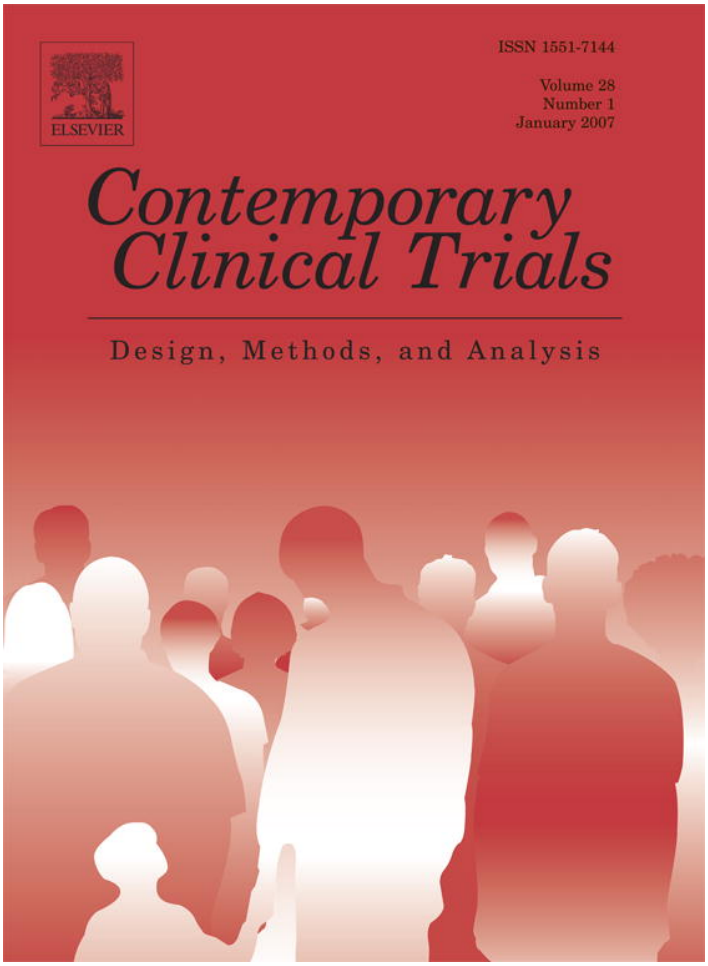

Available online at www.sciencedirect.com
ScienceDirect
Contemporary Clinical Trials 28 (2007) 68–78
Contemporary Clinical Trials
www.elsevier.com/locate/conclintrial

Dependence, hyper-dependence and hypothesis testing in clinical trials

Lemuel A. Moyé *, Sarah Baraniuk

University of Texas School of Public Health, 1200 Herman Pressler E815, Houston, Texas 77025, USA

Received 1 January 2006; accepted 31 May 2006

Abstract

While investigators designing clinical trials face the important issue of endpoint selection, an equally troublesome concern can be the *a priori* selection of the endpoint analysis. In this latter circumstance, there may be only one endpoint of interest in the clinical trial, but several competing endpoint analyses are available (e.g., an analysis of the endpoint that is adjusted for clinical center versus an analysis that is adjusted for geographic region versus an unadjusted analysis). An example that demonstrates the unsatisfactory conclusions that ambiguous choices can produce is offered.

A procedure utilizing conditional probability is provided that permits the conservation of type I error when the investigators have one endpoint and several worthy competitor endpoint analyses that are each prospectively identified and carried out at the trial's conclusion. When the high levels of dependence among these analyses are taken into account, it is possible to carry out the hypothesis tests in a way that 1) provides practicable type I error levels for each analysis, and 2) conserves the familywise type I error. In circumstances in which the endpoint and all members of the family of analyses are selected during the design phase of the trial, this procedure provides confirmatory conclusions as opposed to exploratory findings.

Keywords: Dependence; Endpoint analyses; Clinical trials; Multiplicity; Multiple endpoints

1. Introduction

Randomized clinical trials can be costly, and the natural tendency of many investigators is to increase the yield of these expensive experiments by maximizing the number of statistical evaluations. However, clinical trial investigators have been wisely counseled against yielding to the temptation of executing an undisciplined analysis plan. The requirements of the prospective selection of endpoint analyses in clinical trials have been well described in the literature [1–4], and the standard of good investigator practice mandates the clear choice of the clinical trial's endpoints and analysis plans during the design phase of the study. In fact, specific information is now available to investigators that guides the endpoint selection process during the trial's planning period, assisting them in deciding whether an analysis should be 1) confirmatory and pivotal, 2) supportive, or 3) exploratory [5].

* Corresponding author.
E-mail address: Lemuel.A.Moye@uth.tmc.edu (L.A. Moyé).

1551-7144/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2006.05.010

The acceptance of the aforementioned guidance commonly produces important discussions among the investigators during the design phase of the clinical trial as they labor to identify the primary endpoint analyses of their study. Frequently, such prospective discussions are fruitful, leading to clear decisions that themselves produce lucid analysis plans and, ultimately, unambiguous interpretations [6,7]. Unfortunately, this is not always the case. Clinical scientists can find the endpoint selection process a vexing one, and in the end may be unable to make a final choice of a single endpoint and analysis plan for their study. This is not due to lack of effort on the part of these investigators who wish to reduce any ambiguity in their trial's interpretation. They diligently exclude many candidate endpoints and analyses as they reduce the prospectively declared analysis set to a small number of evaluations. However, at the conclusion of this labor, the investigators may be unable to make a final choice because there is no persuasive or historical basis on which to make a final selection.

The notion of multiplicity has received considerable discussion in the literature. A comprehensive review has been provided by Miller [8]. Adjustments that take into account dependency between the different statistical analyses have been recommended by Tukey et al. [9], and others [10–18]. Sequentially rejective procedures have been reviewed and advocated [19,20], and recent work [21–24] has refined these tools. However, guidelines for investigators who feel compelled to chose a single analysis from multiple admissible ones are missing. Sometimes the research group's collective ambivalence is conveyed in the protocol. The outcome of this circumstance can be unfortunate and disappointing, as illustrated in the following actual example.

In the afternoon session of the 69th meeting of the Oncologic Drugs Advisory Committee meeting on December 6, 2001, consideration was given to expanding the indications for Gliadel, an adjuvant treatment of malignant glioma [25]. Malignant glioma is a brain tumor, diagnosed in approximately 16,500 adult patients per year and from which 13,000 patients die annually [25]. The treatment of choice is surgical extirpation of the tumor mass followed by chemotherapy. Gliadel is a medication combining two antitumor compounds into a wafer that is placed in the tumor cavity of patients who have just had the tumor surgically extracted.

In response to the FDA's concerns that previously submitted clinical studies on the effect of Gliadel in patients with newly diagnosed glioma were too small, the sponsor carried out study T-301. T-301 was a randomized, double-blind, placebo-controlled clinical trial which recruited 240 patients from 42 regional centers in 14 countries, randomly assigning these patients to each of the treatment and control group. T-301 was stratified by center and, as a result, indirectly stratified by country.

The FDA had reviewed the protocol and the statistical analysis plan for T-301 before all patients had completed their follow up and the data were unblinded. The primary, pre-specified efficacy analysis was the comparison of 12 week survival rates using a standard Kaplan–Meier analysis, implementing a log rank test statistic in accordance with the intention-to-treat principle.

The results of the trial demonstrated that the median survival for patients taking Gliadel was 13.9 months (95% CI 12.1 to 15.3) as compared to 11.6 months for the placebo group (95% CI 10.2 to 12.6). The relative risk was 0.77 and the p-value for the log rank test was 0.08, larger than the prospectively set level of 0.05.

Upon the conclusion of the study, an independent statistician was consulted by the sponsor to analyze the T-301 efficacy data. He carried out an alternative evaluation in which the therapy effect was stratified by center, producing a p-value of 0.07; a third analysis stratified by country produced a p-value of 0.03. The sponsor argued that since 1) the study was stratified by center, and 2) that this center-level stratification produced stratification at the country level, then the most appropriate evaluation should be the country-stratified log rank analysis [26]. It was this statistically significant analysis that the sponsor asked the FDA to accept as the definitive evaluation of the primary endpoint of T-301.

The sponsor's advocates argued that, since the study design called for blocking and stratification, then the analysis must be dictated by the intent of the design, even though the prospectively declared analysis plan stated only that a log rank test statistic would be used. They asserted that "…one would expect the analyses to be stratified, and the analysis statistic to be stratified in the same way that the randomization was performed" [25]. The sponsor further argued that *de jure* stratification by center was *de facto* stratification by country, and therefore the most appropriate analysis would call for the use of a log rank test statistic that was stratified by country. The FDA's respondent countered by raising a question and then answering it:

> …should one use a stratified or non-stratified analysis? Which one is more appropriate? Our position is, either one is acceptable as long as you pre-specify one in the protocol [25, page 293]

The FDA further asserted that, if the analysis that was stratified by country was to be the principal analysis, then the fact that it was one of three evaluations requires the application of a multiplicity correction.

In the end, the Advisory Committee voted 7–6 that T-301 was not adequate and well controlled, but did vote that the medication had been shown to be safe and effective.

The *post hoc* decision to adjust the log rank analysis received special attention at the meeting, and is not the focus of this manuscript. Either of the three analyses that were carried out (unadjusted, center-adjusted, or country-adjusted) was acceptable if prospectively specified; the difficult prospective decision that these investigators faced was which of the three should be selected. Since the intervention that they were studying had important clinical relevance, then, from the investigators' perspective, the analysis procedure that measured that intervention's effect should 1) be acceptable to the research and regulatory communities and 2) generate a powerful test statistic best able to appropriately identify a clinically and statistically significant effect. However, in this circumstance, each of the three candidates met criterion (1) and could potentially meet criterion (2). It is the absence of a compelling reason to select one over the other two of these dependent analyses that makes the *a priori* selection of an endpoint and its analysis so challenging.

The purpose of this manuscript is to demonstrate how extreme dependence between prospectively determined primary analyses in a clinical trial can be used to develop decision rules that preclude the need for clinical investigators to choose one and only one of a small number of closely related, acceptable statistical procedures during the trial's design phase.

2. Methodology

The development follows [5]. Assume that, in a randomized clinical trial, there are K prospectively declare primary hypothesis tests $H_1, H_2, H_3, \ldots H_K$. Let H_j denote the jth hypothesis test. For each of these K hypothesis tests, specify the prospectively specified type I error levels $\alpha_1, \alpha_2, \alpha_3, \ldots \alpha_K$. Define T_j for $j=1, 2, 3, \ldots, K$ as a variable that captures whether a type I error has occurred for the jth hypothesis test, i.e., $T_j=0$ if there is no type I error on the jth hypothesis test, and $T_j=1$ if the jth hypothesis test produces a type 1 error. Thus, we can consider K pairs, (H_1, T_1), (H_2, T_2), $(H_3, T_3), \ldots, (H_K, T_K)$, where H_j identifies the statistical hypothesis test and T_j denotes whether a type I error has occurred for that test, i.e., $P[T_j=1]=\alpha_j$.

Using the customary definition of the familywise error as the event that there is at least one type I error among the K prospectively defined primary analyses [27,28], define ξ as the familywise error level, and T_ξ as the variable that denotes whether a familywise type I error level has occurred. Then $\xi=P[T_\xi=1]$, and $P[T_\xi=0]$ is the probability that there were no type I errors among the K hypothesis tests. Therefore

$$P(T_\xi = 0) = P(\{T_1 = 0\} \cap \{T_2 = 0\} \cap \{T_3 = 0\} \cap \ldots \cap \{T_K = 0\}) \tag{1}$$

and

$$P(T_\xi = 1) = 1 - P(\{T_1 = 0\} \cap \{T_2 = 0\} \cap \{T_3 = 0\} \cap \ldots \cap \{T_K = 0\}) = 1 - P\left(\bigcap_{j=1}^{K} T_j = 0\right). \tag{2}$$

When the K individual hypotheses are independent of one another, then $P\left(\cap_{j=1}^{K} T_j = 0\right) = \Pi_{j=1}^{K} P(T_j = 0) = \prod_{j=1}^{K} (1-\alpha_j)$. However, if the K prospectively specified hypothesis tests are dependent, then the evaluation of the expression $P(\cap_{j=1}^{K} T_j = 0)$ becomes more complicated.

We will proceed with our evaluation of the $P(\cap_{j=1}^{K} T_j = 0)$ in stages. Starting with the circumstance for $K=2$, we will contrast the computation of $P\left(\cap_{j=1}^{K} T_j = 0\right)$ in each of the independence and dependence settings. This comparison will permit the definition of a dependence term D that, when specified, can be used to compute ξ when α_1 and α_2 are computed, or, alternatively, to compute α_2 when ξ and α_1 are specified.

In the independence setting for $K=2$, write

$$P[T_1 = 0 \cap T_2 = 0] = P[T_2 = 0 | T_1 = 0] P[T_1 = 0]. \tag{3}$$

This will be a useful equation for us as we develop the notion of dependency in hypothesis testing, since the key to computing the probability of a familywise error $P[T_\xi=0]$ is the computation of the joint probability $P[T_1=0 \cap T_2=0]$. This calculation is straightforward in the independence scenario.

$$P[T_2 = 0 | T_1 = 0] = \frac{P[T_1 = 0 \cap T_2 = 0]}{P[T_1 = 0]} = \frac{(1-\alpha_1)(1-\alpha_2)}{(1-\alpha_1)} = 1-\alpha_2 \tag{4}$$

The opposite circumstance, one of extreme dependence, will be defined as "perfect dependence". Perfect dependence denotes that state between two statistical hypothesis tests in which the occurrence of a type I error for H_1 automatically produces a type I error for statistical hypothesis test H_2. In this situation, the two tests are so intertwined that knowledge that a type I error occurred for the first hypothesis test guarantees that a type I error will occur for the second hypothesis test. Perfect dependence dictates that the conditional probability from Eq. (3) is one, i.e.,

$$P[T_2 = 0|T_1 = 0] = 1. \tag{5}$$

Recalling that $\xi = 1 - P[T_1 = 0 \cap T_2 = 0]$, compute that

$$\xi = 1-P[T_1 = 0 \cap T_2 = 0] = 1-P[T_2 = 0|T_1 = 0]P[T_1 = 0] = 1-(1)(1-\alpha_1) = \alpha_1. \tag{6}$$

Since the occurrence of a type I error on the first statistical hypothesis test implies that a type I error has occurred on the second hypothesis test, the joint occurrence of type I errors is determined by what occurs on H_1. We can, without any loss of generality, order these two hypothesis tests prospectively such $\alpha_1 \geq \alpha_2$. In the setting of perfect dependence, one can execute two hypothesis tests and maintain ξ at its desired level by simply allowing α_2 to take any value such that $\alpha_2 \leq \alpha_1 = \xi$. As an example, consider the hypothetical case of a clinical trial in which there are two prospectively defined primary hypothesis tests H_1 and H_2 with associated test-specific α error levels α_1 and α_2. Choose $\alpha_1 = \alpha_2 = 0.05$. In the familiar case of independence, it is clear that $\xi = 1 - (0.95)(0.95) = 0.0975$. However, under the assumption of perfect dependence ξ remains at 0.05.

In clinical trials, rarely does one have either a collection of prospectively declared primary analyses that are completely independent of one another, or a set of *a priori* analyses that are perfectly dependent. Our goal is to examine the range of dependency between these two extremes, and then compute ξ and α_2 as needed. Since these two extremes reflect the full range of dependence, write

$$1-\alpha_2 \leq P[T_2 = 0|T_1 = 0] \leq 1. \tag{7}$$

We can develop a measure D, which will reflect this level of dependence contained on [0, 1]. The instance when D is zero should correspond to the condition of independence between the statistical hypothesis tests, and identify the situation in which $P[T_2 = 0|T_1 = 0] = 1 - \alpha_2$. Analogously, $D = 1$ will denote perfect dependence, i.e., the case in which the conditional probability of interest $P[T_2 = 0|T_1 = 0]$ attains its maximum value of one. This can be written as

$$1-\alpha_2 \leq P[T_2 = 0|T_1 = 0] \leq 1, 0 \quad \leq \quad D^2 \quad \leq 1. \tag{8}$$

If we are to choose a value of D that will have the aforementioned properties, then we can write D in terms of the conditional probability

$$D = \sqrt{1-\frac{(1-P[T_2 = 0|T_1 = 0])}{\alpha_2}}. \tag{9}$$

In general, we will not use Eq. (9) to compute D. Our ultimate goal is to supply the value of D, and then write the familywise error level in terms of D^2.

$$P[T_2 = 0|T_1 = 0] = (1-\alpha_2) + D^2[1-(1-\alpha_2)] = 1-\alpha_2(1-D^2). \tag{10}$$

The familywise error level for the two statistical hypothesis tests H_1 and H_2 may be written as

$$\begin{aligned} \xi &= 1-P[T_2 = 0 \cap T_1 = 0] \\ \xi &= 1-P[T_2 = 0|T_1 = 0]P[T_1 = 0] = 1-[1-\alpha_2(1-D^2)](1-\alpha_1). \end{aligned} \tag{11}$$

Therefore, the familywise error is formulated in terms involving the test-specific α error rates α_1, α_2 where $\alpha_1 \geq \alpha_2$, and the dependency parameter D.

During the design phase of the trial, as investigators work to select the appropriate levels of the test-specific α error levels for the study, they can first fix ξ, and then choose α_1 and D, moving on to compute the acceptable range of α_2. This is easily accomplished, recalling the assumption that the hypothesis tests are ordered so that $\alpha_1 \geq \alpha_2$.

$$\alpha_2(\max) = \min\left[\alpha_1, \frac{\xi - \alpha_1}{(1-\alpha_1)(1-D^2)}\right]. \tag{12}$$

Eq. (12) provides the maximum value of α_2 that will preserve the familywise error. Denote this maximum value as $\alpha_2(\max)$.

The case for $K=3$ is a straightforward generalization of the consideration for two endpoints and we can carry forward the same nomenclature developed above. We will also assume that $\alpha_1 \geq \alpha_2 \geq \alpha_3$. We may write the familywise type I error as

$$\xi = 1 - P[T_1 = 0 \cap T_2 = 0 \cap T_3 = 0]. \tag{13}$$

where

$$P[T_1 = 0 \cap T_2 = 0 \cap T_3 = 0] = P[T_3 = 0 \cap T_1 = 0 \cap T_2 = 0]P[T_1 = 0 \cap T_2 = 0]. \tag{14}$$

Following the development for the case where $K=2$, write

$$D_{3|1,2} = \sqrt{1 - \frac{(1 - P[T_3 = 0 | T_1 = 0 \cap T_2 = 0])}{\alpha_3}}. \tag{15}$$

$D_{3|1,2}$ measures the degree of dependence between H_3 given knowledge of H_1 and H_2. Solve Eq. (15) for the conditional probability

$$P[T_3 = 0 | T_1 = 0 \cap T_2 = 0] = (1-\alpha_3) + D^2_{3|1,2}[1-(1-\alpha_3)] = 1 - \alpha_3(1 - D^2_{3|1,2}). \tag{16}$$

Now insert the relationship expressed in Eq. (16) into Eq. (14) to find

$$\begin{aligned} \xi &= 1 - P[T_1 = 0 \cap T_2 = 0 \cap T_3 = 0] = 1 - P[T_3 = 0 | T_1 = 0 \cap T_2 = 0]P[T_1 = 0 \cap T_2 = 0] \\ &= 1 - [1 - \alpha_3(1 - D^2_{3|1,2})]P[T_1 = 0 \cap T_2 = 0]. \end{aligned} \tag{17}$$

and

$$\xi = 1 - [1 - \alpha_3(1 - D^2_{3|1,2})][1 - \alpha_2(1 - D^2_{2|1})][1 - \alpha_1]. \tag{18}$$

Solving Eq. (18) for α_3 reveals

$$\alpha_3(\max) = \min\left[\alpha_2, \frac{1 - \frac{1-\xi}{[1-\alpha_1][1-\alpha_2(1-D^2_{2|1})]}}{1 - D^2_{3|1,2}}\right]. \tag{19}$$

Results for the circumstance for $K>3$ are available [5].

3. Results

This methodology was motivated by consideration of the dilemma of investigators who are commonly unable to choose one from several admissible statistical procedures for the primary analysis of the study. Consideration of the similarity of p-values produced from these analyses suggested an important type I error rate linkage between them leading to the development of the dependence parameter. Discussions concerning the selection of the dependency parameter D are available [5].

Eq. (12) demonstrates that when $K=2$ and α_1 is fixed, the maximum value of α_2 increases as D increases. A particularly relevant issue from the development of the preceding section is the range of maximum values of α_2 that are

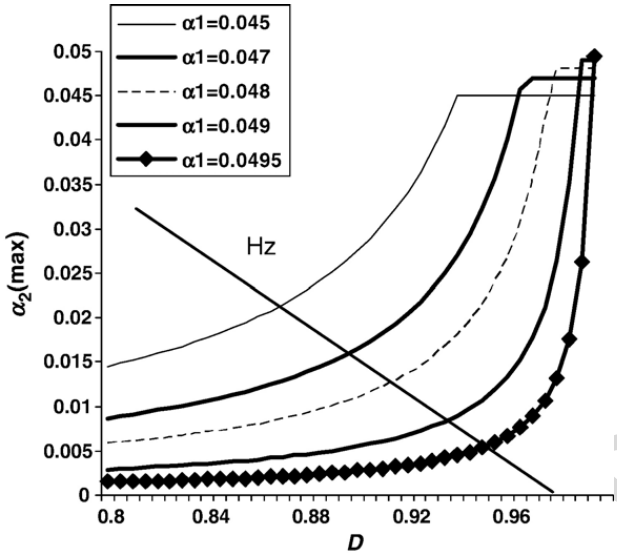

Fig. 1. Relationship between α_2 (max) and the dependency parameter D as a function of α_1 when the familywise error rate is 0.05. The region to the right of line Hz is the zone of hyper-dependence.

permitted when the level of D is high (Fig. 1). As an example, for $\alpha_1 = 0.047$, the maximum value of α_2 rapidly grows as D increases beyond 0.90. For each value of α_1 one can identify a range of high values of D above which the growth of the maximum value of α_2 accelerates. This range of D will be denoted as the zone of hyper-dependence, in which the relationship between the occurrence of type I errors for H_1 and H_2 are so great that type I error is essentially conserved even when the values of both α_1 and α_2 are large.

A similar evaluation reveals that, in the case for $K = 3$ prospectively declared analyses, large values of D produce relatively large values of α_1, α_2, and α_3 (Fig. 2).

Fig. 2 demonstrates the relationship between the maximum value of the test specific alpha level for H_3, α_3, as a function of the dependency parameter $D_{3|1,2}$ for different values of α_2. For each of the five curves in this figure, the familywise error rate is 0.05 and $D_{2|1} = 0.95$. This hyper-dependent condition between H_1 and H_2 permits a maximum value of α_2 of 0.045. Fig. 2 reveals that the maximum value of α_3 increases as a function of $D_{3|1,2}$. Also, just as there was a zone of hyper-dependence in Fig. 1 for α_2, Fig. 2 reveals that there is a similar zone of hyper-dependence for the maximum value of α_3. However, the value of α_2 is not very critical in ascertaining the hyper-dependence zone for $D_{3|1,2}$ and $\alpha_2 \geq 0.040$.

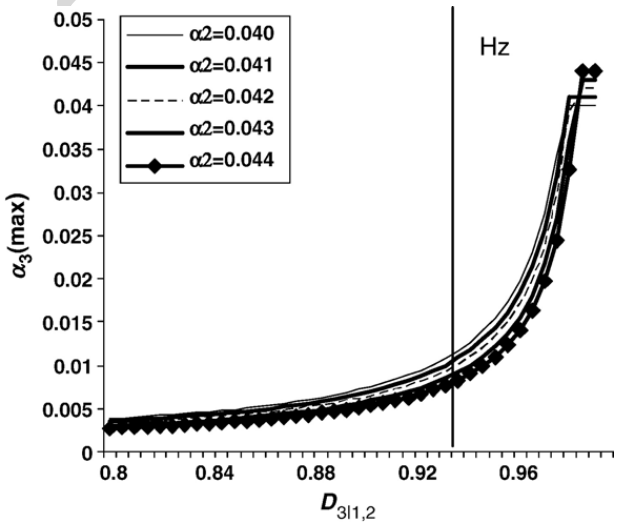

Fig. 2. Relationship between α_3 (max) and $D_{3|1,2}$ as a function of α_2 when the familywise error rate is 0.05 and $\alpha_1 = 0.045$. The region to the right of the line Hz is the zone of hyper-dependence.

Table 1
Alpha allocation in hyper-dependent hypothesis testing example

Primary analyses		Alpha allocation Scenario A	Alpha allocation Scenario B	Alpha allocation Scenario C
	Familywise error rate =	0.05	0.05	0.05
Dependent analysis 1	α_1 =	0.030	0.040	0.045
	$D_{2\|1}$ =	0.40	0.95	0.98
	Alpha error available for α_2 =	0.025	0.040	0.045
Dependent analysis 2	α_2 =	0.024	0.039	0.044
Dependent analysis 3	$D_{3\|1,2}$ =	0.60	0.95	0.99
	Alpha error available for α_3 =	0.001	0.039	0.044
	α_3 =	0.001	0.039	0.044

As an illustration of the use of the methodology presented in this manuscript, consider the circumstance of investigators who have prospectively identified three analyses during the design phase of their randomized, controlled clinical study. The purpose of the trial is to examine the effect of therapy on the total mortality rate of the entire randomized cohort. One of these analyses is to adjust the mortality effect for the recruiting center. A second is a center-adjusted analysis that is also adjusted for a small number of predetermined covariates. The third is an unadjusted analysis. While each of these evaluations will partition random and systematic variability differently, these three evaluations are clearly related to each other. They both evaluate the same endpoint, in the same sample. Even though these analyses are not identical, one would expect (as in the Gliadel example), that the type I error rates that they generate would not be widely separated. The dilemma that faces the investigators is, given that they have declared a maximum type I error level, how can they choose the most appropriate analysis for the situation.

A possible implementation of the methodology proposed in this paper is presented (Table 1).

Table 1 reveals three different schemes for alpha allocation when there are three candidate endpoints. Each scenario conserves the familywise error rate. In Scenario A, the dependency parameters are at moderate levels ($D_{2|1}=0.40$, $D_{3|1,2}=0.60$). In this case, when α_1 is set by the investigators as 0.030, Eq. (12) reveals that a type I error event rate of 0.025 remains for testing analyses 2 and 3. If α_2 is set by the investigators as $\alpha_2=0.024$, then use of Eq. (19) reveals that only 0.001 of the total alpha error rate of 0.05 remains for the third hypothesis test.

Scenarios B and C, operating at hyper-dependency levels of D_{21} and $D_{3|1,2}$, reveal that substantially more conservation of type I error levels can occur. Specifically, scenario C demonstrates that hypothesis testing can effectively occur at the 0.045 level for each of the three hyper-dependent analyses.

4. Discussion

This manuscript addresses the multiple testing issue in clinical trials when evaluations that are being considered for the primary analysis of the study are closely related to each other. The advocated methodology provides a procedure that permits the *a priori* type I error level for each of these analyses to be relatively large, while still conserving the overall type I error rate. Its implementation in the Glidel example would have removed an ambiguity in the interpretation of the statistical results because it would have allowed each analysis to be interpreted at a type I error of between 0.04 and 0.05, removing the investigators to have to prospectively select one analysis from several tightly related ones, each of which would produce similar *p*-values.

The need for investigators to declare their analyses prospectively have been clearly delineated [1,2]. Correction of type I errors for multiple testing is commonly required when there is more than one primary statistical hypothesis. Since this correction typically requires a reduced test-specific type I error threshold, investigators in the design phase of their trial commonly work to identify one and only one primary endpoint for their study. It has been appropriately argued that the type of *post hoc* discussions that occurred in the example provided here can be obviated by a clear *a priori* selection [29]. However, in their attempts to follow this advice, the clinical investigators frequently struggle to collapse a small number of analyses that are admissible as primary analyses down to a single primary evaluation. Their required preliminary discussions for the prospective choice of a single primary analysis are complicated by 1) dependence between endpoint measures, and 2) the lack of a clear clinical rationale to choose one over the other candidate analyses. In the provided example, the standard Kaplan–Meier survival analysis procedure [30], a center stratified log rank test statistic [31], a country-stratified log rank test, and a covariate adjusted analysis [32] were all admissible procedures. Investigators and

even statisticians debate the primacy of one over the other four analyses in a randomized clinical trial, and commonly struggle to reach a consensus. The example provided in the Introduction reveals the confusion that can attend this vexing decision process. While there is commonly useful literature available that can aid investigators in determining when center-adjusted versus unadjusted analyses may be helpful, an issue that has been elaborated in the methodology development of cross over designs e.g., [33–35].

The answer to the question motivating this manuscript, i.e., whether a clinical trial, carried out under the prospective plan of balancing randomization within each of its clinical centers, should execute a stratified analysis on the prospectively declared endpoint, may seems unnecessary to some. They persuasively argue that it is axiomatic in statistics that the efficiency of a research effort is improved when the analysis plan matches the research design. In general, modern health care research designs embed specific features into a research effort, (e.g., the selection of subjects, allocation of therapy, duration of patient follow-up, and choice of the endpoint) with the foreknowledge that the analysis to be executed in the end will be one that incorporates these specific design features into the evaluation. Matching the research design and the subsequent analysis efficiently uses the data to produce the most precise estimate of treatment effects, and among the most powerful hypothesis tests.

This compelling line of reasoning suggests that the best analysis plan in a clinical trial that has chosen to stratify randomization within centers (or within countries) would be a stratified analysis plan. Such a plan successfully and correctly incorporates the within center treatment effect, i.e., the specific effect that the stratified randomization so carefully embedded in the research execution. Applying a stratified analysis to the stratified randomization design produces 1) precise measures of within center treatment effects, 2) a measure of center-adjusted treatment effect, and 3) removal of the center-to-center treatment effect variability from the unexplained variability of the endpoint, producing, *ceteris paribus*, a larger test statistic and a more powerful hypothesis test. In fact, the logic of this approach can appear inescapable. It is this point of view argued by the advocates of Gliadel before the Advisory Committee. Even in the post-hoc light of T-301, this contention did not lose all of its luster.

Advocates of using a non-stratified analysis in the face of a stratified randomization base their argument more on reality than theory. While conceding that the stratified randomization offers a theoretical advantage, they assert that, in order to achieve this advantage in reality, the recruiting centers must randomize sufficient numbers of patients, permitting the within center randomization process to successfully balance therapy allocation within center.

While this is the goal of clinical trials, the experience of many workers suggests that reality is different. Many clinical centers, after promising to recruit a mutually agreed upon minimum number of patients, are simply unable to meet this commitment. Sometimes, these centers can recruit so few patients that stratified randomization is grossly unsuccessful in balancing therapy effect within the study, producing a proportion of clinical centers with unbalanced therapy allocations. Additionally, with lagging recruitment, the trial administration may be compelled to increase the number of centers. This can lead to the inclusion of clinical centers into the research effort that were first excluded either because of inadequate research experience or questions about product quality. Unfortunately, this strategy can exacerbate the problem if these second-tier centers also fail to meet their research quota.

Inadequate planning for this unfortunate eventuality leaves the investigators in the unenviable position of attempting a stratified analysis plan in an unbalanced and unstratified environment, a circumstance that can produce an underpowered analysis [36]. The requirement for a prospective declaration of the endpoint analysis plan precludes the investigators from an *a priori* declaration of a stratified analysis plan, followed by an unstratified analysis if their recruitment plans go awry. Being influenced by such experiences, these workers argue that the research effort is better off by planning for an unstratified analysis. Debates between these two points of view can be vehement, and unsatisfactorily resolved.

The implications of the implementation of the methodology must be clearly elucidated. Carrying out hypothesis testing in the zone of hyper-dependence essentially changes a clinical trial paradigm. Commonly, the persuasive power of a well designed, well executed clinical trial is bolstered when its positive findings for the effect of therapy on a single prospectively declared endpoint are supported by positive findings on other related but secondary endpoints. That persuasive power is vacated when testing occurs among a collection of prospectively declared, hyper-dependent analyses. For example, it is unlikely that the regulatory community would provide three different indications for an intervention that had each of its three prospectively declared analyses on different endpoints identified as statistically significant in the hyper-dependent environment.

Additionally, the value of the dependency parameter has to be determined. While this problem has been discussed in general [5], estimation of the dependency in the setting of multiple analyses of the same endpoint in the same data must be elucidated.

The dependency parameter D, along with the endpoint analysis plans, type I error rates and power must be determined prospectively; its selection is governed by the homogeneity of the treatment effect across centers. The more closely the conditions of the center stratified analysis correspond to those of the non-stratified evaluation, the more informative the one analysis is for the other. This informative condition is produced when the effect of the therapy at each of the centers are close to one another. The measure of the degree to which these are close is the homogeneity of treatment effect. This can be estimated by $D = \frac{v_W}{v_w + v_B}$ where v_W is the within center variability of the treatment effect, and v_B is the between center treatment effect variability. In this setting, large values of D are produced when the between center treatment effect is small relative to the within center treatment effect. What determines the value of D is not endpoint correlation, but the differences between the effect of therapy attained between the unadjusted and center-adjusted effects. Since D must be chosen prospectively, estimates must be available from other research efforts of v_W and v_B. In general, circumstances where the difference in the p-values results in a well conducted, well executed clinical trial is the type of analysis, and the types of analyses are standard and prospectively declared, produce a hyper-dependent environment (i.e., $D > 0.90$).

Specifically, this methodology may also be applied to circumstances where several analysis tools are available, but no analysis tool is particularly preferable. For example, when the investigators are uncertain during the design phase of the research effort whether an analysis adjusted for covariates or unadjusted covariates should be carried out. A protocol may be designed so that three evaluations are prospectively described; 1) an analysis unadjusted for covariates 2) an analysis adjusting for a small number of covariates, and 3) and analysis adjusting for a larger collection of covariates. When all covariates are identified during the design phase, these evaluations are dependent, and the methodology utilized in this manuscript can be applied.

While the usual gatekeeper and fallback procedures e.g., [37] does not explicitly build in dependence, the procedure that we have developed resides within the fallback procedure described in the above paper since it ensures that all hypotheses are carried out while controlling the test specific and overall type I error rate. In addition, the entire development of the use of the dependency parameter occurred within the two-tailed hypothesis testing framework, the one most used in the design of clinical trials, and all hypothesis tests would be carried out using two-tailed p-values. In the one tailed evaluation, the derivation of D and the test specific type I error rates are straightforward in the one-tail testing scenario.

It also must be pointed out that the methodology advocated in this manuscript does not release the investigators from their obligation to declare their evaluations prospectively. The hazards of the alternative philosophy have been clearly delineated [38]. The investigators have the same obligation to choose and defend the value of D as they do for all statistical analysis parameters (e.g., type I and type II error rates, the control group event rate, and the level of efficacy that the clinical trial is designed to detect.). In a regulatory setting, the value of D (and the other aforementioned statistical parameters) would be presented to the regulatory body for discussion during the design phase of the study, giving the sponsor the opportunity to embed the regulators' responses and concerns into their final determination.

Resampling procedures developed by Westfall [39–41] and by Reitmeir and Wassmer [42] have a prominent place in the methodologic literature evaluating the multiple analysis issue. However neither the sequential rejective procedures nor the resampling evaluations allow the investigators to select the type I error rate for each of the small number of primary statistical evaluations. Alternative procedures that permit investigators to carry out multiple analyses on primary endpoints and allow the scientists to retain control of the alpha level threshold have been discussed in the literature [5,43,44] and clinical trials are beginning to develop experience with these procedures currently use related procedures [45]. In fact, clinical trials can carry this out by inducing modest measures of dependency [46]. The work of Benyamini and colleagues [47,48] focusing on the false discovery rate allows workers to control the proportion of hypothesis tests that are false positive. This perspective, which has gained support in the neurosciences, does not lead to tight control of the familywise error rate, and is less applicable in the setting where the primary concern is the control of the overall type I error rate. The evaluations provided in the manuscript do not provide for the type I error rate adjustment that must take place during prospectively planned interim monitoring procedures. However, the *a priori* alpha levels selected would be used as the familywise alpha error rate in computing the boundary values for considering early stopping consideration, using either a Lan and DeMets [49–51] or a conditional power argument [52].

The standard approach of using correlations between the three log-rank tests and adjusting the error rates accordingly is problematic since it requires the correlation structure under the null hypothesis, and that this correlation structure be used to compute a relatively complicated probability, i.e., $P[L_3 \geq 1.96 \mid L_1 \geq 1.96 \cap L_2 \geq 1.96]$ where L_i, $i = 1, 2, 3$ be the three log rank statistics. Given the value of D, it is easy to adjust the type I error rates.

There are potential regulatory implications for the interpretation of a positive trial in the hyper-dependence environment. It is the indication section of the label that describes the benefits of the drug that the FDA and the sponsor reasonably believed would occur in those patients who use the drug as directed. Many times the sponsors of a new intervention express great interest in gaining as many approved indications for its use as possible. This is, in fact, one motivation for implementing a prospectively planned multiple primary analysis mechanism in the design of clinical trials. However, the relevant Code of Federal Regulations (CFR) requires that each indication "shall be supported by substantial evidence of effectiveness". Hyper-dependence among primary analyses would undermine any claim that each of the primary analyses provides substantial evidence of effectiveness. It is therefore difficult to envision that the FDA would provide an indication for each of the positive findings among each of prospectively defined primary, but hyper-dependent analyses produced from a clinical trial. In the case examined in this manuscript, the sponsor could only hope for one indication for the use of therapy.

An important weakness of this approach is the requirement of an accurate selection of D. Overestimation of its value can produced inappropriately high type I error thresholds. To some extent, however, this is offset by the use of D^2 in the type I error computations. In addition, one cannot help but wonder whether the FDA and its advisory committee wasn't "hyper-dependent" on p-values. Clearly the interpretation of a clinical research effort does not turn on the p-value alone, but on the joint consideration of the research effort's design and execution, effect size and effect size variability. In the end, this may have been one the Advisory Committee ultimately considered when they voted that, although T-301 did not reach statistical significance on the per protocol analysis, they voted to approve the compound anyway.

This methodology offered in this manuscript permits the use of multiple, hyper-dependent endpoints in clinical trials in a way that is consistent with the identification of confirmatory results in sample-based clinical research. Additionally, its prospective use in clinical research may permit the standardization of dependent hypothesis testing in the clinical trial community, permitting different levels of dependence for different classes of hypothesis tests. The implications of such a standardization require further investigation.

References

[1] Meinert CL. Clinical trials: design, conduct, and analysis. New York: Oxford University Press; 1986.
[2] Friedman L, Furberg C, DeMets D. Fundamentals of clinical trials. 3rd edition. New York: Springer; 1986.
[3] Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. J Am Med Assoc 1991;266:93–8.
[4] The European Agency for the Evaluation of Medicinal Products. ICH topic E9. Note for guidance on statistical principles in clinical trials. CMPM/ICH/363/96; 1998.
[5] Moyé L. Multiple analyses in clinical trials: fundamentals for investigators. New York: Springer; 2003.
[6] Sacks FM, Pfeffer MA, Moyé LA, et al. The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. N Engl J Med 1996;335:1001–9.
[7] , The HOPE Study Investigators. The HOPE (Heart Outcomes Prevention Evaluation) Study: The design of a large, simple randomized trial of an angiotensin-converting enzyme inhibitor (ramipril) and vitamin E in patients at high risk of cardiovascular events. Can J Cardiol 1996;12:127–37.
[8] Miller RG. Developments in multiple comparisons 1966–1976. J Am Stat Assoc 1977;72:779–88.
[9] Tukey JW, Ciminera JL, Heyse JF. Testing the statistical certainty of a response to increasing doses of a drug. Biometrics 1985;41:295–301.
[10] Dubey SD. Adjustment of p-values for multiplicities of intercorrelating symptoms. Proceedings of the VIth international society for clinical biostatisticians, Germany; 1985.
[11] O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984;40:1079–89.
[12] Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat Med 1997;16:2229–42.
[13] Hochberg Y, Westfall PH. On some multiplicity problems and multiple comparison procedures in biostatistics. In: Sen PK, Rao CR, editors. Handbook of Statistics, vol. 18. Elsevier Sciences B.B.; 2000. p. 75–113.
[14] James S. Approximate multinomial probabilities applied to correlated multiple endpoints in clinical trials. Stat Med 1991:1123–35.
[15] Neuhauser M, Steinijans VW, Bretz F. The evaluation of multiple clinical endpoints with application to asthma. Drug Inf J 1999;33:471–7.
[16] Reitmeir P, Wassmer G. One sided multiple endpoints testing in two-sample comparisons. Commun Stat Simul Comput 1996;25:99–117.
[17] Westfall PH, Ho SY, Prillaman BA. Properties of multiple intersection-union tests for multiple endpoints in combination therapy trials. J Biopharm Stat 2001;11:125–38.
[18] Hochberg Y, Liberman U. An extended Simes' test. Stat Probab Lett 1994;21:101–5.
[19] Zhang J, Qwuan H, Ng J, Stepanavage ME. Some statistical methods for multiple endpoints in clinical trials. Control Clin Trials 1997;18:204–21.
[20] Wright SP. Adjusted P-values for simultaneous inference. Biometrics 1992;48:1005–13.
[21] Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986;73:819–27.
[22] Holm S. A simple sequentially rejective multiple test procedures. Scand J Statist 1979;6:65–70.

[23] Hommel G. A stepwise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 1988;75:383–6.
[24] Shaffer JP. Modified sequentially rejective multiple test procedures. J Am Stat Assoc 1986;81:826–31.
[25] Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research. Transcript of the Oncologic Drug Advisory Committee Sixty-Ninth Meeting, December 6, 2001.
[26] Questions to the FDA Oncology Drugs Advisory Committee. December 6, 2001.
[27] Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley; 1987.
[28] Westfall PH, Young SS. Resampling based multiple testing: examples and methods for *P*-value adjustment. New York: Wiley; 1993.
[29] Moyé LA. *P*-value interpretation in clinical trials. The case for discipline. Control Clin Trials 1999;20:40–9.
[30] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: John Wiley; 1980.
[31] Kleinbaum DG, Kupper LL, Morgenstern H. Epidemdiologic research: principles and quantitative methods. New York: Van Nostrand Reinhold Company; 1982.
[32] Cox DR. Regression models and life-tables. JR Stat Soc B 1972;34:187–220.
[33] Grizzle JE. The two-period change-over design and its use in clinical trials. Biometrics 1965;21:467–80.
[34] Armitage P, Hills M. The two-period crossover trial. The Statistician 1982;32:119–31.
[35] Senn S. Cross-over trials in clinical research. Chicester, New York: John Wiley and Sons; 1993.
[36] Lin Z. The number of centers in a multicenter clinical study: effects on statistical power. Drug Inf J 2000;54:379–96.
[37] Wiens B, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. J Biopharm Stat 2005;15:929–42.
[38] Moyé LA. The perils of nonprospectively planned research. Part 1: drawing conclusions from sample-based research. Am Clin Lab: April 2001 2001:34–6.
[39] Westfall PH, Young SS, Wright SP. Adjusting *p*-values for multiplicity. Biometrics 1993;49:941–5.
[40] Westfall PH, Young S. *P*-value adjustments for multiple tests in multivariate binomial models. J Am Stat Assoc 1989;84:780–6.
[41] Westfall PH, Krishnen A, Young SS. Using prior information to allocate significance levels for multiple endpoints. Stat Med 1998;17:2107–19.
[42] Reitmeir P, Wassmer G. Resampling-based methods for the analysis of multiple endpoints in clinical trials. Stat Med 1999;18:3453–62.
[43] Dunnett CW, Tamhane AC. Multiple testing to establish superiority/equivalence of a new treatment compared with k standard treatments. Stat Med 1997;16:2489–506.
[44] Cheung SH, Holland B. Extension of Dunnett's multiple comparison procedure to the case of several groups. Biometrics 1991;47:21–32.
[45] Davis BR, Cutler JA, Gordon DJ, et al. Rationale and design for the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). Am J Hypertens 1996;9:342–60.
[46] The Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial Investigators. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without st-segment elevation. N Engl J Med 2001;345:494–502.
[47] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy Stat Soc, Ser B. 57:289–300.
[48] Y. Benjamini, D. Yekutieli. The Control of the false discovery rate in multiple testing under dependency. Tech Report, School of Mathematical Sciences, Tel Aviv University. April 27, 2001. Available at http://www.math.tau.ac.il/~benja/.
[49] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659–63.
[50] Lan KKG, DeMets DL. Changing frequency of interim analysis in sequential monitoring. Biometrics 1989;45:1017–20.
[51] Lan KKG, DeMets DL. Group sequential procedures. Calendar vs. information time. Stat Med 1989;8:1191–8.
[52] Lan KK, Wittes J. The B-value. A tool for monitoring data. Biometrics 1988;44:579–85.