoint (e.g., effect size estimate) or subgroup (e.g., expected population size) to prospectively include in the alpha allocation scheme but the clinical trial will provide critical information about the endpoint or subgroup.

With respect to optimal application of chain procedures, it is important to note that the class of p-value-based chain procedures includes as special cases a few popular testing procedures. For example, the fixed-sequence and fallback procedures may

each be expressed as p-value-based serial chain procedures via appropriate specification of hypothesis weights and transition matrix. In addition, the Holm procedure may be expressed as a p-value-based cyclical chain procedure. Each of the special cases may be prescribed by the given chain testing optimization algorithm when the associated application dictates. Thus, the introduction of the class of chain procedures and the associated optimization algorithm advances both the armamentarium of multiple testing procedures and the appropriate application of commonly used, currently available procedures. In addition, extensions of the metrics introduced in Section 5 are possible. For example, if there is a considerable amount of uncertainty with respect to the assumptions regarding the magnitude of the treatment effect for the various tests, the set of possible chain procedures may be constrained to those which ensure testing of particular (or all) hypotheses in the sequence, i.e., the weights for the particular hypotheses may be constrained to be strictly positive and bounded. Moreover, an optimization metric may include conditional probabilities, rather than only the marginal probabilities, at different stages in the testing sequence. In other words, one may want to evaluate the probability of rejecting H_j given failure to reject H_i for all $i \neq j$ in the application-specific optimization algorithm. The adaptability of the optimization method ensures that the final chosen chain procedure fits the needs of the specific problem. The general algorithm provided in Section 5 is readily suited to the broad possibilities of optimzation metrics.

As noted in Sections 3 and 4, parametric chain procedures offer increased power over p-value-based chain procedures and the parametric methods may be constructed whenever the joint distribution of the test statistics is known. For multiplicity problems involving multiple populations evaluated via a common endpoint, in particular, the parametric chain procedure is readily constructed (since the correlations are direct functions of the sample sizes). The optimization considerations presented in Section 5 apply fully to all chain procedures, including parametric procedures. As the relative amount of literature focused on parametric multiple testing procedures is somewhat limited, there are no known special cases of parametric chain procedures to point out. It was noted in Section 3, that the parametric chain procedure with appropriate transition parameters provides a *modified* parametric fallback procedure. This procedure is not identical to, and tends to be more powerful than, the parametric fallback procedure offered by Huque and Alosh (2008) which, as the authors pointed out, relies on asymptotics for control of the Type I error rate.

The class of chain testing procedures both advances the armamentarium of available procedures and supports appropriate use of currently available testing procedures. The flexibility of this class of testing procedures makes them suitable for broad use within clinical trials. Lastly, the presentation of optimization metrics and considerations should have broad impact for clinical trial researchers considering complex multiple testing problems regardless of the procedure ultimately used.

Acknowledgements

We would like to thank the Associate Editor and an anonymous reviewer for their comments that helped improve the presentation of the material.

References

[1] Benjamini, Y., Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics*. 24, 407–418.

[2] Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*. 28, 586–604.

[3] Burman, C.-F., Sonesson, C., Guilbaud, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine*. 28, 739–761.

[4] Dmitrienko, A., Offen, W.W., Westfall, P.H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine*. 22, 2387–2400.

[5] Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., Offen, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*. SAS Press: Cary, NC.

[6] Dmitrienko, A., Tamhane, A.C., Wiens, B.L. (2008). General multistage gatekeeping procedures. *Biometrical Journal*. 50, 667-677.

[7] Dmitrienko, A., Bretz, F., Westfall, P.H., Troendle, J., Wiens, B.L., Tamhane, A.C., Hsu, J.C. (2009). Multiple testing methodology. *Multiple Testing Problems in Pharmaceutical Statistics*. Dmitrienko, A., Tamhane, A.C., Bretz, F. (editors). Chapman and Hall/CRC Press, New York.

[8] Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*. 50, 1096–1121.

[9] Genz, A., Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Springer Verlag, Heidelberg.

[10] Grechanovsky, E., Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference*. 76, 79–91.

[11] Hochberg, Y., Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

[12] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika*. 75, 800–802.

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

[14] Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47, 663–685.

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

[16] Huque, M.F., Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference*. 138, 321–335.

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

[18] Naik, U.D. (1975). Some selection rules for comparing p processes with a standard. *Communications in Statistics. Series A*. 4, 519–535.

[19] Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson J.D., Krishen, A., Liu, T., Ryder, S., Sankoh, A.J., Wang, J., Yeh, C.H. (2007). Multiple co-primary endpoints: Medical and statistical solutions. A report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America. *Drug Information Journal*. 41, 31–46.

[20] Senn, S., Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*. 6, 161–170.

[21] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 63, 655–660.

[22] Wang, S., O'Neill, R.T., Hung H.M.J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*. 6, 227–244.

[23] Westfall, P.H., Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, New York.

[24] Westfall, P. H., Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference*. 99, 25–40.

[25] Wiens, B. (2003). A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*. 2, 211–215.

[26] Wiens, B., Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*. 15, 929–942.

[27] Zhao, Y.D., Dmitrienko, A., Tamura, R. (2010). On optimal designs of clinical trials with a sensitive subgroup. *Statistics in Biopharmaceutical Research*. 2, 72–83.

Appendix

General parametric chain procedures

General parametric chain procedures in problems with a data-driven or fixed testing sequence are constructed using the closure principle. Consider the closed family associated with the null hypotheses $H_1, \ldots, H_m$, which consists of all non-empty intersections of these m null hypotheses. To define a closed testing procedure, it is sufficient to define α-level local tests for all intersection hypotheses. Given the local tests, the closed testing procedure rejects a null hypothesis if and only if it rejects all intersection hypotheses including this particular null hypothesis.

Serial chain procedures. Select an intersection hypothesis H_I from the closed family, $I \subseteq M$. A parametric chain procedure is defined as the closed testing procedure which rejects H_I if at least one of the following conditions is satisfied:

$$v_i(I)t_i \geq c_i(I),\ i \in I,$$

where $v_i(I)$, $i \in I$, are hypothesis weights and $c_i(I)$, $i \in I$, are critical values. The hypothesis weights are computed sequentially using the following algorithm (note that $\delta_i(I) = 1$ if $i \in I$ and 0 otherwise).

Step $k = 1$**.** Let $s_1(I) = w_1$ and $v_1(I) = \delta_1(I)s_1(I)$.

Step $k = 2, \ldots, m$**.** Let

$$s_k(I) = w_k + \sum_{i=1}^{k-1}(1 - \delta_i(I))g_{ik}s_i(I)$$

and $v_k(I) = \delta_k(I)s_k(I)$.

Further, the critical values $c_i(I)$, $i \in I$, are defined to ensure that the test for H_I is an α-level or more conservative test. Let $\ell(I)$ denote the smallest index $i \in I$ for which $v_i(I) > 0$. The critical values are computed sequentially using the algorithm given below.

Step $k = 1, \ldots, m$**.** Select all intersection hypotheses H_I with $\ell(I) = k$ and consider the following three cases:

- Case 1. If $I = \{k, \ldots, m\}$, let $c_i(I) = d_k$, $i \in I$, where d_k is found from

$$P\left(\bigcap_{i \in I}(v_i(I)T_i < d_k)\right) = 1 - \alpha.$$

- Case 2. If $I = \{k\}$, let $c_k(I) = d_k$.
- Case 3. Otherwise, let $c_k(I) = d_k$ and $c_i(I) = d(I)$, $i \in I^*$, where $I^* = I \backslash \{k\}$. Here $d(I)$ is found from

$$P\left((v_k(I)T_k < d_k) \bigcap_{i \in I^*} (v_i(I)T_i < d(I))\right) = 1 - \alpha.$$

Since each intersection hypothesis is tested using an α-level or more conservative test, the closed testing procedure defined above controls the FWER in the strong sense at the α level.

Cyclical chain procedures. Consider an arbitrary intersection hypothesis

$$H_I = \bigcap_{i \in I} H_i,\ I \subseteq M,$$

where M is the index set containing indices of all null hypotheses. A parametric chain procedure is defined as the closed testing procedure which rejects the intersection hypothesis H_I if at least one of the following conditions is satisfied:

$$v_i(I)t_i \geq c(I),\ i \in I,$$

where $v_i(I)$, $i \in I$, are intersection-specific hypothesis weights and $c(I)$ is an intersection-specific critical value. The hypothesis weights are defined using an algorithm similar to that developed in Bretz et al. (2009) for Bonferroni-based chain procedures. In this algorithm $I_0 = M$, $v_i(I_0) = w_i$, $i = 1, \ldots, m$, and $g_{ij}(I_0) = g_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, m$. Further, $\ell = m - |I|$, where $|I|$ is the number of elements in the index set I, and $\{i_1, \ldots, i_\ell\}$ are the indices not included in I, i.e., $M \backslash I = \{i_1, \ldots, i_\ell\}$.

Step $k = 1, \ldots, \ell - 1$**.** Let $I_k = I_{k-1} \backslash \{i_k\}$ and

$$v_i(I_k) = w_i + v_{i_k} g_{i_k,i}(I_{k-1}),\ i \in I_k,$$
$$g_{ij}(I_k) = \frac{g_{ij} + g_{i,i_k} g_{i_k,j}}{1 - g_{i,i_k} g_{i_k,i}},\ i \in I_k,\ j \in I_k.$$

Step ℓ**.** Let $I_\ell = I_{\ell-1} \backslash \{i_\ell\}$ and note that $I_\ell = I$. The hypothesis weights $v_i(I)$, $i \in I$, are defined as follows:

$$v_i(I) = w_i + v_{i_\ell} g_{i_\ell,i}(I_\ell),\ i \in I.$$

The critical value $c(I)$ is found from

$$P\left(\bigcap_{i\in I}(v_i(I)T_i < c(I))\right) = 1-\alpha$$

and thus the intersection hypothesis H_I is tested using an α-level local test. Since each local test is an α-level test, the closed testing procedure controls the FWER in the strong sense at the α level.

Further, using the monotonicity property formulated in Bretz et al. (2009), it can be shown that parametric chain procedures are, in fact, based on the following stepwise algorithm. In this algorithm $T_1,\ldots,T_m$ have the same joint distribution as $t_1,\ldots,t_m$ under the global null hypothesis.

Step 1. Let $I_1 = M$, where $M = \{1,\ldots,m\}$ is the index set containing indices of all null hypotheses. The hypothesis weights and transition matrix used in Step 1 are given by

$$w_i(I_1) = w_i,\ i\in I_1,\quad g_{ij}(I_1) = g_{ij},\ i\in I_1,\ j\in I_1.$$

Find the critical value $c(I_1)$ from

$$P\left(\bigcup_{i\in I_1}(w_i(I_1)T_i \geq c(I_1))\right) = \alpha.$$

Define the weighted test statistics

$$t_i^*(I_1) = w_i(I_1)t_i,\ i\in I_1,$$

and let i_1 be the index of the largest test statistic. Reject H_{i_1} if $t_{i_1}^*(I_1)\geq c(I_1)$ and go to Step 2. Accept all null hypotheses otherwise.

Step $k = 2,\ldots,m-1$**.** Let I_k denote the index set containing indices of all remaining null hypotheses. Update the hypothesis weights and transition matrix to account for the rejection of the null hypothesis $H_{i_{k-1}}$ in Step $k-1$, i.e., let

$$\begin{aligned} w_i(I_k) &= w_i + w_{i_{k-1}}g_{i_{k-1},i}(I_{k-1}),\ i\in I_k,\\ g_{ij}(I_k) &= \tfrac{g_{ij}+g_{i,i_{k-1}}g_{i_{k-1},j}}{1-g_{i,i_{k-1}}g_{i_{k-1},i}},\ i\in I_k,\ j\in I_k. \end{aligned}$$

Using these hypothesis weights, find the critical value $c(I_k)$ from

$$P\left(\bigcup_{i\in I_k}(w_i(I_k)T_i \geq c(I_k))\right) = \alpha$$

and define the weighted test statistics

$$t_i^*(I_k) = w_i(I_k)t_i,\ i \in I_k.$$

Let i_k be the index of the largest test statistic. Reject H_{i_k} if $t_{i_k}^*(I_k) \geq c(I_k)$ and go to Step $k+1$. Accept the null hypotheses H_i, $i \in I_k$, otherwise.

Step m**.** Let i_m denote the index of the last remaining null hypothesis and $I_m = \{i_m\}$. Rejected this null hypothesis if $t_{i_m} \geq c(I_m)$, where the critical value is found from $P(T_{i_m} \geq c(I_m)) = \alpha$, and accept it otherwise.

Computation of proportional power

Consider the three-hypothesis problem introduced in Section 6. Let $R_i = \{\text{Reject } H_i\}$, $A_i = \{\text{Accept } H_i\}$, $i = 1, 2, 3$. Weighted proportional power for the parametric chain procedure is given by

$$v_1\psi_1 + v_2\psi_2 + v_3\psi_3,$$

where the probabilities of rejecting H_1, H_2 and H_3 are found using the following equations:

$$\psi_1 = P(R_1) = P(t_1 \geq c_1^*),$$

$$\begin{aligned}\psi_2 &= P(R_2) = P(R_1R_2) + P(A_1R_2)\\ &= P(t_1 \geq c_1^*, t_2 \geq c_2^*) + P(t_1 < c_1^*, t_2 \geq c_3^*),\end{aligned}$$

$$\begin{aligned}\psi_3 &= P(R_3) = P(R_1R_2R_3) + P(A_1R_2R_3) + P(R_1A_2R_3) + P(A_1A_2R_3)\\ &= P(t_1 \geq c_1^*, t_2 \geq c_2^*, t_3 \geq c_4^*) + P(t_1 < c_1^*, t_2 \geq c_3^*, t_3 \geq c_6^*)\\ &+P(t_1 \geq c_1^*, t_2 < c_2^*, t_3 \geq c_5^*) + P(t_1 < c_1^*, t_2 < c_3^*, t_3 \geq c_7^*).\end{aligned}$$

Table 1. Closed testing representation of the serial p-value-based chain procedure in a problem of testing three null hypotheses H_1, H_2 and H_3. Each row defines an intersection hypothesis in the closed family and associated local test. The closed testing procedure rejects a hypothesis if it rejects all intersection hypotheses including the selected hypothesis.

Intersection hypothesis	Local test (rejection rule)
$H_1 \cap H_2 \cap H_3$	$p_1 \leq \alpha w_1$ or $p_2 \leq \alpha w_2$ or $p_3 \leq \alpha w_3$
$H_1 \cap H_2$	$p_1 \leq \alpha w_1$ or $p_2 \leq \alpha w_2$
$H_1 \cap H_3$	$p_1 \leq \alpha w_1$ or $p_3 \leq \alpha(w_3 + w_2 g_{23})$
H_1	$p_1 \leq \alpha w_1$
$H_2 \cap H_3$	$p_2 \leq \alpha(w_2 + w_1 g_{12})$ or $p_3 \leq \alpha(w_3 + w_1 g_{13})$
H_2	$p_2 \leq \alpha(w_2 + w_1 g_{12})$
H_3	$p_3 \leq \alpha$

Table 2. Closed testing representation of the serial parametric chain procedure in a problem of testing three null hypotheses H_1, H_2 and H_3. Each row defines an intersection hypothesis in the closed family and associated local test. The closed testing procedure rejects a hypothesis if it rejects all intersection hypotheses including the selected hypothesis.

Intersection hypothesis	Local test (rejection rule)
$H_1 \cap H_2 \cap H_3$	$w_1 t_1 \geq c_1$ or $w_2 t_2 \geq c_1$ or $w_3 t_3 \geq c_1$
$H_1 \cap H_2$	$w_1 t_1 \geq c_1$ or $w_2 t_2 \geq c_2$
$H_1 \cap H_3$	$w_1 t_1 \geq c_1$ or $(w_3 + w_2 g_{23}) t_3 \geq c_3$
H_1	$w_1 t_1 \geq c_1$
$H_2 \cap H_3$	$(w_2 + w_1 g_{12}) t_2 \geq c_4$ or $(w_3 + w_1 g_{13}) t_3 \geq c_4$
H_2	$(w_2 + w_1 g_{12}) t_2 \geq c_4$
H_3	$t_3 \geq c_5$

Table 3. Closed testing representation of the cyclical parametric chain procedure in a problem of testing three null hypotheses H_1, H_2 and H_3. Each row defines an intersection hypothesis in the closed family and associated local test. The closed testing procedure rejects a null hypothesis if it rejects all intersection hypotheses including the selected hypothesis.

Intersection hypothesis	Local test (rejection rule)
$H_1 \cap H_2 \cap H_3$	$w_1 t_1 \geq c_{123}$ or $w_2 t_2 \geq c_{123}$ or $w_3 t_3 \geq c_{123}$
$H_1 \cap H_2$	$(w_1 + w_3 g_{31}) t_1 \geq c_{12}$ or $(w_2 + w_3 g_{32}) t_2 \geq c_{12}$
$H_1 \cap H_3$	$(w_1 + w_2 g_{21}) t_1 \geq c_{13}$ or $(w_3 + w_2 g_{23}) t_3 \geq c_{13}$
H_1	$t_1 \geq c_1$
$H_2 \cap H_3$	$(w_2 + w_1 g_{12}) t_2 \geq c_{23}$ or $(w_3 + w_1 g_{13}) t_3 \geq c_{23}$
H_2	$t_2 \geq c_2$
H_3	$t_3 \geq c_3$

Table 4. P-value cutoffs and weighted proportional power for the optimal serial parametric chain (PC), p-value-based fallback (NF) and parametric fallback (PF) procedures in the clinical trial example based on a one-sided $\alpha = 0.025$.

Condition	P-value cutoff		
	PC procedure	NF procedure	PF procedure
P-value cutoff for H_1			
	0.0222	0.0125	0.0125
P-value cutoff for H_2			
H_1 is rejected	0.0249	0.0213	0.0213
H_1 is not rejected	0.0060	0.0088	0.0132
P-value cutoff for H_3			
H_1 and H_2 are rejected	0.0249	0.0250	0.0250
H_1 is rejected and H_2 is not rejected	0.0001	0.0034	0.0047
H_1 is not rejected and H_2 is rejected	0.0029	0.0125	0.0125
H_1 and H_2 are not rejected	0.0001	0.0034	0.0047
Power (%)	83.8	80.1	81.2

Table 5. Weighted proportional power for the original optimal serial parametric chain procedure (Procedure A) under three alternative scenarios and weighted proportional power of scenario-specific optimal serial parametric chain procedures.

Effect size			Power (%)	
Group 1	Group 2	Group 3	Procedure A	Optimal procedure
0.30	0.45	0.45	91.9	92.4
0.25	0.45	0.50	87.4	87.4
0.30	0.40	0.50	92.1	92.1

Figure 1. Serial chain testing procedure in a problem of testing three null hypotheses H_1, H_2 and H_3.

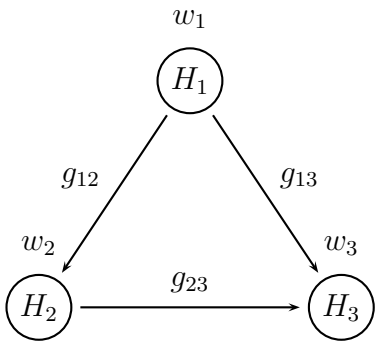

Figure 2. Cyclical chain testing procedure in a problem of testing three null hypotheses H_1, H_2 and H_3.

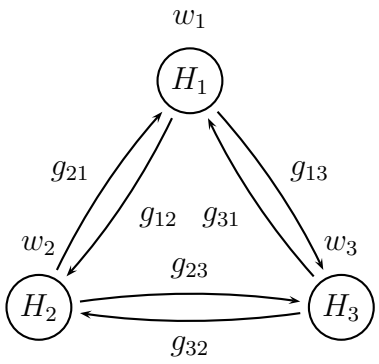

Figure 3. Optimal serial chain testing procedure in the clinical trial example.