POWER CALCULATIONS IN CLINICAL TRIALS WITH COMPLEX CLINICAL OBJECTIVES

Alex Dmitrienko∗, Gautier Paux† and Thomas Brechenmacher‡

ABSTRACT

Over the past decade, a variety of powerful multiple testing procedures have been developed for the analysis of clinical trials with multiple clinical objectives based, for example, on several endpoints, dose-placebo comparisons and patient subgroups. Sample size and power calculations in these complex settings are not straightforward and, in general, simulation-based methods are used. In this paper, we provide an overview of power evaluation approaches in the context of clinical trials with multiple objectives and illustrate the key principles using case studies commonly seen in the development of new therapies.

1. Introduction

It is becoming increasingly common to study complex/multivariate clinical objectives in clinical trials, especially in confirmatory Phase III trials (Dmitrienko, D’Agostino and Huque, 2013). In a confirmatory setting, clinical trial sponsors are often interested in pursuing multiple regulatory claims for new treatments based on evaluating several endpoints, two or more doses compared to a control or characterizing the treatment effect in multiple subgroups. As the definition of success in these clinical trials is based on multiple significance tests, the probability of erroneously claiming the effectiveness of a new treatment, i.e., the Type I error rate, must be controlled by applying a multiple testing procedure (MTP). Over the past decades, a variety of powerful MTPs have been developed and many of them have found applications in clinical trials. This includes, for example, the Holm, Hochberg and Hommel procedures, as well as other methods that allow the sponsor to specify the relative importance of different objectives, e.g., the fixed-sequence, fallback and chain procedures (Wiens, 2003; Wiens and Dmitrienko, 2005; Millen and Dmitrienko, 2011) and gatekeeping procedures (Dmitrienko and Tamhane, 2009; Dmitrienko, D’Agostino and Huque, 2013). To provide an example of a successful application of a powerful MTP in confirmatory Phase III settings, Hommel-based gatekeeping procedures were used in the lurasidone clinical trials in patients with schizophrenia (Meltzer et al., 2011; Nasrallah et al., 2013). This gatekeeping procedure was utilized to protect the overall Type I error rate in a complex multiplicity problem which involved evaluation of the efficacy of two or three doses of lurasidone based on a primary clinical endpoint and one or two key secondary endpoints.

An important aspect of designing clinical trials with simple or complex clinical objectives includes the determination of the sample size (number of patients in fixed-design trials

∗Quintiles, Inc., 4820 Emperor Blvd, Durham, NC 27703, U.S.A.
E-mail: alex.dmitrienko@quintiles.com
†Institut de Recherches Internationales Servier (I.R.I.S.), 50 Rue Carnot, 92150 Suresnes, France
‡Novartis, 10 rue Lionel Terray, 92500 Rueil-Malmaison, France
Key words: Clinical trials; Familywise error rate; Multiple testing procedures; Power; Sample size; Simulation.
or number of events in event-driven trials). In a traditional setting with a single clinical objective, sponsors must ensure that the primary statistical test has desirable power (for example, 80%) to detect a treatment effect of a given size. In the context of clinical trials with several potential regulatory claims based on multiple clinical objectives, the traditional approaches to sample size and power calculation need to be extended in several directions (Dmitrienko and D’Agostino, 2013). The calculations should reflect the multiple testing strategy to be used; in particular, the standard power function needs to be replaced by a multivariate success criterion. In general, closed-form expressions for success criteria cannot be derived and sample size/power calculations are typically performed using simulation methods.

In this paper, we examine the current methodology for power and sample size calculations in clinical trials that are designed to address multiple objective. It is worth noting that, while we emphasize applications to confirmatory Phase III trials, the sample principles can be applied to Phase II trials. We focus on the problem of quantitative evaluation of all factors that affect the probability of success in a clinical trial. We emphasize the importance of a systematic comparison of trial designs and analysis methods, including appropriate multiplicity adjustments, and discuss the selection of relevant success criteria that are best aligned with the trial’s overall goals.

This paper is organized as follows. Section 2 introduces common multiplicity problems arising in clinical trial settings as well as case studies that will be used throughout this paper. Section 3 defines the main classes of multiple testing procedures used in clinical trial applications, including traditional procedures based on nonparametric, semiparametric and parametric approaches, as well as more advanced gatekeeping procedures. Main components of power calculations in clinical trials with multiple objectives are discussed in Section 4. Section 5 reviews the key principles of power calculations in clinical trials with emphasis on a general approach to systematic quantitative assessment of all key factors on the trial’s operating characteristics known as the Clinical Scenario Evaluation approach. Sections 6, 7 and 8 illustrate power calculations in clinical trials with complex objectives using the three case studies introduced in Section 2.

2. Multiplicity problems in clinical trials

It was pointed out in the introduction that “multivariate” clinical objectives are commonly employed in clinical trials to characterize different aspects of the new treatment’s efficacy profile and pursue associated regulatory claims. Analysis of these objectives requires special attention to the associated multiplicity issues and application of appropriate statistical methods aimed at controlling the overall Type I error rate. This section provides a brief introduction to multiplicity problems arising in clinical trials and defines a hypothesis-testing framework which will be used in the subsequent sections.

Several case studies will be utilized throughout this paper to illustrate multiplicity problems arising in clinical trial settings:

- Case study 1 (multiple endpoints). A clinical trial with the primary endpoint accompanied by several key secondary endpoints.
- Case study 2 (multiple dose-control comparisons). A clinical trial with multiple doses of a new treatment compared to a control treatment.
- Case study 3 (multiple dose-control comparisons and multiple endpoints). A clinical
trial with multiple endpoints (primary and key secondary endpoints) evaluated at several doses of a new treatment.

Specific examples of clinical trials with the multiplicity problems listed above are provided in Sections 6, 7 and 8.

A common feature of the multiplicity issues in the three case studies is that performing multiple tests in a univariate manner based on unadjusted treatment effect $p$-values increases the overall probability of an incorrect conclusion (Type I error rate). To define a formal framework and introduce an appropriate definition of the error rate, we will consider a general hypothesis-testing problem with $m$ null hypotheses of no treatment effect. The hypotheses for which the sponsor wishes to make formal claims are denoted by

$$H_1, \ldots, H_m.$$  

To provide a quick example, the null hypotheses in Case study 2 are defined in terms of the treatment differences for the individual dose-control comparisons. Suppose that the efficacy of $m$ doses is evaluated compared to a control and a larger value of the treatment difference indicates a beneficial effect. Let $\theta_i$ denote the true treatment effect in the $i$th dosing group relative to the control, $i = 1, \ldots, m$. The null hypotheses of no treatment effect are defined as follows:

$$H_1 : \theta_1 \leq \delta, \ldots, H_m : \theta_m \leq \delta,$$

where $\delta$ defines the minimal clinically important difference. This parameter is typically set to 0 when superiority assessments are performed and to a negative value which represents a relevant margin in non-inferiority testing problems.

The setting presented above includes a single set (or family) of null hypotheses and can be viewed as a one-dimensional multiplicity problem with a single source of multiplicity. It is important to contrast this setting with a more advanced setting which involves two or more families of null hypotheses associated with several sources of multiplicity. This multidimensional setting is typically encountered in clinical trials with hierarchically ordered endpoints, for example, in Case study 3. Consider again the set of $m$ null hypotheses of interest, i.e., $H_1, \ldots, H_m$. These null hypotheses are grouped into $k$ ordered families:

- Family $F_1 : H_1, \ldots, H_{M_1}$,
- Family $F_2 : H_{M_1+1}, \ldots, H_{M_2}$,
- \ldots
- Family $F_k : H_{M_{k-1}+1}, \ldots, H_{M_k}$.

Here $m_i$ denotes the number of null hypotheses in Family $F_i$ with $M_i = m_1 + \ldots + m_i$, $i = 1, \ldots, k$, and $m_1 + \ldots + m_k = m$.

The ordered families are examined in a sequential manner starting with Family $F_1$. Family $F_i$ serves as a gatekeeper for Family $F_{i+1}$, $i = 1, \ldots, k - 1$, which means that the rejection of the hypotheses included in $F_{i+1}$ is contingent upon the outcome of the treatment effect tests carried out in $F_i$.

To illustrate the process of defining hypothesis families in a multiplicity problem with ordered clinical endpoints, consider Case study 1. The primary endpoint test in this case study determines the overall trial outcome and provides the basis for the main regulatory claim. Given this, it is natural to put the associated null hypothesis in the primary family.
(F₁). The additional endpoints (key secondary endpoints) provide supportive evidence of treatment efficacy and can be placed in one or more secondary families (F₂ through Fₖ). A similar approach can be used in Case study 3 to account for the relative importance of the primary and key secondary endpoints. Several families of null hypotheses can be set up as follows: the null hypotheses corresponding to the primary dose-control comparisons are placed in the primary family (F₁) and the null hypotheses based on the secondary endpoints are included in the other families.

In order to support regulatory claims in a confirmatory trial or, in general, a Phase III development program, the program’s sponsor is required to protect the overall Type I error rate at the trial level. The specific definition of error rate control used in Phase III trials is known as the strong control of the familywise error rate (FWER). This definition is applied to a given family of null hypotheses (H₁, ..., Hₘ) and states that the probability of erroneously rejecting any true null hypothesis in this family must not exceed a pre-defined level (e.g., one-sided α = 0.025) regardless of the actual subset of the true null hypotheses. Formally, this means that

\[ P(\text{Reject at least one null hypothesis } Hᵢ, i ∈ T | Hᵢ, i ∈ T, \text{ are true}) ≤ α \]

for any index set of true null hypotheses \( T ∈ \{1, \ldots, m\} \).

A key feature of this definition is that, under strong FWER control, the probability of an erroneous conclusion is protected for any set of null hypotheses. This implies that the probability of incorrectly making a regulatory claim is controlled for any set of clinical objectives.

Multiple testing procedures (MTPs) that provide strong FWER control in the general multiplicity problems introduced above are discussed in Section 3. Section 3.2 introduces main classes of MTPs used in traditional multiplicity problems with a single family of null hypotheses. Further, MTPs designed for addressing the more advanced multiplicity problems with several families of null hypotheses, termed gatekeeping procedures, are discussed in Section 3.3.

3. Multiple testing procedures

In this section, we provide a high-level summary of MTPs commonly used in clinical trials. More details on the individual MTPs introduced below can be found in recently published review papers and tutorials, for example, Dmitrienko, D’Agostino and Huque (2013) and Dmitrienko and D’Agostino (2013).

3.1. Classification of multiple testing procedures

A variety of MTPs can be applied in a general clinical trial setting and the selection of the most appropriate procedure or procedures is typically driven by the amount of available clinical information (information on relevant logical restrictions) and statistical information (distributional information). Table 1 defines a classification scheme which groups commonly used MTPs into six categories based on the imposed logical restrictions and information on the joint distribution of the hypothesis test statistics.

A general description of the two factors in this classification scheme is provided below.

Logical restrictions

A priori hypothesis ordering is not always feasible in a clinical trial if relevant clinical information is not available to arrange the individual objectives in the order of importance. For
example, in a clinical trial with multiple dose-placebo comparisons such as Case study 2, the sponsor may not be able to determine which dose will have the highest efficacy at the planning stage and thus it may be challenging to order the null hypotheses in a logical way. In that case, it is most appropriate to apply single-step MTPs such as the Bonferroni and Dunnett procedures (all hypotheses are tested simultaneously) or data-driven MTPs such as the Hochberg and stepwise Dunnett procedures (hypotheses are tested in the order determined by the significance of the hypothesis test statistics). If reliable clinical information is available to set up logical restrictions among the null hypotheses and establish a clinically meaningful testing sequence, procedures that assume a pre-specified testing sequence can be used.

**Distributional information**

Another very important consideration in the choice of an MTP is the information on the joint distribution of the hypothesis test statistics. Depending on the amount of statistical information, two main classes can be defined: *p*-value-based (nonparametric or semiparametric) and parametric MTPs. Beginning with the first class, nonparametric MTPs rely on the univariate *p*-values without making any distributional assumptions whereas semiparametric MTPs need additional assumptions (but not a full specification of) on the joint distribution. Lastly, parametric MTPs require a fully specified joint distribution of the hypothesis test statistics, e.g., a multivariate normal distribution.

### 3.2. Single-family multiplicity problems

In traditional multiplicity problems with a single family of null hypotheses $H_1, \ldots, H_m$ (see, for example, Case study 2), several powerful MTPs can be used to control the FWER, depending on the clinical and statistical information available to the clinical trial’s sponsor.

If no information is readily available to define logical restrictions among the null hypotheses, single-step or stepwise procedures with a data-driven testing sequence are commonly used. The Bonferroni procedure is the most basic multiplicity adjustment based on a single-step testing algorithm. This procedure tends to be conservative when the number of null hypotheses is large or when the hypotheses are strongly correlated. For this reason,

---

1It needs to be pointed out that some MTPs such as the Holm, stepwise Dunnett, parametric fallback and parametric chain procedures can be defined as procedures with a data-dependent or pre-specified testing sequence. As shown in Dmitrienko and D’Agostino (2013, Appendix A), the underlying testing algorithms can be formulated as algorithms with a fixed or flexible testing sequence.
the Bonferroni procedure is not commonly used in clinical trial applications but serves as a foundation for the construction of more powerful MTPs. Stepwise procedures such as the Holm, Hochberg and Hommel procedures are popular MTPs in this particular setting because they take advantage of a data-driven testing sequence. This class of procedures examines the null hypotheses based on the ordered \( p \)-values and uses a monotonically increasing sequence of adjusted significance levels. The resulting MTPs are more powerful than single-step methods. Decision rules used by these MTPs are discussed below.

The Holm procedure (Holm, 1979) is based on a step-down algorithm to test the null hypotheses; namely, it begins with the most significant \( p \)-value and stops when a null hypothesis cannot be rejected (the untested hypotheses are then automatically accepted). This procedure is derived from the Bonferroni procedure and does not require any assumptions on the joint distribution of the hypothesis test statistics. By contrast, the Hochberg procedure (Hochberg, 1988) uses a step-up testing sequence and starts by examining the null hypothesis corresponding to the least significant \( p \)-value. This procedure stops as soon as a null hypothesis is rejected and the remaining hypotheses are automatically rejected. The Holm and Hochberg procedures use the same sequence of adjusted significance levels but the latter is uniformly more powerful as it is based on (a simplified version of) the Simes global test, which is more powerful than the Bonferroni procedure. The Hommel procedure (Hommel, 1988) is another stepwise Simes-based MTP based on a step-up testing algorithm. The Hommel procedure is uniformly more powerful than the Hochberg procedure since it is derived from the regular Simes global test.

It is important to note that, unlike Bonferroni-based procedures that control the FWER in the strong sense in any setting, the Simes-based procedures (Hochberg and Hommel procedures) are semiparametric and impose additional distributional assumptions to protect the FWER. In order to control the FWER, the hypothesis test statistics must follow a multivariate normal distribution with non-negative pairwise correlations (Sarkar, 2008). In confirmatory clinical trials, these conditions are met in several settings such as trials with multiple dose-control comparisons (see Case study 2), trials with two or more patient populations (e.g., overall population of patients and pre-defined target subpopulations) or trials with multiple endpoints that are known to be non-negatively correlated.

The MTPs discussed above utilize decision rules based on marginal \( p \)-values and either make no assumptions about the joint distribution of the hypothesis test statistics (nonparametric procedures) or make fairly “loose” distributional assumptions (semiparametric procedures) to preserve the FWER. Parametric procedures rely on a full specification of the joint distribution which leads to a power gain compared to nonparametric or semiparametric procedures. These MTPs can be used in multiplicity problems where parametric assumptions are justified. Clinical trials where the test statistics follow a multivariate normal or \( t \) distribution (e.g., trials with multiple dose-control comparisons as in Case study 2) serve as a good example. The most common parametric procedure is the single-step Dunnett procedure (Dunnett, 1955) which assumes that the test statistics follow a multivariate \( t \) distribution. Consider a clinical trial with \( m + 1 \) treatment arms (\( m \) dose groups and a control arm) and assume a balanced design with \( n \) patients per treatment arm. The Dunnett procedure tests the \( m \) null hypotheses corresponding to the \( m \) dose-control contrasts. Let \( t_1, \ldots, t_m \) denote the associated test statistics. The decision rule for the null hypothesis \( H_i \) is based on comparing \( t_i \) to a common critical value. This value is the \((1 - \alpha)\)-quantile of the distribution of the maximum test statistic, i.e., \( \max(T_1, \ldots, T_m) \), where \( T_1, \ldots, T_m \) follow the same distribution as \( t_1, \ldots, t_m \) under the global null hypothesis. Dunnett-adjusted criti-
cal values are smaller than Bonferroni-adjusted critical values and therefore this parametric procedure is more powerful than the Bonferroni procedure. It is to be noted that stepwise versions of the Dunnett procedure exist and can be viewed as parametric versions of the Holm procedure (step-down Dunnett procedure, Dunnett and Tamhane, 1991) and Hochberg procedure (step-up Dunnett procedure, Dunnett and Tamhane, 1992). Both stepwise MTPs are more powerful than the single-step Dunnett procedure.

In some cases, relevant clinical information may be used to pre-specify the testing sequence before the data analysis. *A priori* ordering of the null hypotheses may be defined based on clinical information collected earlier in the development program (e.g., from Phase II trial data) or based on the clinical importance of the trial objectives (e.g., primary and key secondary endpoint(s)). Several methods have been developed that allow the trial’s sponsor to incorporate this information into the testing strategy while controlling the FWER in the strong sense. The most intuitive MTP with a pre-specified testing sequence is the fixed-sequence procedure which uses a straightforward approach to testing a set of ordered null hypotheses \( H_1, \ldots, H_m \). The testing algorithm is built around the sequentially rejective rule, i.e., testing begins with \( H_1 \) at the full \( \alpha \) level and continues until a null hypothesis cannot be rejected at the same full \( \alpha \) level; the untested null hypotheses are accepted. Although this testing approach has found numerous applications in confirmatory clinical trials, the underlying sequentially rejective rule can be too restrictive. It is easy to see that, if a test placed early in the sequence is underpowered, the testing algorithm will likely stop early.

A more flexible approach is supported by the class of nonparametric (Bonferroni-based) chain procedures (Bretz et al, 2009; Millen and Dmitrienko, 2011) that enable complex decision trees. Each chain procedure is uniquely defined by two sets of procedure parameters, initial hypothesis weights and initial transition parameters, defined as follows:

- **Initial hypothesis weights:**
  
  \[ w_1, \ldots, w_m, \]

  where

  \[ w_i \geq 0, \ i = 1, \ldots, m, \]

  \[ \sum_{i=1}^{m} w_i = 1. \]

- **Initial transition matrix:**

  \[
  \begin{pmatrix}
  g_{11} & \cdots & g_{1m} \\
  \cdots & \cdots & \cdots \\
  g_{m1} & \cdots & g_{mm}
  \end{pmatrix}
  \]

  with

  \[ g_{ij} \geq 0, \ i = 1, \ldots, m - 1, \ j = i + 1, \ldots, m, \]

  \[ g_{ij} = 0, \ i = 1, \ldots, m, \ j = 1, \ldots, i, \]

  \[ \sum_{j=1}^{m} g_{ij} = 1, \ i = 1, \ldots, m. \]

These procedure parameters are defined using the \( \alpha \) allocation and \( \alpha \) propagation rules, respectively. Unlike the fixed-sequence procedure which uses a restrictive \( \alpha \) propagation rule, a general chain procedure offers more flexibility by defining an arbitrary set of transition parameters that govern the process of redistributing the error rate among the non-rejected null hypotheses. Specifically, after a null hypothesis is rejected, the hypothesis weights...
and transition parameters are updated using the algorithm described in Bretz et al (2009). Popular MTPs such as the Bonferroni, Holm, fixed-sequence and fallback procedures (Wiens, 2003; Wiens and Dmitrienko, 2005) can be constructed using the general chain approach.

As an illustration, consider the hypothesis-testing problem in Case study 2 with three active doses (Dose L, Dose M and Dose H). Let $H_1$, $H_2$ and $H_3$ denote the null hypotheses associated with the tests of Dose H, Dose M and Dose L versus Placebo. Suppose that, based on the Phase II data, the trial’s sponsor expects the highest efficacy at Dose H and similar efficacy at Doses L and M. It may be risky to apply the fixed-sequence testing approach as it forces the sponsor to test $H_2$ and $H_3$ sequentially and cannot account for the fact that Dose L and Dose M will likely be equally effective. A more relevant approach to multiplicity adjustment in this case will be to set up a chain procedure which first tests $H_1$ and then splits the remaining error rate between $H_2$ and $H_3$. The associated $\alpha$ propagation rule will help the sponsor ensure that the testing strategy is consistent with the logical relationships based on the available clinical information. Examples of chain procedures will be provided in Sections 6, 7 and 8.

It is worth reiterating that the chain procedures introduced above are nonparametric and do not impose any assumptions on the joint distribution of the hypothesis test statistics. When the joint distribution is fully specified, a parametric version of the chain procedures can be constructed (Millen and Dmitrienko, 2011). The class of parametric chain procedures includes as special cases the single-step Dunnett, step-down Dunnett and parametric fallback procedures (Huque and Alosh, 2008). Parametric chain procedures are more powerful than their nonparametric counterparts and should be preferred in clinical trial settings with a known joint distribution of the test statistics.

3.3. Multi-family multiplicity problems

In clinical trials with at least two families of null hypotheses, such as clinical trials with primary and key secondary endpoints or clinical trials evaluating the treatment effect in the overall population as well as pre-specified target subgroups, logical relationships among the null hypotheses can be quite complex. Multiple testing procedures that account for several sources of multiplicity in these increasingly more complex settings need to be considered in this setting. A broad class of gatekeeping procedures has been developed in the literature to allow a trial’s sponsor to control the FWER in the strong sense within and across the families while maximizing power by accounting for the hierarchical structure of the null hypotheses.

Depending on the logical relationships across the null hypotheses, several types of gatekeeping procedure can be used. First of all, chain procedures introduced in Section 3.2 can be applied to certain types of settings with several sources of multiplicity. Further, continuing to MTPs that were specifically developed for multi-family multiplicity problems, it is helpful to differentiate between serial gatekeeping procedures and parallel gatekeeping procedures. These procedures rely on straightforward gatekeeping conditions, i.e., specific types of logical relationships among the null hypotheses in different families (Dmitrienko and Tamhane, 2009). Consider the $k$ families of null hypotheses $(F_1, \ldots, F_k)$ defined in Section 2. A serial gatekeeping procedure allows the testing of the null hypotheses in Family $F_j$ if all hypotheses are rejected in Family $F_{j-1}$, $j = 2, \ldots, k$. This approach can be used in a clinical trial with two co-primary endpoints P1 and P2 (Family $F_1$) and two key secondary endpoints S1 and S2 (Family $F_2$). The serial gatekeeping condition states that, to proceed to the endpoint tests in Family $F_2$, both primary endpoint tests must be significant and the associated null hypotheses of no effect must be rejected. The parallel gatekeeping condition
is less restrictive and supports testing of the null hypotheses in Family $F_j$, $j = 2, \ldots, k$, if at least one hypothesis is rejected in the immediately preceding family. A parallel gatekeeping procedure can be used in a setting with multiple primary and key secondary endpoints where each primary endpoint can lead to a regulatory claim. This means that the secondary endpoint tests can be carried out if at least one primary endpoint test is significant. A general method for setting up parallel gatekeeping procedures was developed in Dmitrienko, Tamhane and Wiens (2008) and Dmitrienko, Kordzakhia and Tamhane (2011).

Serial and parallel gatekeeping procedures are in fact special cases of a more general class of procedures known as mixture-based gatekeeping procedures (Dmitrienko and Tamhane, 2011, 2013). Gatekeeping procedures in this class support very general logical relationships among the null hypotheses in different families. A common setting where mixture-based gatekeeping procedures provide an advantage over the simple gatekeeping procedures defined above is a setting with a hierarchical set of trial endpoints. Consider, for example, a clinical trial which is designed to evaluate the treatment effect of several dose levels on multiple endpoints, i.e., a single primary endpoint and a key secondary endpoint (see Case study 3). As it is not clinically relevant to evaluate a treatment effect on an endpoint if no effect was seen at the same dose on a higher-ranked endpoint, logical restrictions exist within each dose-control comparison sequence. For a specific dose, if the null hypothesis related to one endpoint cannot be rejected, the null hypotheses associated with the endpoints placed later in the sequence are automatically accepted. A class of gatekeeping procedures, termed multiple-sequence gatekeeping procedures, that take into account complex logical relationships of this kind was studied in Dmitrienko, Kordzakhia and Brechenmacher (2015). An example of a multiple-sequence gatekeeping procedure will be provided in Section 8.

It is important to note that gatekeeping procedures control the FWER not only within each family of null hypotheses but also globally, i.e., across the families. Within each family, local FWER control is provided by a pre-specified multiple testing procedure known as a component procedure. Any MTP which controls the FWER in the strong sense, such as the nonparametric, semiparametric or parametric procedures introduced in the Section 3.2, can be used as a component within a gatekeeping procedure. However, the procedures specified in Families $F_1, \ldots, F_{k-1}$ must be separable, meaning that a procedure should allow a positive fraction of the Type I error rate to be transferred to the next family when one or more hypotheses are rejected in the current family. The Holm, Hochberg, Hommel and stepwise Dunnett procedures are not separable but their truncated versions with a pre-specified truncation parameter $\gamma$ ($0 \leq \gamma < 1$) are separable and can be used for building general mixture-based gatekeeping procedures. For more information on truncated MTPs, see Dmitrienko, Tamhane and Wiens (2008).

With this multitude of multiple testing procedures, it is critical to ensure that the selected multiplicity adjustment method is fully aligned with the clinical objectives of a given trial and provides adequate power for each objective. Section 4 introduces a general framework for power evaluation in clinical trials with multiple objectives.

### 4. Power calculations in clinical trials with multiple objectives

To define a general framework for power evaluation in clinical trials with complex clinical objectives, it is instructive to begin with a simple setting involving a single objective, for example, a clinical trial with a single primary endpoint. Using the known distribution of the primary endpoint, it is typically straightforward to derive the power function, i.e., compute the probability of a statistically significant treatment effect as a function of key
parameters such as the total sample size, true treatment difference, variability parameters, randomization ratio and Type I error rate. The power function is evaluated analytically or, alternatively, an approximate analytical expression may be utilized. Let $\psi(n)$ denote the value of power as a function of the total sample size $n$. Given the desirable value of Type II error rate, e.g., $\beta$, the sample size in the clinical trial is computed from $\psi(n) = 1 - \beta$.

The problem of sample size calculations in a trial with a single objective very often admits a closed-form solution.

When this simple setting is extended to a more general setting with multiple clinical objectives, several conceptual and implementation challenges arise. First of all, the focus shifts from a univariate distribution of a single primary test statistic to a multivariate distribution of several test statistics associated with the multiple objectives. With multiple sets of possible trial outcomes, the trial’s sponsor can consider multiple definitions of power or, more generally, multiple success criteria. Commonly used criteria, including exceedance criteria (e.g., disjunctive and conjunctive power) and expectation criteria (e.g., weighted power) are defined below.

The following notation will be used to introduce these success criteria. In a general problem of testing a family of $m$ null hypotheses ($H_1, \ldots, H_m$), let $r_i$ denote an indicator variable with $r_i = 1$ if the null hypothesis $H_i$ is rejected and 0 otherwise. Further, let $\psi_i(n)$ denote the marginal power function of the $i$th test, i.e., the probability to reject the null hypothesis $H_i$ as a function of the sample size. Note that $E(r_i) = P(\text{Reject } H_i) = \psi_i(n)$, $i = 1, \ldots, m$.

Exceedance success criteria are very popular in clinical trials. They are formulated in terms of the probability of exceeding a certain number of rejections in the family of null hypotheses. The most broadly used exceedance criterion is the disjunctive criterion. Its criterion function is equal to the probability of rejecting one or more null hypotheses:

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

A straightforward extension of the regular disjunctive criterion is the subset disjunctive criterion, which is defined as follows:

$$\psi_{SD}(n) = P \left( \sum_{i \in I} r_i \geq 1 \right).$$

Here the sum is computed over a pre-specified index set $I \subset \{1, \ldots, m\}$. In other words, the criterion is based on the probability of at least one rejection in a certain subset of the original family of null hypotheses.

Another popular exceedance criterion is the conjunctive criterion, which is defined in terms of the probability to reject all null hypotheses, i.e.,

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

The subset conjunctive criterion can be derived similarly to the subset disjunctive criterion.

Unlike exceedance criteria, expectation success criteria focus on the expected number of clinically relevant events, e.g., the expected number of rejected null hypotheses. An example is the weighted power criterion, which is defined as a weighed sum of the marginal power functions:
Power Calculations in Clinical Trials

\[
\psi_W(n) = \sum_{i=1}^{m} v_i \psi_i(n),
\]

where \(v_1, \ldots, v_m\) are the importance parameters that quantify the relative importance or value of the null hypotheses within the family. It is assumed that \(v_i \geq 0, i = 1, \ldots, m,\) and \(v_1 + \ldots + v_m = 1.\) When the importance parameters are equal to each other, i.e., \(v_1 = \ldots = v_m = 1/m,\) weighted power is proportional to the expected number of rejected null hypotheses. Indeed,

\[
mp\psi_W(n) = \sum_{i=1}^{m} \psi_i(n) = \sum_{i=1}^{m} P(\text{Reject } H_i) = E \left( \sum_{i=1}^{m} r_i \right).
\]

With unequal importance parameters, one or more null hypotheses may have more impact on the criterion than the others. For example, consider the problem of testing three null hypotheses and let \(v_1 = 1/2\) and \(v_2 = v_3 = 1/4.\) In this case, \(4\psi_W(n) = E(2r_1 + r_2 + r_3)\) and thus the null hypothesis \(H_1\) is “counted twice” when the expected number of rejections is computed. In other words, a rejection of the null hypothesis \(H_1\) is as important as a simultaneous rejection of the other two null hypotheses.

When the overall success in a clinical trial is defined in terms of “winning” on at least one clinical objective, i.e., rejecting one or more null hypothesis in the prospectively defined family, the disjunctive criterion is arguably the most commonly used success criterion. However, when selecting a criterion to be used in sample size calculations, it is important to ensure that it is consistent with the trial’s clinical objectives. As pointed out in several publications (see, for example, Millen and Dmitrienko, 2011), a potential limitation of the disjunctive and related success criteria is that they are symmetric in the sense that the null hypotheses are treated as being interchangeable. However, clinical objectives and associated null hypotheses are not interchangeable in multiple settings. For example, in a clinical trial with two patient populations (overall patient population and target subpopulation), the rejection of the null hypothesis of no overall effect may be more desirable compared to the null hypothesis of no effect in the subpopulation. In this case, the two null hypotheses are not interchangeable and “asymmetric” success criteria such as the weighted power criterion are more relevant than popular symmetric criteria.

From the implementation perspective, the trial’s sponsor needs to gather additional information to fully specify all parameters of the underlying multivariate distribution. In addition to the parameters defined in the univariate setting, for example, the standardized treatment differences (effect sizes) for the individual tests, pairwise correlations need to be specified. A more important implementation consideration is related to the fact that analytical expressions for popular success criteria are rarely available in a multivariate setting. One of the few exceptions is the case of basic single-step multiple testing procedures, including the Bonferroni procedure (see Section 3.2). The key property of single-step multiple testing procedures is that each test is carried out at a fixed significance level and thus it is easy to evaluate the success criteria defined above.

As an illustration, consider a clinical trial with two primary null hypotheses and assume that the test statistics associated with the two hypotheses (denoted by \(Z_1\) and \(Z_2\)) follow a bivariate normal distribution. Under this assumption, disjunctive power for the Bonferroni procedure can be computed analytically and the sample size can be found from the following equation:

\[
P(Z_1 \geq z_{\alpha/2} \text{ or } Z_2 \geq z_{\alpha/2}) = 1 - \beta.
\]
This probability is evaluated under the alternative hypothesis of a beneficial treatment effect and depends on the assumed treatment effects as well as the correlation between the test statistics. A similar approach can be used with other success criteria, including conjunctive and weighted power.

When more complex multiple testing procedures are considered, the test-specific significance levels are no longer fixed. As explained in Sections 3.2 and 3.3, stepwise procedures such as chain procedures are built around the concept of $\alpha$ propagation, which means that the significance levels of the individual tests are updated after each rejection. As a result, it becomes quite challenging to extend the analytical approach outlined above to powerful stepwise procedures. Simulation-based methods are commonly employed to perform power evaluations for the efficient multiple testing procedures defined in Section 3. A general framework for comprehensive simulation-based evaluation in drug development, which is directly applicable to clinical trials with multiple objectives, is presented in Section 5.

5. Clinical scenario evaluation approach

An important limitation of a standard sample-size computation approach is that its focus is quite narrow. Sample size calculations in clinical trials generally pursue a simple goal of computing a quick estimate of the required number of patients and often force the trial’s sponsor to make a variety of simplifying assumptions. In particular, very few options for the trial design or analysis method are typically explored and limited attention is paid to the selection of a success criterion. At a more fundamental level, closed-form solutions are normally preferred, which imposes additional restrictions on important design elements such as patient accrual and dropout patterns.

A desirable alternative to standard approaches to sample size determination in clinical trials with a basic “one-dimensional” goal will be a broad “multi-dimensional” approach to a quantitative assessment of the operating characteristics of several candidate analysis methods under multiple candidate trial designs to arrive at a solution which is consistent with the trial’s clinical objectives and maximizes a relevant success criterion or utility function (Dmitrienko and D’Agostino, 2013). These general goals can be accomplished within the Clinical Scenario Evaluation (CSE) framework (Benda et al, 2010; Friede et al, 2010). Conceptually similar approaches to quantitative evaluation of multiple analysis strategies were also proposed in Dmitrienko et al (2011a) and Millen and Dmitrienko (2011).

The CSE approach recognizes that sample size calculation and power evaluation in clinical trials are inherently high-dimensional statistical problems. This approach helps decompose a complex problem of examining an overwhelming number of options by identifying the main components of the evaluation process. These components are termed models, i.e., data models (assumptions), analysis models (options) and evaluation models (metrics), as defined below.

- Data models define the process of generating trial data (e.g., sample sizes, outcome distributions).
- Analysis models define the statistical methods applied to the trial data (e.g., statistical tests, multiplicity adjustments).
- Evaluation models specify the measures for evaluating the performance of the analysis strategies (success criteria such as disjunctive or weighted power).
The general CSE approach enables a clinical trial sponsor to transition from basic sample size calculations to clinical trial optimization. Clinical trial optimization can be understood as a comprehensive approach to quantifying the impact of applicable design scenarios and analysis strategies on the selected evaluation criteria and identifying the parameter configurations that lead to optimal performance. It is important to note that the success criteria are not generally uniquely defined and multiple criteria may need to be utilized to help “quantify” the clinical objectives of a particular trial.

The optimization exercise should be performed in parallel with sensitivity assessments to understand the effect of design and/or analysis model parameters on the overall conclusions. It is imperative to ensure that the conclusions are robust to reasonable deviations from the underlying statistical assumptions. Sensitivity analyses play a key role in a traditional clinical trial setting with a single clinical objective and are even more relevant in trials with multiple objectives due to a complex interplay of individual design and analysis parameters.

Key principles of the CSE approach, including optimal selection of relevant design and analysis parameters as well as sensitivity assessments, are illustrated in Sections 6, 7 and 8 using the three case studies introduced in Section 2. Section 6 uses a clinical trial example with several endpoints to demonstrate the importance of evaluating power for all relevant clinical objectives, including the primary and secondary endpoints. An additional important aspect of the CSE framework presented in this section is the evaluation of appropriate success criteria for an entire development program. Further, a clinical trial example with multiple dose-control tests is used in Section 7 to discuss optimal selection of target parameters of the selected analysis method (multiplicity adjustment). Lastly, Section 8 also focuses on clinical trial optimization and compares the performance of two candidate analysis methods (multiple testing procedures) in the context of a trial with two families of clinical objectives (evaluation of the primary and key secondary endpoints at two dose levels).

An R package (Mediana package) was developed to provide a general software implementation of the CSE approach. The current version of the Mediana package supports a broad set of clinical trial designs and analysis methods. A variety of other features will be added in the future, including adaptive trial designs, support to Bayesian methods, etc. For more information on the Mediana package, see http://biopharmnet.com/mediana

The clinical trial simulations presented in Sections 6, 7 and 8 were performed using the Mediana package based on 10,000 simulation runs.

6. Case study 1: Multiple endpoints

This section provides examples of power calculations in clinical trials where multiple endpoints are utilized to characterize the efficacy of a new treatment. While standard sample size calculations focus on a single evaluation model (power of the primary test), it is advisable to explore all relevant evaluation models, e.g., evaluate power of the key secondary endpoint tests as well. The sample size may be increased to ensure that critically important secondary analyses are adequately powered.

A trial’s sponsor may consider other evaluation models such as power based on the overall probability of success across a Phase III development program. Most commonly, two or more clinical trials are included in a development program to assess the safety and effectiveness of a new treatment. It is often sensible and even desirable to compute the probability of success at the program level by applying appropriate success criteria across...
the trials (Wiens, 2011). The CSE approach can be applied to pursue two related goals in this example:

- Goal 1. Power evaluation within each trial.
- Goal 2. Power evaluation across the development program.

6.1. Clinical trial example

The main considerations in the multiple-endpoint setting will be illustrated using a development program in systemic lupus erythematosus (SLE). Two confirmatory clinical trials will be conducted within this program (Trials A and B) to evaluate the effects of a novel treatment on signs and symptoms of SLE. Within each trial, a single dose of the treatment will be tested versus a placebo. The efficacy evaluation will be performed based on the primary endpoint (Endpoint E1) and two key secondary endpoints (Endpoints E2 and E3). The endpoints are defined as follows:

- Endpoint E1: Proportion of patients who achieve symptomatic improvement as measured by a SLE responder index.
- Endpoint E2: Change from baseline in the physician global assessment score.
- Endpoint E3: Change from baseline in the SLEDAI (Systemic Lupus Erythematosus Disease Activity Index) score.

The primary endpoint is a dichotomous endpoint and the key secondary endpoints are normally distributed.

The key secondary endpoints will be analyzed after the primary trial’s objective is met, i.e., the treatment is demonstrated to provide a statistically significant improvement over placebo based on Endpoint E1. Therefore, it is natural to put Endpoint E1 in the first family. The key secondary endpoints are considered equally important in terms of the associated regulatory claims and will be placed in the second family:

- Family $F_1$ (primary family): Endpoint E1 (corresponding null hypothesis of no treatment effect is denoted by $H_1$).
- Family $F_2$ (secondary family): Endpoints E2 and E3 (corresponding null hypotheses are denoted by $H_2$ and $H_3$).

The resulting multiplicity problem serves as an example of a two-dimensional multiplicity problem. The two families define the first source of multiplicity. Secondly, two tests will be carried out within the secondary family, which induces additional multiplicity.

6.2. Multiplicity adjustment

As was indicated in Section 3.3, a common approach to addressing multiplicity in trials with complex clinical objectives is to construct a gatekeeping procedure which accounts for the hierarchical structure of the individual objectives. In this case, it is most appropriate to set up a two-stage gatekeeping procedure which first examines the primary family and, if the primary objective is met, proceeds to the secondary family. A single test is included in Family $F_1$ and thus no multiplicity adjustment is required within this family. Further, any MTP which ensures local FWER control can be used within Family $F_2$. Since the null hypotheses within this family are equally important, the Holm, Hochberg or Hommel procedures will
be preferable to the fixed-sequence procedure. To improve the probability of success in the secondary family, it would have been ideal to apply the more powerful Hochberg or Hommel procedures in Family $F_2$ (note that the Hochberg and Hommel procedures are equivalent to each other in this case since there are two hypotheses in Family $F_2$). However, as pointed in Section 3.2, the Hommel procedure strongly controls the FWER only if the endpoint test statistics follow a multivariate normal distribution with non-negative correlations. It is reasonable to assume a bivariate normal distribution; however, the correlation coefficient is typically unknown. As a consequence, the sponsor will have to rule out semiparametric as well as parametric procedures and resort to a Bonferroni-based procedure, e.g., the Holm procedure, in Family $F_2$.

The resulting Holm-based gatekeeping procedure serves as an important component of the analysis model in this clinical trial example and is based on the following algorithm. Let $p_i$ denote the $p$-value computed from an appropriate statistical test for the hypothesis $H_i$, $i = 1, 2, 3$.

- **Family $F_1$.** $H_1$ is rejected if $p_1 \leq \alpha$.

- **Family $F_2$.** The null hypotheses $H_2$ and $H_3$ are tested using the Holm procedure at the full $\alpha$ level if $H_1$ is rejected. Let $p_{(2)} < p_{(3)}$ denote the ordered $p$-values in the secondary family and, similarly, let $H_{(2)}$ and $H_{(3)}$ denote the ordered null hypotheses. The null hypothesis $H_{(2)}$ is rejected if $p_{(2)} \leq \alpha/2$. Further, the null hypothesis $H_{(3)}$ is rejected if $H_{(2)}$ is rejected and $p_{(3)} \leq \alpha$.

It is worth noting that, in this simple clinical trial example, the Holm-based gatekeeping procedure is, in fact, equivalent to a nonparametric (Bonferroni-based) chain procedure. The $\alpha$ allocation and $\alpha$ propagation rules used in this chain procedure are displayed in Figure 1. As explained in Section 3.2, the $\alpha$ allocation rule defines the initial weights of the null hypotheses. Since the primary family serves as a gatekeeper for the secondary family, the weight of $H_1$ is set to 1 and the weights of $H_2$ and $H_3$ are set to 0, i.e., $w_1 = 1$ and $w_2 = w_3 = 0$. The $\alpha$ propagation rule defines the process of transferring specific fractions of the overall error rate from a rejected null hypothesis to the remaining non-rejected null hypotheses. In this case, the error rate released after the rejection of $H_1$ is split equally between the secondary null hypotheses ($g_{12} = g_{13} = 0.5$) and, after a secondary

![Diagram](image)

**Fig. 1:** Visual representation of the $\alpha$ allocation and $\alpha$ propagation rules used in the non-parametric chain procedure (Case study 1).
null hypothesis is rejected, the available error rate is carried over to the other secondary null hypothesis \((g_{23} = g_{32} = 1)\).

### 6.3. Clinical scenario evaluation

Assessment of the operating characteristics of the Holm-based gatekeeping procedure within each Phase III trial can be performed using the general principles presented in Section 5. The trial’s sponsor can first evaluate the marginal power functions of the three endpoint tests and select the sample size which guarantees a desirable probability of detecting a significant effect on the primary endpoint as well as the two secondary endpoints. Considering other success criteria, it should be noted that the disjunctive power function for the set of three endpoints is simply equal to the marginal power function of the primary endpoint since Family \(F_1\) serves as gatekeeper for Family \(F_2\). The subset disjunctive criterion applied to the secondary family, i.e., the probability of rejecting at least one null hypothesis in \(F_2\), may be more informative.

It is also recommended to evaluate the performance of the selected analysis model (Holm-based gatekeeping procedure) across the two clinical trials. This evaluation will provide information on the overall chances of meeting the overall goals of the development program. Based on current regulatory requirements, a strict consistency criterion is applied to primary endpoints: a beneficial effect must be demonstrated in each trial within a program. Analysis of the secondary endpoints can be performed using the same consistency criterion or, alternatively, a pooled analysis can be considered. Wang et al. (2013) argued that the former approach may be overly conservative. As stated in Wang et al. (2013), the advantages of the pooled analysis include improved power of the secondary analyses and reduced probability of observing inconsistent results across the trials.

To illustrate the process of applying the CSE approach in this clinical trial example, simulations were performed under the following set of statistical assumptions that defined the data model. Common improvement rates for the primary endpoint (Endpoint E1) in the placebo and treatment arms were assumed in Trials A and B, i.e., 40% of patients were expected to achieve symptomatic improvement in the placebo arm and 55% in the treatment arm \((p_1 = 0.40\) and \(p_2 = 0.55\)). This translated into the following standardized treatment effect (effect size):

\[
\frac{p_2 - p_1}{\sqrt{\hat{p}(1 - \hat{p})}} = 0.3,
\]

where \(\hat{p} = (p_2 + p_1)/2\). Further, the effect sizes for Endpoints E2 and E3 were assumed to be 0.24 and 0.20, respectively, in both trials. Each effect size was defined as the mean treatment difference divided by the common standard deviation. Note that a simplified setting with a single data model was considered in this case study. Multiple assumption sets should be explored in real clinical trial applications.

A copula-based approach was used for modeling the joint distribution of the three endpoints, i.e., a latent trivariate normal distribution was assumed and the first component was converted to a dichotomous variable to represent the primary endpoint. For simplicity, a common correlation coefficient was assumed in the trivariate distribution \(\rho\) and was initially set to 0.3.

Several sets of clinical trial simulations were performed to study the relevant operating characteristics. The first set of simulations focused on Goal 1, i.e., on determining the sample size within a single trial (Trial A or Trial B). To help provide a balanced view of the probability of winning on the primary endpoint as well as probability of establishing...
significant treatment effects for the key secondary endpoints, the following power functions were evaluated over a broad range of sample sizes:

- Marginal power in Family $F_1$ (probability of rejecting $H_1$).
- Disjunctive power in Family $F_2$ (probability of rejecting $H_2$ or $H_3$).

Table 2 presents the marginal power and disjunctive power functions for $n = 240$, $260$ and $280$ patients per treatment arm. Based on the marginal power function, the probability of meeting the trial’s primary objective was very high (it was over 90% with $n = 240$) and thus the sponsor could consider smaller sample sizes. This argument focuses on the primary endpoint and it is recommended to also examine the probability of achieving success in the secondary family. Focusing on $n = 240$ per arm, the secondary endpoint tests were underpowered; specifically, the probabilities of rejecting $H_2$ and $H_3$ were 61.1% and 48.4%, respectively (these values are not displayed in Table 2). Switching to disjunctive power helped improve the probability of success; however, the disjunctive power in Family $F_2$ was still below the 80% threshold. If it is critical to ensure a reasonably high probability of detecting significant treatment effects in the secondary family, the sample size per arm could be increased to 260 or even 280 patients.

The next set of simulations dealt with Goal 2, i.e., the evaluation of probabilities of success at the program level. Two different approaches were considered:

- Compute the probability of a statistically significant effect simultaneously in both trials.
- Apply appropriate success criteria based on the pooled data from the two trials.

As explained above, the first approach is mandated in the primary analysis and it is advisable to use the second approach in the analysis of key secondary endpoints.

Beginning with the primary analysis, Table 2 displays simultaneous marginal power in Family $F_1$, i.e., the probability of meeting the primary objective in both trials. It follows from the table that the success probability at the program level was quite high (it exceeded 80%).

Simulation results related to the performance of the secondary endpoint tests across the trials are also presented in Table 2. This table provides information on the following power functions:

| Criterion                          | Sample size per treatment arm |
|------------------------------------|-------------------------------|
|                                    | $n = 240$ | $n = 260$ | $n = 280$ |
| Primary endpoint test (Family $F_1$) |          |           |           |
| Marginal power (%)                 | 90.7      | 92.6      | 94.1      |
| Simultaneous marginal power (%)    | 82.8      | 86.2      | 89.0      |
| Secondary endpoint tests (Family $F_2$) |          |           |           |
| Disjunctive power (%)              | 68.7      | 73.6      | 77.7      |
| Simultaneous disjunctive power (%) | 48.0      | 55.1      | 61.5      |
| Pooled disjunctive power (%)       | 97.0      | 98.2      | 98.8      |
| Pooled conjunctive power (%)       | 83.8      | 87.3      | 89.9      |
• Simultaneous disjunctive power in Family $F_2$ (probability of rejecting $H_2$ or $H_3$ simultaneously in Trial A and Trial B).

• Pooled disjunctive power in Family $F_2$ (probability of rejecting $H_2$ or $H_3$ using the pooled data from Trial A and Trial B).

• Pooled conjunctive power in Family $F_2$ (probability of rejecting $H_2$ and $H_3$ using the pooled data from Trial A and Trial B).

Table 2 shows that simultaneous disjunctive power in Family $F_2$ ranged between 48% and 62% and thus the probability of winning on at least one secondary endpoint simultaneously in both trials was clearly very low. By contrast, the pooled analysis approach resulted in a much higher probability of success. The probability of establishing a beneficial treatment effect on one secondary endpoint or both secondary endpoints (pooled disjunctive power) was greater than 90% over the sample size range. It is also helpful to note that, when a pooled analysis of the two secondary endpoints was performed, the program’s sponsor could expect a high probability of establishing a significant effect simultaneously on both secondary endpoints (pooled conjunctive power). This probability was greater than 80% with $n = 240$ per arm and very close to 90% with $n = 280$ per arm.

6.4. Sensitivity assessments

As indicated in Section 5, sensitivity assessments play a central role in the CSE framework. To give an example of implementing sensitivity analyses, power functions in the secondary family depend on the key data model parameters (effect sizes for Endpoints E2 and E3) as well as additional parameters such the correlations among the three endpoints. The available information on plausible values of correlation coefficients is limited in most real-life situations. Given this, it is recommended to evaluate the marginal power functions of the two secondary tests and the disjunctive power function in Family $F_2$ over the entire range of the correlation coefficients.

For the sake of illustration, a common correlation coefficient ($\rho$) was assumed in the underlying trivariate normal distribution. Since correlations in clinical trial applications are typically positive and rarely exceed 0.6 (Sankoh, D’Agostino and Huque, 2003), $\rho$ ranged between 0 and 0.6 in this example. Additional simulations were run and demonstrated that choosing extreme values of the common correlation coefficient had a relatively small impact on disjunctive power in the secondary family. For example, with $n = 280$ patients per arm, the disjunctive power decreased from 82.0% at $\rho = 0$ to 77.7% at $\rho = 0.3$ and lastly to 73.1% at $\rho = 0.6$. This simple sensitivity assessment helps increase the general confidence level by showing that the performance of the selected gatekeeping procedure in the secondary family was fairly robust to the somewhat arbitrary choice of the correlation coefficients.

7. Case study 2: Multiple dose-control comparisons

This section focuses on the following goals that illustrate important aspects of the general CSE approach:

• Goal 1. Overall power evaluation and sample size determination.

• Goal 2. Optimal selection of the target parameters.
7.1. Clinical trial example and multiplicity adjustment

To illustrate the key ideas, consider a clinical trial with multiple dose-control comparisons (Case study 2). In this setting, multiplicity is induced by the three dose-placebo comparisons: Dose L versus placebo, Dose M versus placebo, Dose H versus placebo. Let \( H_3, H_2 \) and \( H_1 \) denote the corresponding null hypotheses of no effect.

An analysis model can be formulated in terms of any of the MTPs introduced in Section 3.2, including nonparametric, semiparametric and fully parametric procedures, can be applied to handle error rate inflation in this problem. For the sake of concreteness, we will focus on nonparametric (Bonferroni-based) chain procedures and discuss optimal selection of the target parameters, i.e., the hypothesis weights and transition parameters used in this multiplicity adjustment.

As a specific example, we will use a Phase III trial for the treatment of patients with major depressive disorder. The primary endpoint is the change from baseline to the end of the treatment phase in the Montgomery-Asberg Depression Rating Scale (MADRS) total score. This endpoint is assumed to follow a normal distribution.

The largest standardized treatment difference (effect size) in this trial is expected at Dose H and the treatment is anticipated to be less effective at Doses L and M. With these assumptions, it is reasonable to consider a chain procedure defined in Figure 2. This chain procedure uses the decision rules that are conceptually similar to those used in Case study 1 (see Figure 1). Specifically, testing begins with \( H_1 \) and, after this null hypothesis is rejected, the associated error rate will be split between \( H_2 \) and \( H_3 \). The key difference, however, is that the parameters of the chain procedure in this case study are no longer fixed. The hypothesis weights and transition parameters will serve as the target parameters that need to be optimally selected using available clinical information. As shown in Figure 2, \( 0 < w \leq 1 \) is the weight of the hypothesis \( H_1 \), i.e., \( w_1 = w \). The remaining weight is split between the other hypotheses: \( w_2 = w_3 = (1 - w)/2 \). Further, \( 0 \leq g \leq 1 \) is the transition parameter which determines the fraction of the error rate used for testing \( H_1 \) to be transferred to the hypothesis \( H_2 \) after \( H_1 \) is rejected. The remaining fraction will be carried over to \( H_3 \). In other words, \( g_{12} = g \) and \( g_{13} = 1 - g \).

While it may be considered "natural" to allocate the same weights to the three null hypotheses (i.e., let \( w = 1/3 \)), it is important to remember that assigning equal weights to the null hypotheses of interest must be viewed as the sponsor’s choice rather than a default.

Fig. 2: Visual representation of the \( \alpha \) allocation and \( \alpha \) propagation rules used in the nonparametric chain procedure (Case study 2).
(Wiens, Dmitrienko and Marchenko, 2013). It will be in the sponsor’s best interest to explore multiple sets of hypothesis weights or, ideally, explicitly search for the set of weights which optimizes an appropriate success criterion. The same approach can be applied to the more general problem of dealing with several sets of target parameters, including transition parameters. In this case, a multivariate grid search can be employed to identify an optimal set of procedure parameters (Millen and Dmitrienko, 2011) as part of the CSE approach.

7.2. Optimality criteria for procedure parameter selection

As indicated above, an important goal considered in this section is optimal selection of the target parameters in the analysis model, i.e., the \( w \) and \( g \) parameters of the chain procedure. A critical consideration in any optimization problem is the choice of the optimality criterion. The trial’s overall objective is formulated in terms of demonstrating that one or more doses of the new treatment are effective. The associated success criterion is the disjunctive criterion, i.e.,

\[
\psi_D(w, g) = P(\text{Reject } H_1 \text{ or } H_2 \text{ or } H_3).
\]

While it is reasonable to use disjunctive power to perform general power evaluations in the trial and determine the sample size, the disjunctive criterion may not be suitable for parameter selection. Recall that, with larger values of \( w \), the initial weight of the hypothesis \( H_1 \) is high. As a result, \( H_1 \) essentially serves as a gatekeeper for the other two hypotheses, i.e., the rejection of at least one hypothesis is virtually equivalent to the rejection of \( H_1 \). The general behavior of the disjunctive power function will be very close to that of the marginal power function for \( H_1 \) but the marginal power functions for \( H_2 \) or \( H_3 \) will have little influence on the overall criterion. This implies that the optimality criterion based on disjunctive power will be driven mostly by \( w \) and virtually unaffected by the choice of \( g \).

An alternative approach is to consider an optimality criterion based on the weighted power function which is defined as a weighted sum of the marginal power functions. The weighted power criterion in this example is given by

\[
\psi_W(w, g) = v_1 P(\text{Reject } H_1) + v_2 P(\text{Reject } H_2) + v_3 P(\text{Reject } H_3),
\]

where \( v_1, v_2 \) and \( v_3 \) are prospectively defined positive coefficients that quantify the relative importance of establishing a significant treatment effect at each dose \( (v_1 + v_2 + v_3 = 1) \). Unlike the disjunctive criterion, this particular criterion accounts for the probability of rejecting each individual hypothesis and thus serves as a more reliable tool for assessing the impact of the two target parameters \( (w \text{ and } g) \) on the overall probability of success.

7.3. Clinical scenario evaluation

In what follows, we provide a summary of the simulation study which was carried out to address the goals stated at the beginning of this section, i.e., Goal 1 (general power evaluation) and Goal 2 (optimal selection of the target parameters).

Clinical trial simulations were run under the dose-response scenarios (data models) presented in Table 3. For simplicity, we focused on only two data models. The assumed standardized treatment differences (effect sizes) for the three dose-placebo comparisons were denoted by \( \theta_1, \theta_2 \) and \( \theta_3 \), respectively. Under Scenario 1, Doses L and M were assumed to be equally effective and Scenario 2 assumed a linear dose-response relationship.

As a starting point related to Goal 1, it is helpful to determine a general range of sample sizes in this clinical trial example. Disjunctive power was used in the initial set of calculations and, with \( n = 250 \) patients per treatment arm, the minimum disjunctive power
Table 3: Assumed dose-response scenarios in Case study 2.

| Scenario     | Dose L versus Placebo | Dose M versus Placebo | Dose H versus Placebo |
|--------------|------------------------|------------------------|------------------------|
| Scenario 1   | $\theta_3 = 0.20$      | $\theta_2 = 0.20$      | $\theta_1 = 0.30$      |
| Scenario 2   | $\theta_3 = 0.15$      | $\theta_2 = 0.25$      | $\theta_1 = 0.30$      |

over all combinations of the $w$ and $g$ parameters was 81.7% for Scenario 1 and 84.9% for Scenario 2. Since this lowest power level may be much less than the power level associated with an optimal configuration of the procedure parameters, this approach to selecting a starting value of the sample size is likely to be overly conservative. With a smaller sample size ($n = 200$), disjunctive power ranged from 71.7% to 84.7% under Scenario 1 and from 75.5% to 84.5% under Scenario 2.

Using the selected sample size ($n = 200$ patients per treatment arm), additional simulations were performed to address Goal 2, i.e., to identify an optimal combination of the $w$ and $g$ parameters. The evaluation was based on the weighted power function $\psi_W(w,g)$ defined above. The three null hypotheses were assumed to be equally important, i.e., $v_1 = v_2 = v_3 = 1/3$. As noted in Section 4, it is easy to interpret the value of weighted power in this case: $3\psi_W(w,g)$ is equal to the expected number of rejected null hypotheses. Figure 3 presents a visual summary of the simulation results, namely, a “heat map”, under Scenario 1 (top panel) and Scenario 2 (bottom panel). Each cell in the heat map shows the expected number of rejected null hypotheses as a function of $w = 0.1, \ldots, 1$ and $g = 0, 0.1, \ldots, 1$.

Considering the first dose-response scenario in the top panel of Figure 3, the optimal number of rejected null hypotheses was 1.69, which corresponds to the weighted power criterion of 56.5%. The expected number of rejections was maximized for several combinations of the $w$ and $g$ parameters, e.g., for $w = 1$ and $g$ ranging between 0.2 and 0.8. For the sake of example, one can select $w = 1$ and $g = 0.5$. This particular set of target parameters defines a chain procedure which assigns all of the weight to the null hypothesis $H_1$ (Dose H versus placebo), i.e.,

$$w_1 = 1, \ w_2 = 0, \ w_3 = 0.$$ 

After $H_1$ is rejected, the procedure splits the error rate equally between the other two null hypotheses, i.e.,

$$g_{12} = 0.5, \ g_{13} = 0.5, \ g_{23} = 1, \ g_{32} = 1.$$ 

This optimal testing strategy looks quite reasonable given that the assumed effect size for the Dose M-placebo test ($\theta_2$) is equal to that for the Dose L-placebo test ($\theta_3$) under Scenario 1. Further, as shown in the bottom panel of Figure 3, the largest expected number of rejections (1.77) was achieved with $w = 1$ and $g = 1$ under Scenario 2. The associated chain procedure is the fixed-sequence procedure with

$$w_1 = 1, \ w_2 = 0, \ w_3 = 0$$

and

$$g_{12} = 1, \ g_{13} = 0, \ g_{23} = 1, \ g_{32} = 1.$$
Fig. 3: Heat maps of the expected number of rejected null hypotheses as a function of the $w$ and $g$ parameters in the chain procedure under Scenario 1 (top panel) and Scenario 2 (bottom panel) with 200 patients per treatment arm in Case study 2. The expected number of rejected null hypotheses is displayed within each cell.

This chain procedure allocates all of the weight to the null hypothesis $H_1$ and transfers this weight to $H_2$ after $H_1$ is rejected. If $H_2$ is also rejected, the weight is carried over to $H_3$. To see why this testing strategy maximizes the expected number of rejections, note that $\theta_2$
is much greater than $\theta_3$ and therefore assigning a positive weight to $H_3$ after rejecting $H_1$ will result in a suboptimal approach.

Given the optimal set of target parameters listed above, it is also useful to identify

---

Fig. 4: Heat maps of the nearly optimal combinations of the $w$ and $g$ parameters in the chain procedure under Scenario 1 (top panel) and Scenario 2 (bottom panel) with 200 patients per treatment arm in Case study 2. A filled cell identifies a nearly optimal parameter combination.
parameter sets that lead to nearly optimal performance. The nearly optimal performance can be defined in terms of the largest reduction in the optimality criterion which is not considered practically important. To give a quick example, a 5% reduction on a relative scale may be viewed as unimportant for practical purposes. For instance, the largest expected number of rejected hypothesis is 1.69 under Scenario 1 and, applying a 5% rule, the trial’s sponsor can treat any combination of the target parameters with the expected number of rejections above 1.61 as being nearly optimal. These combinations of the $w$ and $g$ parameters are identified in Figure 4 under Scenario 1 (top panel) and Scenario 2 (bottom panel) using filled cells. It follows from the figure that chain procedures with $w \geq 0.5$ and any value of the $g$ parameter were not much different from the optimal procedure under Scenario 1. The bottom panel of Figure 4 indicates that most of the parameter combinations in the upper right corner of the heat map resulted in nearly optimal chain procedures under Scenario 2.

The discussion presented above focused on only one set of sample sizes ($n = 200$ per treatment arm) and it will be informative for the trial’s sponsor to assess the effect of the sample size on the optimal values of the $w$ and $g$ parameters of the chain procedure. As an illustration, Table 4 presents a summary of additional simulations with other sample size sets and demonstrates that the sample size has no impact on the optimal combination of the procedure parameters. Under Scenario 1, the expected number of rejected hypotheses was maximized for several different combinations of the target parameters; however, the performance of the chain procedure with $w = 1$ and $g = 0.5$ was similar to the performance of the optimal chain procedure for each sample size set ($n = 220$ and $n = 240$). Under Scenario 2, the optimal combinations of the target parameters were identical across the sample size sets. Of note, as expected, an increase in the sample size had a positive impact on all evaluation criteria (disjunctive power, conjunctive power and expected number of rejected hypotheses).

Table 4: Summary of optimal parameter selection in Case study 2.

| Criterion                                      | Sample size per treatment arm |  |
|------------------------------------------------|-------------------------------|--|
|                                                | $n = 200$                      | $n = 220$ | $n = 240$ |
| **Scenario 1**                                 |                               |           |           |
| Minimum disjunctive power over all combinations of target parameters (%) | 71.7                          | 76.0      | 80.2      |
| Optimal set of target parameters $w = 1, g = 0.3$ | $w = 1, g = 0.6$              | $w = 1, g = 0.4$ |
| Expected number of rejected hypotheses with the optimal set of target parameters | 1.69                          | 1.82      | 1.94      |
| Expected number of rejected hypotheses with $w = 1, g = 0.5$ | 1.69                          | 1.82      | 1.94      |
| Disjunctive power with the optimal set of target parameters (%) | 84.7                          | 87.8      | 90.7      |
| **Scenario 2**                                 |                               |           |           |
| Minimum disjunctive power over all combinations of target parameters (%) | 75.5                          | 80.1      | 83.8      |
| Optimal set of target parameters $w = 1, g = 1$ | $w = 1, g = 1$                | $w = 1, g = 1$ |
| Expected number of rejected hypotheses with the optimal set of target parameters | 1.77                          | 1.89      | 1.98      |
| Disjunctive power with the optimal set of target parameters (%) | 84.5                          | 88.0      | 90.8      |
7.4. Sensitivity assessments

As in Case study 1 (Section 6), it is important to assess the sensitivity of the conclusions (optimal values of the \( w \) and \( g \) parameters in the chain procedure) to the assumptions made in the simulation study, e.g., the shape of the dose-response relationship in this dose-finding trial. Consider, for example, the optimal chain procedure in Scenario 1. Since this chain procedure was optimized for a particular dose-response scenario, it is reasonable to hypothesize that the procedure would no longer perform as well if the true dose-response relationship were different from the assumed relationship. For instance, a chain procedure with a different set of the target parameters would likely be optimal if the dose-response function were flat or umbrella-shaped. However, it is unclear how quickly the performance of the optimal chain procedure in Scenario 1 will begin to deteriorate if the true dose-response function starts deviating from the assumed dose-response function. A sensitivity assessment is designed to answer this question and quantify the influence of the “true” underlying dose-response function on the performance of an optimal chain procedure in this setting.

Recall that the sensitivity assessment setting in Section 6 was performed in a very straightforward manner since it focused on a single parameter. The effect of the assumed value of the common correlation coefficient \( \rho \) on the performance of the selected analysis method (multiple testing procedure) was examined by computing an appropriate evaluation criterion (disjunctive power) under the extreme values of \( \rho \). This simple approach can not be directly applied in this case study since it is not clear what the “extreme values” of a dose-response function are.

An alternative approach was adopted to conduct a sensitivity assessment in this clinical trial example. The performance of the optimal chain procedure under a particular dose-response function was evaluated by “randomly perturbing” the assumed dose-response relationship and computing the weighted power criterion (or, equivalently, the expected number of rejected null hypotheses) for each new set of effect sizes for the dose-placebo contrasts. To define an algorithm used in this assessment, consider, for example, the dose-response function under Scenario 1. The corresponding effect sizes are given by \( \theta_1 = 0.3 \), \( \theta_2 = 0.2 \) and \( \theta_3 = 0.2 \) and therefore this dose-response function can be uniquely identified by the effect size vector

\[
\theta = (0.3, 0.2, 0.2).
\]

Using this notation, sensitivity assessment was performed as follows:

- Beginning with the assumed dose-response function, which corresponded to the effect size vector

\[
\theta = (\theta_1, \theta_2, \theta_3) = (0.3, 0.2, 0.2),
\]

multiple sets of effect sizes that defined the “true” underlying dose-response functions were generated. Each dose-response function was identified using the effect size vector

\[
\theta_i^* = (\theta_{1i}^*, \theta_{2i}^*, \theta_{3i}^*),
\]

where \( i = 1, \ldots, k \) and \( k \) was a sufficiently large number, say, \( k = 50 \) or \( 100 \). Here \( \theta_{ij}^* \) defined the mean treatment effect at the \( j \)th dose in the \( i \)th effect size set and followed a normal distribution with the mean \( \theta_j \) and standard deviation \( \sigma, j = 1, 2, 3 \). The value of \( \sigma \) quantified the amount of uncertainty or amount of deviation from the assumed dose-response function.
• Considering each “true” dose-response function, 10,000 sets of treatment effect \( p \)-values were generated and the optimal chain procedure based on the assumed dose-response function was applied to each set of \( p \)-values to perform a multiplicity adjustment. The success criterion (expected number of rejected null hypotheses) was computed for each “true” dose-response function.

• The resulting values of the success criterion were summarized across the \( k \) “true” dose-response functions using descriptive statistics.

To illustrate this sensitivity assessment approach, the dose-response function assumed under Scenario 1, i.e., \( \theta = (0.3, 0.2, 0.2) \), was considered and two sets of 50 “true” dose-response functions were randomly generated. The uncertainty parameter \( \sigma \) was set to 0.025 in the first set and to 0.05 in the second set. Noting that the mean effect size across the three dose-placebo tests was 0.23, the two values of \( \sigma \) corresponded to 10% and 20% coefficients of variation, respectively. The two sets defined two sensitivity assessment scenarios with slight and moderate deviations from the assumed dose-response function.

Table 5 lists standard descriptive statistics for the expected number of rejected null hypotheses under the two sensitivity assessment scenarios with \( n = 200 \) patients per arm. The calculations were performed using the optimal chain procedure with \( w = 1 \) and \( g = 0.5 \). Based on the assumed dose-response function, the expected number of rejections was 1.69 and, as shown in Table 5, the expected number of rejections computed from the two sets of “true” dose-response functions was slightly lower. In addition, the expected number of rejections slowly decreased as the variability in the true effect sizes for the three dose-placebo contrasts increased. For example, the mean expected number of rejected null hypotheses dropped to 1.65 with \( \sigma = 0.025 \) and 1.61 with \( \sigma = 0.05 \). The two values corresponded to the mean weighted power of 55% and 54%, respectively. As indicated above, the performance of the optimal chain procedure was anticipated to worsen since the dose-response functions considered in the sensitivity assessment were different from the assumed function; however, the overall pattern of performance loss was unknown. This sensitivity assessment provides a quantitative answer to this important question.

The proposed approach to sensitivity assessment helps the trial’s sponsor determine how much performance loss should be expected under the optimal combination of the target parameters when the true underlying dose-response function is different from the assumed function. For the sake of simplicity, only two values of \( \sigma \) were considered in this illustration and, to perform a comprehensive sensitivity assessment, it is recommended to examine a broader range of the uncertainty parameter.

Table 5: Results of the sensitivity assessment under Scenario 1 with 200 patients per treatment arm in Case study 2.

| Descriptive statistic | Deviation from the assumed dose-response function |
|----------------------|-----------------------------------------------|
|                      | Slight \((\sigma = 0.025)\) | Moderate \((\sigma = 0.05)\) |
| Expected number of rejected null hypotheses |
| First quartile       | 1.52     | 1.37     |
| Median               | 1.65     | 1.60     |
| Mean                 | 1.65     | 1.61     |
| Third quartile       | 1.78     | 1.84     |
8. Case study 3: Multiple dose-control comparisons and multiple endpoints

The CSE approach presented in Section 7 focuses on power calculations that are accompanied by fine-tuning a single analysis model (multiplicity adjustment based on a chain procedure) aimed at identifying optimal values of the chain procedure’s parameters. This section extends this evaluation exercise by including several candidate analysis models in the context of a problem with two sources of multiplicity. Specifically, the CSE approach will focus on these two goals in this case study:

- Goal 1. Selection of the best analysis model (most powerful multiple testing procedure).
- Goal 2. Optimal selection of target parameters in the identified analysis model.

8.1. Clinical trial example

A complex setting with several candidate analysis models will be illustrated using a Phase III clinical trial in patients with schizophrenia. The trial will be designed to evaluate the efficacy and safety of two doses (Dose L and Dose H) of a novel treatment compared to placebo. The efficacy evaluation will be performed using the following normally distributed endpoints:

- Endpoint E1 (primary endpoint): Change from baseline in the Positive and Negative Syndrome Scale total score.
- Endpoint E2 (key secondary endpoint): Change from baseline in the Clinical Global Impression-Severity score.

The two endpoint tests define the first source of multiplicity. The second source is defined by the two dose-placebo comparisons. As indicated above, the endpoints will be evaluated at Dose L and Dose H. The corresponding dose-placebo tests will be denoted by LP1, HP1, LP2 and HP2, respectively. For example, Test LP1 is based on the comparison between Dose L and placebo on Endpoint E1. The resulting four endpoint tests (and null hypotheses of no effect) are naturally grouped into two families of null hypotheses:

- Family $F_1$: Tests LP1 and HP1 (corresponding null hypotheses of no treatment effect are denoted by $H_1$ and $H_2$).
- Family $F_2$: Tests LP2 and HP2 (corresponding null hypotheses are denoted by $H_3$ and $H_4$).

It is important to note that clinically relevant logical restrictions exist among the four null hypotheses. Testing begins with the primary hypotheses in Family $F_1$ and the secondary hypotheses in $F_2$ will be tested as follows:

- $H_3$ will be tested only if $H_1$ is rejected.
- $H_4$ will be tested only if $H_2$ is rejected.

In other words, if a significant treatment effect is established at a particular dose on the primary endpoint, the secondary endpoint will be tested at the same dose. On the other hand, if a primary test is not significant, the corresponding secondary test is considered to be clinically irrelevant and the associated null hypothesis will be automatically accepted. The resulting logical relationships among the four null hypotheses define two testing sequences and are displayed in Figure 5.
8.2. Multiplicity adjustments

A two-stage gatekeeping procedure can be applied to address Type I error rate inflation in this multiplicity problem with two families of null hypotheses. When selecting a gatekeeping procedure, it is critical to ensure that the decision rules used in the procedure are aligned with the logical relationships depicted in Figure 5. A fairly broad class of procedures satisfy this requirement and, for the sake of example, we will focus on the following gatekeeping procedures in this case study:

- Procedure B: Bonferroni-based chain procedure (see Section 3.2).
- Procedure H: Hochberg-based multiple-sequence gatekeeping procedure (see Section 3.3).

As shown below, both gatekeeping procedures account for the logical relationships.

The decision rules used in Procedure B are displayed in Figure 6. This figure defines the procedure’s $\alpha$ allocation and $\alpha$ propagation rules. Specifically, positive weights are initially assigned to the primary null hypotheses ($w_1 = w_2 = 1/2$) and zero weights are assigned to the secondary null hypotheses ($w_3 = w_4 = 0$). After a primary hypothesis is rejected,
the associated fraction of the overall error rate will be transferred to the corresponding secondary hypothesis ($g_{13} = g_{24} = 1$).

It is worth noting that the testing strategy presented above could be enhanced by adding a re-testing step. Suppose, for example, that the hypothesis $H_1$ is rejected but $H_2$ is not in the first step of the testing algorithm. Further, if $H_3$ is rejected, the associated fraction of the error rate can be propagated back to $H_2$. This null hypothesis will then be re-tested at a higher significance level, which may lead to a rejection. To implement this strategy, non-negative values need to be assigned to the transition parameters $g_{32}$ and $g_{41}$, e.g., $g_{32} = g_{41} = 1$.

Secondly, Procedure H is a multiple-sequence gatekeeping procedure which applies a truncated version of the Hochberg procedure with a pre-specified truncation parameter $\gamma$ (0 $\leq$ $\gamma$ < 1) in $F_1$ and regular Hochberg procedure in $F_2$ (as emphasized in Section 3.3, a truncated Hochberg procedure must be used in the primary family because the regular Hochberg procedure is not separable). Let $p_i$ denote the $p$-value for the hypothesis $H_i$, $i = 1, 2, 3, 4$. Procedure H is based on the following testing algorithm (Dmitrienko et al, 2011a):

- **Family $F_1$.** The primary null hypotheses are tested using the truncated Hochberg procedure at the full $\alpha$ level. Let $p_{(1)} < p_{(2)}$ denote the ordered $p$-values in this family and let $H_{(1)}$ and $H_{(2)}$ denote the associated ordered null hypotheses. Both $H_{(1)}$ and $H_{(2)}$ are rejected if $p_{(2)} \leq \alpha(1 + \gamma)/2$ or only $H_{(1)}$ is rejected if $p_{(1)} \leq \alpha/2$ and $p_{(2)} > \alpha(1 + \gamma)/2$.

- **Family $F_2$.** The secondary null hypotheses are tested using regular Hochberg procedure at the full $\alpha$ level if both primary null hypotheses are rejected. Let $p_{(3)} < p_{(4)}$ denote the ordered $p$-values and let $H_{(3)}$ and $H_{(4)}$ denote the ordered null hypotheses. Both $H_{(3)}$ and $H_{(4)}$ are rejected if $p_{(4)} \leq \alpha$ or only $H_{(3)}$ is rejected if $p_{(3)} \leq \alpha/2$ and $p_{(4)} > \alpha$. Further, if only $H_1$ is rejected in the primary family, $H_3$ is rejected if $p_3 \leq \alpha(1 - \gamma)/2$ and $H_4$ is automatically accepted. Likewise, if only $H_2$ is rejected in the primary family, $H_4$ is rejected if $p_4 \leq \alpha(1 - \gamma)/2$ and $H_3$ is accepted.

In general, a higher value of the truncation parameter $\gamma$ helps improve power in the primary family but may reduce power in the secondary family. Optimal selection of this target parameter will be discussed later in this section.

Further, just like Procedure B, Procedure H does not include an option to re-test a non-rejected primary null hypothesis after testing is completed in the secondary family. The trial’s sponsor could consider extending the Hochberg-based gatekeeping procedure defined above to implement a re-testing strategy; see, for example, Dmitrienko et al (2011a).

### 8.3. Clinical scenario evaluation

The CSE approach was applied in this example using the statistical assumptions for the primary and secondary endpoint tests presented in Table 6. The two endpoints were assumed to follow a bivariate normal distribution. The table defines two scenarios for the effect sizes (effect sizes for the four tests are denoted by $\theta_1, \ldots, \theta_4$, respectively). Under Scenario 1, the effect sizes for the primary tests were generally comparable to those for the secondary tests. Scenario 2 assumed lower effect sizes in the secondary family compared to the primary family. Lastly, the correlation in the bivariate normal distribution was set to 0.3.
Table 6: Assumed treatment effects for the primary tests (Tests LP1 and HP1) and secondary tests (Tests LP2 and HP2) in Case study 3.

| Scenario | Effect size | Family $F_1$ | Test LP1 | Test HP1 | Family $F_2$ | Test LP2 | Test HP2 |
|----------|-------------|--------------|----------|----------|--------------|----------|----------|
| Scenario 1 | $\theta_1 = 0.23$ | $\theta_2 = 0.27$ | $\theta_3 = 0.25$ | $\theta_4 = 0.25$ |
| Scenario 2 | $\theta_1 = 0.23$ | $\theta_2 = 0.27$ | $\theta_3 = 0.20$ | $\theta_4 = 0.20$ |

As in Case study 2 (see Section 7), it is recommended to begin with determining a general range of sample sizes to be used as part of the CSE approach. The trial’s outcome will be declared positive if a beneficial effect on the primary endpoint is established at Dose L or Dose H. This definition of success translates into the disjunctive criterion applied to Family $F_1$ (probability of rejecting $H_1$ or $H_2$). A simple simulation-based assessment showed that, with $n = 280$ patients per treatment arm, disjunctive power for Procedure B and Procedure H with the truncation parameter $\gamma = 0.8$ was equal to 89.0% and 89.6% under Scenario 1 and 89.1% and 89.8% under Scenario 2, respectively\(^2\) (since the performance of Procedure H in the primary and secondary families depends on the truncation parameter $\gamma$, this parameter was initially set to 0.8 to improve power in the primary family). Using this sample size, the trial’s sponsor can perform a comprehensive comparison of the two gatekeeping procedures (Goal 1) and select their parameters to arrive at the most efficient multiplicity adjustment (Goal 2). Once the general performance of the multiple testing procedure is well understood, the sponsor can consider additional adjustments to the sample size to ensure a desirable probability of success.

Beginning with Goal 1, recall that, in this clinical trial example, it is desirable to reject at least one null hypothesis in the primary family and, for label claim purposes, it may also be desirable to reject the corresponding null hypothesis in the secondary family. This criterion provides motivation for the use of disjunctive power in Family $F_1$ as well as $F_2$. As was pointed out in Section 7, it is good practice to investigate other success criteria, including conjunctive power in both families. Conjunctive power is an appropriate criterion to consider if the trial’s sponsor is interested in evaluating the probability of detecting a significant treatment effect at both doses, which is helpful for characterizing the dose-response relationship and performing risk-benefit assessments.

To support a comprehensive assessment of the two gatekeeping procedures, Table 7 provides a summary of the following key operating characteristics with $n = 280$ patients per arm:

- Disjunctive power in Family $F_1$ (probability of rejecting $H_1$ or $H_2$).
- Conjunctive power in Family $F_1$ (probability of rejecting $H_1$ and $H_2$).
- Disjunctive power in Family $F_2$ (probability of rejecting $H_3$ or $H_4$).
- Conjunctive power in Family $F_2$ (probability of rejecting $H_3$ and $H_4$).

\(^2\)It should be noted that the small differences in the disjunctive and conjunctive power for the primary endpoint tests observed between Scenarios 1 and 2 for each procedure are due to random variation.
Table 7: Summary of power evaluation with 280 patients per treatment arm in Case study 3.

| Criterion                  | Treatment effect scenario | Scenario 1 | Scenario 2 |
|----------------------------|---------------------------|------------|------------|
|                            | Procedure B (γ = 0.8)     | Procedure H (γ = 0.8) | Procedure B | Procedure H |
| Primary endpoint tests     | Disjunctive power (%)     | 89.0       | 89.1       | 89.8       |
|                            | Conjunctive power (%)     | 61.7       | 61.4       | 70.7       |
| Secondary endpoint tests   | Disjunctive power (%)     | 78.4       | 76.2       | 62.4       | 60.4       |
|                            | Conjunctive power (%)     | 43.3       | 56.4       | 27.0       | 39.0       |

Considering the primary endpoint tests, recall that Procedure B applies the basic Bonferroni procedure and Procedure H uses the truncated Hochberg procedure with γ = 0.8 in Family $F_1$. It follows from Table 7 that Procedure H was virtually equivalent to Procedure B in terms of disjunctive power in the primary family under both scenarios. However, when conjunctive power in Family $F_1$ was considered, Procedure H provided substantial improvement over the other gatekeeping procedure. The gain in conjunctive power was due to an important feature of the Hochberg procedure. The Hochberg procedure is derived from the Simes global test and, as a consequence, it maximizes the probability of simultaneous rejection of all null hypotheses within a family compared to the basic Bonferroni procedure. Further, power comparisons in the secondary family revealed that an application of Procedure B resulted in higher disjunctive power than Procedure H under both scenarios. This is driven by the fact that a fairly large value of the truncation parameter γ was utilized in the primary family which led to reduced power of the secondary tests. However, as in Family $F_1$, Procedure H provided an important advantage over Procedure B in terms of conjunctive power in Family $F_2$. Lastly, lower disjunctive power and conjunctive power were observed for both gatekeeping procedures under Scenario 2 since this scenario assumed smaller effect sizes for both dose-placebo tests.

### 8.4. Selection of optimal procedure parameters

The next CSE goal in this case study deals with optimal selection of target parameters in the two analysis models (parameters of gatekeeping procedures). Procedure H will be used to illustrate the process and optimization will be performed with respect to the truncation parameter γ. This parameter helps control the balance between the levels of disjunctive or conjunctive power in the primary and secondary families.

When the target parameter is set to a very large value, say, 0.95 (note that γ cannot be set to 1 to ensure the procedure is separable), the procedure used in Family $F_1$ is very close to the regular Hochberg procedure, which increases power of the primary tests and decreases power of the secondary tests. On the other hand, with γ = 0, the truncated Hochberg procedure simplifies to the Bonferroni procedure, which leads to lower power in Family $F_1$ but may improve power in Family $F_2$. To find a “sweet spot”, it is helpful to examine an appropriate criterion (disjunctive or conjunctive power) in each family as a function of the target parameter.

Using $n = 280$ patients per treatment arm, Figure 7 displays a bivariate plot of disjunctive power in Family $F_1$ versus disjunctive power in Family $F_2$ under Scenario 1 (left panel)
Fig. 7: Relationship between disjunctive power in the primary family (Family $F_1$) and disjunctive power in the secondary family (Family $F_2$) for Procedure H under Scenario 1 (left panel) and Scenario 2 (right panel) with 280 patients per treatment arm in Case study 3. The labels indicate the values of the truncation parameter $\gamma$.

and Scenario 2 (right panel) as a function of $\gamma^3$ (as stated above, the truncation parameter $\gamma$ must be less than 1 and this $\gamma = 1$ is excluded from consideration). The general patterns (i.e., the relationship between the two power functions) were quite similar in the two panels. In both cases, selecting a large value of $\gamma$ resulted in higher power in the primary family compared to $\gamma = 0$. However, the power gain was trivial (about 1% improvement) and was easily offset by a tangible power gain in the secondary family. Under Scenario 1, disjunctive power in Family $F_2$ increased by 17.6% as the target parameter changed from 1 to 0. An overall assessment of the power in the two families suggests that $\gamma = 0$ provides an optimal balance between disjunctive power in Family $F_1$ and disjunctive power in Family $F_2$ in both scenarios.

It is helpful to note that a conceptually similar optimization algorithm can be constructed using the conjunctive or weighted power criteria defined in Section 4. As an example, Figure 8 shows bivariate plots based on conjunctive power in the two families under Scenario 1 (left panel) and Scenario 2 (right panel) as a function of the target parameter ($\gamma$) with $n = 280$ patients per arm. Conjunctive power was maximized in both families if $\gamma = 1$ since both null hypotheses must be rejected in Family $F_1$ in order to reject both null hypotheses in Family $F_2$. If the trial’s sponsor is interested in achieving a “compromise” between the disjunctive and conjunctive criteria in this clinical trial example, $\gamma$ could be set, for example, to 0.8 because this value of the target parameter guarantees reasonably high levels of disjunctive and conjunctive power.

Further, with weighted power, the trial’s sponsor can assign importance parameters to the two families and then select the target parameter by maximizing an appropriately defined weighted power function, e.g., a weighted sum of the disjunctive power functions in Families $F_1$ and $F_2$.

$^3$The curves in Figure 7 lack smoothness due to random variation.
Finally, as in Case study 2, it is important to consider sensitivity analyses aimed at assessing the impact of important parameters of the data model such as the effect sizes and correlation between the two endpoints on the overall conclusions presented above.

9. Summary

The main goal of this paper was to discuss key considerations in the general problem of power evaluation and sample size calculations in trials with multiple clinical objectives. Analysis of multiple objectives induces multiplicity and we provided a summary of statistical methods that are commonly used to perform multiplicity adjustments in confirmatory Phase III trials. We demonstrated the main differences between power evaluation in a simple clinical trial setting with a single objective and more challenging settings with multiple objectives, including the choice of an appropriate definition of a successful trial outcome (success criterion). Popular success criteria were introduced and their properties were discussed. We also defined a comprehensive approach aimed at examining the impact of all relevant factors (including parameters of data and analysis models) on the selected success criteria (Clinical Scenario Evaluation approach or CSE approach). This approach was implemented using a new powerful R package (Mediana package) which facilitates systematic quantitative assessment of commonly used trial designs and analysis methods.

Examples of power evaluation were provided based on three case studies that are representative of increasingly more complex multiplicity problems encountered in Phase III clinical trials. These case studies illustrated the importance of general principles of the CSE approach.

Beginning with the end in mind, all three case studies emphasized the importance of selecting one or more success criteria that best fit the clinical objectives of a given trial. Certain widely used criteria such as disjunctive power are often misapplied and may lead to
incorrect recommendations for the choice of the overall sample size or analysis methods (see, for example, Case study 2). Clinical trial researchers are encouraged to carefully choose the most appropriate definition of a successful outcome in each trial.

Evaluations in all three case studies were performed under several plausible sets of data model parameters such as the assumed treatment effects for each individual objective. A closely related task is sensitivity assessment. An “optimal” trial design or analysis method is defined locally rather than globally in the sense that it is conditional upon a particular set of parameters. It is critical to assess the robustness of the selected design and analysis methodology with respect to statistical assumptions. We provided examples of sensitivity assessments in a univariate setting (i.e., with respect to a single data model parameter) in Case study 1 and a multivariate setting in Case study 2.

As the complexity of a multiplicity problem grows, so does the number of applicable multiple testing procedures. Case study 3 demonstrated the value of comparing the operating characteristics of several candidate analysis strategies (multiplicity adjustment strategies) and selecting the analysis method which provides the best overall performance. Another important consideration in the context of confirmatory clinical trials is the evaluation of success probabilities across multiple clinical trials within a development program. Program-wise power calculations and consistency assessments across several trials included in a program were discussed in Case study 1.

Finally, in this paper we focused on a frequentist formulation of the problem of power evaluation and clinical trial optimization in settings with multiple objectives. As a potential extension of this general framework, a clinical trial’s sponsor can consider assessing probability of success based on Bayesian approaches. This topic has attracted much attention in the clinical trial literature — see, for example, Chuang-Stein (2006), Hobbs and Carlin (2008) and Wang (2015). By utilizing Bayesian methods, clinical trial researchers can account for the variability of unknown design parameters which can be obtained from historical information, i.e., previously conducted studies. Bayesian methodology can be easily integrated into the general CSE approach and will lead to a more reliable assessment of a clinical trial’s operating characteristics.

REFERENCES

Benda, N., Branson, M., Maurer, W., Friede, T. (2010). Aspects of modernizing drug development using clinical scenario planning and evaluation. *Drug Information Journal*. 44, 299–315.

Hobbs, B.P., Carlin, B.P. (2008). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*. 18, 54–80.

Brechenmacher, T., Xu, J., Dmitrienko, A., Tamhane, A.C. (2011). A mixture gatekeeping procedure based on the Hommel test for clinical trial applications. *Journal of Biopharmaceutical Statistics*. 21, 748–767.

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

Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statistics*. 5, 305–309.

Dmitrienko, A., Tamhane, A.C., Wiens B. (2008). General multistage gatekeeping procedures. *Biometrical Journal*. 50, 667–677.

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 Sciences*. 1, 1–22.
Statistics. Dmitrienko, A., Tamhane, A.C., Bretz, F. (editors). Chapman and Hall/CRC Press, New York.

Dmitrienko, A., Millen, B., Brechenmacher, T., Paux, G. (2011a). Development of gatekeeping strategies in confirmatory clinical trials. *Biometrical Journal*. 53, 875–893.

Dmitrienko, A., Kordzakhia, G., Tamhane, A.C. (2011b). Multistage and mixture parallel gatekeeping procedures in clinical trials. *Journal of Biopharmaceutical Statistics*. 21, 726–747.

Dmitrienko, A., Tamhane, A.C. (2009). Gatekeeping procedures in clinical trials. *Multiple Testing Problems in Pharmaceutical Statistics*. Dmitrienko, A., Tamhane, A.C., Bretz, F. (editors). Chapman and Hall/CRC Press, New York.

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

Dmitrienko, A., Tamhane, A.C. (2013). General theory of mixture procedures for gatekeeping. *Biometrical Journal*. 55, 402–419.

Dmitrienko, A., D’Agostino, R.B. (2013). Tutorial in biostatistics: Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*. 32, 5172–5218.

Dmitrienko, A., D’Agostino, R.B., Huque, M.F. (2013). Key multiplicity issues in clinical drug development. *Statistics in Medicine*. 32, 1079–1111.

Dmitrienko, A., Kordzakhia, G., Brechenmacher, T. (2015). Mixture-based gatekeeping procedures for multiplicity problems with multiple sequences of hypotheses. *Journal of Biopharmaceutical Statistics*. To appear.

Dunnett, C.W., Tamhane, A.C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine*. 10, 939–947.

Dunnett, C.W., Tamhane, A.C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*. 87, 162–170.

Friede, T., Nicholas, R., Stallard, N., Todd, S., Parsons, N.R., Valdes-Marquez, E., Chataway, J. (2010). Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug Information Journal*. 44, 713–718.

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

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

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

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.

Meltzer, H.Y. et al. (2011). Lurasidone in the treatment of schizophrenia: a randomized, double-blind, placebo- and olanzapine-controlled study. *American Journal of Psychiatry*. 168, 957–967.

Millen, B., Dmitrienko, A. (2011). Chain procedures: A class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research*. 3, 14–30.

Nasrallah, H.A. et al. (2013). Lurasidone for the treatment of acutely psychotic patients with schizophrenia: A 6-week, randomized, placebo-controlled study. *Journal of Psychiatric Research*. 47, 670–677.
Sankoh, A.J., D’Agostino, R.B., Huque, M.F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine*. 22, 3133–3150.

Sarkar, S.K. (2008). On the Simes inequality and its generalization. *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*. Balakrishnan, N., Pena, E.A., Silvapulle, M.J. (editors). Institute of Mathematical Statistics: Beachwood, Ohio.

Wang, M.-D. (2015). Applications of probability of study success in clinical drug development. *Applied Statistics in Biomedicine and Clinical Trials Design*. Chen, Z., Liu, A., Qu, Y., Tang, L., Ting, N., Tsong, Y. (editors). Springer, New York.

Wang, S.J., Bretz, F., Dmitrienko, A., Hsu, J., Hung, J., Huque, M., Koch, G. (2013). Panel forum on multiple comparison procedures: A commentary from a complex trial design and analysis plan. *Biometrical Journal*. 55, 275–293.

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

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

Wiens, B.L. (2011). Multiple testing procedures in two pivotal clinical trials for approval of a new pharmaceutical product. *Journal of Biopharmaceutical Statistics*. 21, 669–681.

Wiens, B., Dmitrienko, A., Marchenko, O. (2013). Selection of hypothesis weights and ordering when testing multiple hypotheses in clinical trials. *Journal of Biopharmaceutical Statistics*. 23, 1403–1419.

(Received: November 21, 2014, Accepted: February 24, 2015)