Drug Information Journal
http://dij.sagepub.com/

A Statistical Framework for Decision Making in Confirmatory Multipopulation Tailoring Clinical Trials

Brian A. Millen, Alex Dmitrienko, Stephen Ruberg and Lei Shen

Drug Information Journal published online 6 August 2012
DOI: 10.1177/0092861512454116

The online version of this article can be found at:
http://dij.sagepub.com/content/early/2012/08/01/0092861512454116

Published by:
SAGE
http://www.sagepublications.com

On behalf of:

Drug Information Association

Drug Information Journal
00(0) 1-10

DOI: 10.1177/0092861512454116
http://dij.sagepub.com

A Statistical Framework for Decision Making in Confirmatory Multipopulation Tailoring Clinical Trials

Brian A. Millen, PhD[1], Alex Dmitrienko, PhD[2], Stephen Ruberg, PhD[1], and Lei Shen, PhD[1]

Abstract

This article focuses on statistical analysis of clinical trials pursuing tailored therapy objectives, wherein evaluation of treatment effect occurs in the overall population as well as in a predefined subpopulation(s). The design and analysis principles presented provide a framework for decision making based on these novel multipopulation tailoring trial designs, considering the particular case of confirmatory trials. These principles include traditional multiple testing considerations, as well as 2 new analysis principles.

Keywords

subgroup analysis, tailored therapy, type I error rate, influence condition, interaction condition

1. Introduction

The promise of tailored therapeutics and personalized medicine has resulted in increased attention on evaluation of treatment effects in focused subpopulations in clinical trials. The subpopulations of interest may be defined by demographics, clinical markers, genetic markers, or a combination of these. Although subgroup analyses that explore treatment effects in subpopulations (defined by demographics and other characteristics) has been a standard part of clinical trial analysis plans for decades, the intent of those analyses historically had not been to test a priori hypotheses regarding treatment effect in the subpopulation(s). Instead, these analyses were exploratory and hypothesis generating in nature; the inference set for these trials was the overall patient population only.

As interest in focused subpopulations advanced, single population tailoring trials became more common. In these trials, evaluation of treatment effect in a targeted subpopulation was the primary objective, and so-called enrichment designs[1] were employed. The Herceptin program[2] provides an early and prominent example. More recent examples include the Xalkori registration trials,[3] the Zelboraf registration trials,[4] and the Alimta program for nonsquamous non–small cell lung cancer.[5] At times, such trials were undertaken only after negative results were obtained in overall population trials and the subpopulation was hypothesized based on exploratory subgroup analyses from the negative trial(s). In other cases, overall population trials were omitted altogether in favor of target subpopulation trials. In these cases, treatment effect in the remainder of the population is unknown.

Today, clinical trials with more complex objectives, such as evaluating treatment effects in focused subpopulations, as well as in the broader overall population, are conducted to realize the promise of tailored therapeutics. These *multipopulation* tailoring trials offer several advantages over the single population trials. As these trials provide inference of treatment effect for multiple (overlapping) populations within a single trial, they are more efficient than the traditional approach of conducting multiple single population trials. In addition, as the trials are prospectively designed to allow multiple population inference, these trials are more informative regarding treatment effect than traditional single population trials. Because of their efficiency and richness of information, trial sponsors are

[1] Eli Lilly and Company, Indianapolis, IN, USA
[2] Quintiles, Overland Park, KS, USA

Portions of this work were presented at the 2011 DIA/FDA Tailored Therapeutics Conference (Bethesda, MD, USA) and at the 2011 Annual DIA Meeting (Chicago, IL, USA).

Submitted 09-February-2012; accepted 06-Jun-2012.

Corresponding Author:
Brian A. Millen, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA
Email: millen_brian_a@lilly.com

Table 1. Overview of multiple testing for various types of clinical trials.

Trial Description	Trial's Primary Objective	Multiple Testing Methods (for populations)
Traditional trial	Demonstrate effect of treatment in overall (broad) patient population	N/A
Single population trial (subpopulation)	Demonstrate effect of treatment in a focused (predefined) subpopulation	N/A, or Fixed sequence procedure. The test of treatment effect in the overall population is a secondary objective (ie, pursued only after primary objective is met).
Multipopulation tailoring trial	Demonstrate effect of treatment in the overall population and/or a predefined subpopulation	Nonparametric and Parametric procedures included in Table 2-2.

increasingly conducting multipopulation trials. The SATURN trial[6] is one example.

Multipopulation tailoring trials are the focus of this article. Specifically, we focus on confirmatory clinical trials in which the primary inference is of treatment effect for the overall population as well as a prospectively defined subpopulation. Without loss of generality, we assume the primary inference is of treatment efficacy. In these trials, 3 distinct positive inferences may result: efficacy in the overall population, efficacy in the predefined subpopulation (with insufficient evidence to conclude efficacy in the overall population), or efficacy in both the overall population and the predefined subpopulation. Each of these inferential outcomes provides the foundation for registration of a new treatment in the appropriate population(s). In addition, information regarding heterogeneity of effect across the overall population is available from the trials.

Given the novelty of such clinical trial and drug registration approaches, there have been few publications addressing the unique statistical considerations for these trials, particularly with a focus on registration of a new treatment. Moreover, alignment on appropriate statistical methods for the design and analysis of such trials is lacking. Wang, O'Neill, and Hung[7] and Wang, Hung, and O'Neill[8] discussed general principles of design and analysis of clinical trials with a predefined subpopulation. Song and Chi[9] and Alosh and Huque[10] presented consistency-ensured multiple testing methods for primary inference from such trials. Zhao, Dmitrienko, and Tamura[11] introduced a more complete treatment with design and analysis considerations, including a novel multiple testing method. Simon and his colleagues offered landmark papers on signature adaptive design and related designs.[12,13] Nonetheless, more discourse is needed, particularly with respect to regulatory approval considerations based on statistical results. This has been a central theme of several conferences and regulatory (eg, FDA and EMA) presentations, including the EMA Expert Workshop on Subgroup Analysis held in November 2011.

In this article, we present a statistical framework to support decision making in clinical trials wherein the primary objective is to assess the effect of treatment in both the overall population and, without loss of generality, a (single) prespecified subgroup. We introduce key, broadly applicable principles that need to be considered when drawing inference from such trials, along with the associated statistical analysis and design considerations. In section 2, we present multiplicity considerations. In sections 3 and 4, we introduce the influence and interaction conditions, respectively. In section 5, we provide decision principles based on the analyses outlined in sections 2 through 4. In section 6, we briefly discuss associated design considerations. We provide an illustrative example in section 7, and we close with a general discussion in section 8.

2. Multiplicity Considerations

Multipopulation tailoring trials give rise to multiplicity considerations because of the fact that the trial's sponsor is given multiple opportunities to claim a significant treatment effect. That is, the sponsor can claim a treatment benefit in the overall population and, independent of the outcome in the overall population, a "win" can be claimed in the subpopulation. In order to address the associated inflation of the type I error rate because of multiple testing, the sponsor needs to prospectively define the family of null hypotheses to be tested and a multiplicity adjustment. The null hypotheses of interest include the null hypothesis of no effect in the overall population and the null hypothesis of no effect in the subpopulation (see Table 1).

The objective of a multiplicity adjustment is to protect the familywise error rate in the strong sense,[14] which subsequently enables the sponsor to make conclusions about the treatment's efficacy in the overall population and subpopulation. Several methods for performing multiplicity adjustments may be employed in this setting. Examples include Bonferroni-based and parametric fallback and related procedures,[15-17] Bonferroni-based and parametric chain procedures,[18-20]

Table 2. Basic considerations for implementation of available multiple testing procedures in multipopulation tailoring trials.

Multiple Testing Methodology	Specify Ordered Testing Sequence?	Specify α Allocation?	Incorporates α Propagation?	Comments
Simple nonparametric procedures				
Bonferroni	N	Y	N	Lack of α propagation reduces power.
Hochberg and Hommel	N	Y	N	Uniformly more powerful than Bonferroni procedure but lack flexibility.
Fallback and chain (Bonferroni based)	Y	Y	Y	Uniformly more powerful than Bonferroni procedure (reject all hypotheses of Bonferroni procedure and potentially more). Chain procedures increase flexibility through researcher-defined α-propagation rules.
More powerful parametric procedures				
Fallback and chain (parametric)	Y	Y	Y	Increased power over Bonferroni-based procedures. Leverage known correlations of test statistics for the overlapping populations.
Consistency-ensured fallback	Y	Y	Y	Adds constraint to fallback so that subpopulation inference is dependent on overall population result.
Feedback	Y	Y	Y	Increased power over Bonferroni-based procedures (correlations of test statistics for the overlapping populations are taken into account).

Table 3. Summary of the potential outcomes from the primary hypothesis tests in a multipopulation tailoring trial.

	Outcome of Primary Hypothesis Tests			
	Outcome 1	Outcome 2	Outcome 3	Outcome 4[a]
Statistical Significance: Overall Population	Yes	No	Yes	No
Statistical Significance: Predefined Subpopulation	No	Yes	Yes	No

[a]No conclusions drawn in the case of outcome 4.

feedback procedure,[11] and general Bonferroni-based adjustments, as considered by Freidlin and Simon[12] and others. While each of the procedures provides the opportunity to test effect in each population, they differ in some key characteristics which must be considered prior to choosing the methodology to employ. A detailed review of these methods is out of scope for this article; however, a high-level comparison of the methods must include the following: each of the methods requires distributing the allowable type I error rate (alpha) among the hypotheses corresponding to the populations of interest; the methods differ in whether, or how, results of one test affect the other (ie, alpha propagation rules) and whether testing occurs in a prespecified sequence (see Table 2).

As a result, inference will differ with the selection of the procedure. This is true of multiple testing, in general, and is not limited to tailoring trials. The example provided in section 4 uses a fallback procedure. As a result of testing the null hypotheses of no effect in the overall population and predefined subpopulation, 3 primary conclusions are possible (see Table 3):

- Outcome 1: Beneficial effect only in the overall population.
- Outcome 2: Beneficial effect only in the predefined subpopulation, with insufficient evidence to conclude efficacy in the overall population.
- Outcome 3: Beneficial effect in both populations.

(Note: the fourth possible outcome is failure to reject both null hypotheses [corresponding to the overall population and the predefined subpopulation]. Strictly speaking, no conclusion is drawn in this case.)

A few key questions may remain beyond the primary conclusions from the multiple testing method, which are important for sponsor and regulatory decision making. For example, with

Table 4. Assessment of the influence condition.

Population	Relative Size (%)	Effect Size
Scenario 1: Influence condition is not satisfied		
Predefined subpopulation	60	1.0
Complementary subpopulation	40	–0.25
Overall population	100	0.5
Scenario 2: Influence condition is satisfied		
Predefined subpopulation	30	1.0
Complementary subpopulation	70	0.285
Overall population	100	0.5

outcome 3 above, should the treatment be available for the overall population (with simple broad indication labeling), or for the predefined subpopulation only, or available for the overall population with labeling detailing enhanced effect seen in the predefined subpopulation? To facilitate decision making beyond the primary outcomes noted above, assessment of the *influence* and *interaction* conditions is needed. These 2 important analytical conditions are introduced in the next 2 sections. Throughout, the following notation will be used: O = overall population, A = predefined subpopulation, and A^c = complementary subpopulation; $O = A \cup A^c$.

3. Influence Condition

The influence condition states that to enable overall population labeling, the beneficial effect of treatment must not be limited to only the predefined subpopulation.

The influence condition is a "natural" requirement to support a broad (overall population) indication based on a tailoring clinical trial with a positive result for the predefined subpopulation as well as the overall population. When it is not satisfied, the subpopulation results unduly influence the overall population inference. To illustrate this point, consider the hypothetical examples presented in Table 4. In scenario 1, a true overall population effect size is positive, although the true effect size in the complementary subpopulation is negative (ie, there is a detrimental effect of treatment) and the true effect size in the predefined subpopulation is positive. (Here, effect size is defined as the ratio of the mean treatment difference to the common standard deviation.) In this case the effect size in the overall population is positive, equal to 0.5, solely because of the influence of the predefined subpopulation. In this example, the influence condition is not satisfied. Despite the positive average effect size in the overall population, the true population corresponding to positive effect is the predefined subpopulation. In this case, membership in the predefined subpopulation may be used in determining appropriate candidates for the treatment.

Scenario 2 in Table 4 presents another hypothetical example. In this example the predefined subpopulation makes up 30% of the overall population and the effect sizes in the target and complementary subpopulations are given by 1.0 and 0.285, respectively. The effect size in the overall population is 0.5 and overall positive effect is supported by positive effect sizes in each subpopulation, which means that the influence condition is satisfied.

It is important to note that the influence condition does not require comparable effects in the predefined subpopulation (A) and its complement (A^c). The condition simply ensures that a positive effect in a subpopulation does not mask a negative effect in the complementary subpopulation. In this sense, the influence condition may be thought of as a restriction against qualitative interaction. A qualitative interaction is present if the true treatment difference is positive in the predefined subpopulation and negative in the complementary subpopulation.

In practice true effect sizes, like the ones provided in Table 4, are unknown. Thus, it is important to develop reasonable methods by which to assess the influence condition. We present a discussion of a few possibilities below along with our recommendations of appropriate approaches.

A straightforward approach to evaluation of the influence condition may be found in testing for qualitative interaction, for example, by using the Gail-Simon qualitative interaction test.[21] A significant result by a qualitative interaction test would provide evidence of violation of the influence condition. While this testing approach directly addresses the question of interest, the power is likely to be low.[22] Furthermore, the related scenario of a positive effect in the predefined subpopulation and a zero effect in the complementary subpopulation is not directly addressed by this testing approach.

An alternative, yet extreme (conservative), statistical approach for assessing the influence condition is simple hypothesis testing for efficacy in the predefined subpopulation and its complement. A statistically significant treatment effect in both the target and complementary subpopulations would clearly satisfy the influence condition. However, failure to attain a significant treatment difference in the complementary subpopulation may reflect only an underpowered test. In fact, one would expect this test to be underpowered in a typical clinical trial. Given these considerations, this hypothesis testing-based approach for assessment of the influence condition is not recommended.

Estimation-based approaches offer an alternative to approaches based on hypothesis testing for assessment of the influence condition. Rather than resulting in P values and corresponding decision rules, these approaches provide estimates as information to aid regulators and sponsors in the evaluation of the influence condition. Below, we formulate 2 estimation-based approaches to enable assessment of the influence condition. One approach uses frequentist estimation, while the other uses Bayesian estimation.

To introduce the estimation-based approaches, let $\delta(A)$ and $\delta(A^c)$ denote the true effect sizes for A and A^c, respectively. Their sample estimates are denoted by $d(A)$ and $d(A^c)$, respectively. For continuous endpoints, the effect size is defined as the ratio of the mean treatment difference to the common standard deviation. For binary endpoints, the effect size is defined as the ratio of the difference in proportions to the standard deviation. For time-to-event endpoints, the effect size is defined as the log-hazard ratio between the treatment groups.

According to the frequentist approach, the influence condition is satisfied if $d(A^c) > e$, where e is a nonnegative constant, which represents a minimal threshold of clinical relevance. Note that this constant defines a population threshold of minimal clinical relevance, which is generally less than the threshold of minimal clinical relevance in an individual, because of between-subject variability in the population. While the inherent uncertainty associated with point estimates and their use for decision making is evident, this may be accounted for in the trial design stage. That is, an additional criterion for sample size determination based on the influence criterion may be, for example, $\Pr(d(A^c) > e) \geq 80\%$. In addition, confidence intervals may be reported.

An alternative to simple frequentist estimation is computation of a Bayesian posterior probability to quantify the support for the influence condition. For example, one may compute the posterior probability that the true effect size in the complementary subpopulation is greater than e, that is, $\Pr(\delta(A^c) > e)$. Higher probabilities demonstrate greater likelihood, given the data, that a positive treatment effect exists in the complementary subpopulation and that the influence condition is satisfied. The posterior probability and, furthermore, a summary of the entire posterior probability distribution of the effect size in the complementary subpopulation would provide regulators and sponsors with sufficient information to evaluate the influence condition.

We advocate estimation-based approaches for assessment of the influence condition. In particular, the Bayesian posterior estimation is attractive to support decision making based on the influence condition.

4. Interaction Condition

The interaction condition states that to support enhanced labeling for the predefined subpopulation (A), the treatment effect in the predefined subpopulation (ie, $\delta(A)$) should be appreciably greater than the treatment effect in the complementary subpopulation (ie, $\delta(A^c)$).

The interaction condition plays a key role in regulatory decisions to warrant a broad population indication with enhanced labeling for a predefined subpopulation. If, indeed, the effect in the subpopulation (A) is comparable to that of the complementary subpopulation (A^c), the additional result does not provide information that is valuable to the end users (patients and prescribers) and there is no need to report information at the subpopulation level. As with the influence condition, both hypothesis testing-based and estimation-based approaches can be used for assessment of the interaction condition. We advocate addressing the assessment of the interaction condition as an estimation problem.

An obvious way to assess the interaction condition from a hypothesis testing perspective is to rely on traditional statistical tests of quantitative interaction, for example, ANOVA-based interaction tests for normally distributed variables or the Breslow-Day test for binary variables. One may declare the interaction condition satisfied if the P value obtained from a statistical test of the treatment-by-marker interaction is significant (assuming appropriate directionality). However, this approach is not recommended because of power concerns and the associated potential to assume homogeneity when diseases and individuals are generally not homogeneous.

Another approach that has been advocated by others is to compare treatment effects in the predefined and complementary subpopulations by comparing their treatment effect P values. A more significant P value in the predefined subpopulation may provide evidence of enhanced effect in the predefined subpopulation compared to the overall population or complementary subpopulation. However, such comparisons are problematic, since P values are functions of sample size and, thus, populations with very different effect sizes may have comparable P values purely because of the sample sizes involved. Moreover, use of P values in this context is complicated by the choice of the multiple testing procedure applied to the trial.

An alternative, estimation-based approach for assessing the interaction condition is via direct comparison of effect size estimates in the target and complementary subpopulations. In particular, assessment of the interaction condition can be based on the ratio of the estimated effect sizes in the 2 subpopulations. Using the notation introduced in section 3, the interaction condition is satisfied when $d(A) / d(A^c) > \lambda$, where λ is a prespecified constant greater than 1. This constant defines a lower bound on the magnitude of improvement in the predefined subpopulation (A) to the complementary subpopulation (A^c) that would be clinically relevant. Furthermore, likelihoods associated with the estimation-based evaluation of the interaction condition may be assessed at the study design stage. For instance, the sample size in the predefined subpopulation may be chosen to ensure that the relative criterion is satisfied with a prespecified probability, for example, $\Pr(d(A) / d(A^c) > \lambda) \geq 80\%$.

Finally, as in section 3, a Bayesian approach can be used to quantify the likelihood, given the available data, that the effect size in the predefined subpopulation is greater than

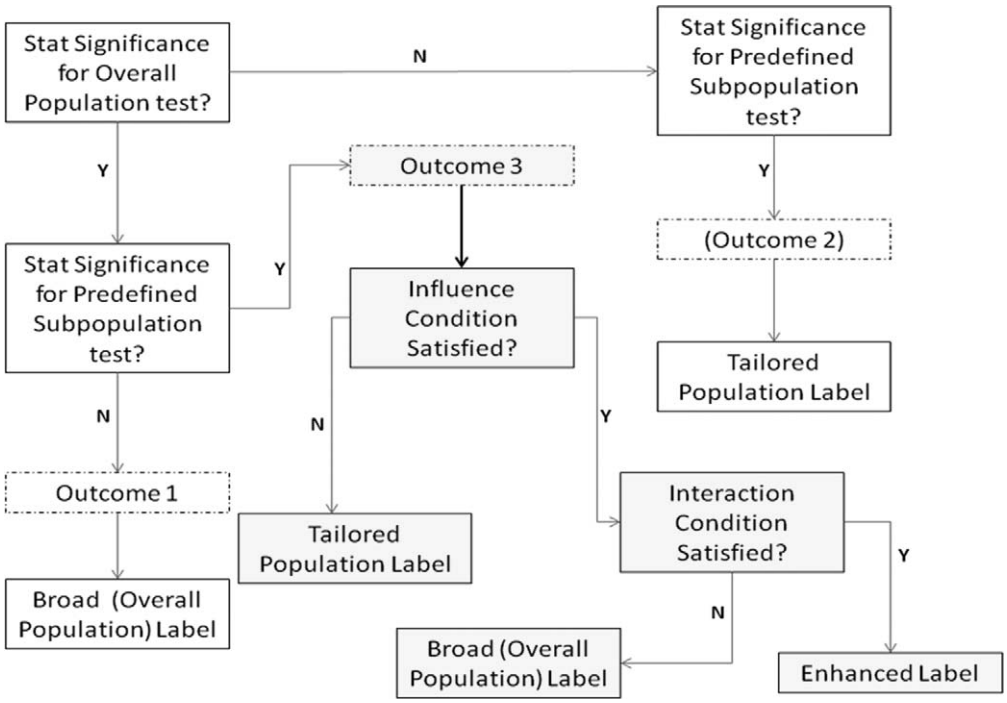

Figure 1. The decision-making process in clinical trials with tailoring objectives based on the influence and interaction conditions.

the effect size in the complementary subpopulation or that $\delta(A) / \delta(A^c) > \lambda$. If the posterior probability of this outcome is sufficiently high, this will provide support that the interaction condition is satisfied.

5 Decision Rules

To facilitate the process of developing decision rules based on the interaction and influence conditions, consider first the case when a significant treatment effect is present only in the predefined subpopulation. In this case there is clearly no need to assess the influence condition, and, similarly, assessment of the interaction condition becomes irrelevant. Inference from the trial is limited to the predefined subpopulation. Of course, as in any trial, exploratory subgroup analyses within the predefined subpopulation may be conducted for hypothesis generation and consistency assessments. Similarly, if there is a significant treatment effect only in the overall population (ie, outcome 1), exploratory subgroup analyses may be conducted. In particular, an assessment of the influence condition may be considered, although the decision pathway is less clear in this case.

The interaction and influence conditions are both relevant in the case of joint primary inference of effect in the overall and predefined subpopulations (ie, outcome 3). The decision diagram presented in Figure 1 provides suggested conceptual decision guidelines for this scenario, based on outcomes of the analyses in the overall and predefined subpopulation. Beginning with the influence condition, if this condition is not satisfied, the beneficial treatment effect is limited to only the predefined subpopulation, and it is natural to consider a simple label that reflects the tailored indication. If the influence condition is met, the interaction condition is assessed next. If the treatment effect in the predefined subpopulation is comparable to that in the overall population, the interaction condition is not satisfied and thus the broad indication (overall population) would be appropriate. On the other hand, if the treatment provides a substantial improvement in the predefined subpopulation compared to the overall population, it is sensible to consider an enhanced label with the broad indication and additional labeling that reflects the enhanced effect in the predefined subpopulation.

To illustrate the decision-making process in a clinical trial with tailoring objectives, consider the 3 scenarios presented in Figure 2. The trial was conducted to evaluate the efficacy of a single dose of a new treatment versus a placebo. Scenario 1 represents the case when a positive treatment effect is present in the predefined subpopulation but there is no treatment benefit in the complementary subpopulation. In this case the influence condition is not satisfied and the tailored indication is considered. Furthermore, the same magnitude of beneficial treatment effect is observed in the 2 subpopulations in scenario 2, which implies that the influence condition is satisfied. However, it is clear that the interaction condition is not satisfied, and

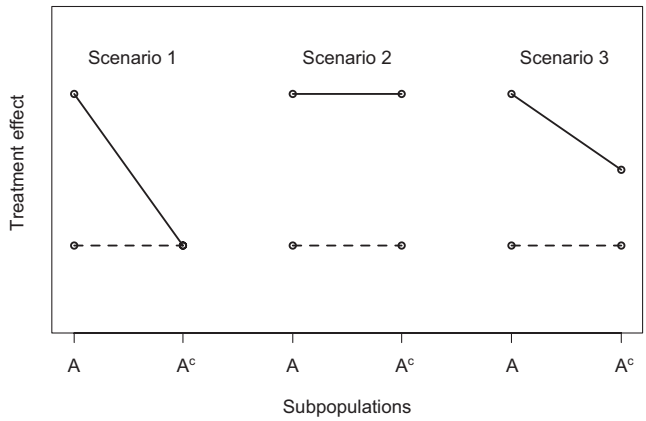

Figure 2. Assessment of the influence and interaction conditions in 3 scenarios. Treatment groups: solid line, active treatment; dashed line, placebo. Subpopulations: A, predefined subpopulation; A^c, complementary subpopulation. Outcomes: scenario 1, the influence condition is not satisfied and the interaction condition is satisfied; scenario 2, the influence condition is satisfied and the interaction condition is not satisfied; scenario 3, both conditions are satisfied.

there is no need to report the subpopulation results in the label. The last scenario serves as an example of the trial outcome with both conditions satisfied. An enhanced label will appropriately capture a positive effect in the overall population and an additional treatment benefit in the predefined subpopulation in this scenario.

6. Design Considerations

Determination and justification of sample size is a primary statistical design consideration for any trial. Although the sample size determination may be based on power for the primary endpoint for the overall population in a traditional trial, the sample size determination for the multipopulation tailoring trial will need to incorporate the additional complexity of the primary objective. At the simplest level, the clinical trial sponsor may wish to ensure adequate marginal power for the treatment effect tests in the overall population as well as the predefined subpopulation. Alternatively, the sponsor may wish to optimize the probability of detecting a significant effect in either or both populations, or the probability of obtaining a significant treatment effect in the predefined subpopulation when the overall population result is negative. A discussion of available metrics may be found in Millen and Dmitrienko.[18]

Note that sample sizes based solely on probabilities of significant results, no matter how complex, may not sufficiently meet all the needs of the trial or may result in inefficient trial designs. Attention must be paid to ensure adequate sample size to satisfy assessment of the influence and interaction conditions, if necessary. That is, the sponsor should evaluate and attempt to optimize the probabilities of satisfying the 2 conditions, for example, $\Pr(d(A^c) > e)$ and $\Pr(d(A) / d(A^c) > \lambda)$, with given sample sizes. All of these metrics may be readily evaluated via simulation, using minimal assumptions, enabling appropriate sizing of clinical trials with tailoring objectives.

In addition, when appropriate, potential enrichment of the predefined subpopulation needs to be taken into account. For example, if the required relative size of the subpopulation is 20% of the trial but the prevalence of patients included in the subpopulation is only 10%, the sponsor can utilize an enrichment study design in which the subpopulation is "oversampled" for inclusion in the trial. See, for example, Zhao, Dmitrienko, and Tamura.[11]

7. Clinical Trial Example

The design considerations presented in section 6 will be illustrated using an example based on the SATURN trial.[18] Consider a phase III clinical trial in patients with advanced non–small cell lung cancer. This trial will be conducted to evaluate the efficacy and safety profiles of a single dose of the new treatment compared to placebo in the overall population as well as a prospectively defined subpopulation of patients. The predefined subpopulation consists of EGFR (epidermal growth factor receptor) immunohistochemistry-positive patients. The relative size of the predefined subpopulation is expected to be 60%. The primary endpoint is progression-free survival (PFS).

The sponsor is planning to pursue regulatory claims in the overall population as well as the predefined subpopulation and a multiplicity adjustment based on the fallback procedure will be performed. The weights of the overall population and predefined subpopulation in the fallback procedure will be set to 0.8 and 0.2. In other words, the treatment effect will be first tested in the overall population at a 2-sided .04 level. If the treatment effect in the overall population is significant, the test in the predefined subpopulation will be carried out at a 2-sided .05 level; a 2-sided .01 level will be used otherwise.

Figures 3 and 4 display the power curves under the assumption that the hazard ratio in the complementary subpopulation is 0.8 and the hazard ratio in the predefined subpopulation is 0.65 or 0.7. Power of the PFS analysis is computed as the probability to detect a significant treatment effect in the overall population, predefined subpopulation, and either overall population or predefined subpopulation. Figure 3 shows that the total sample size of 400 patients provides 90% probability of demonstrating a statistically significant improvement in PFS in the overall population or predefined subpopulation when the hazard ratio in the predefined subpopulation is 0.7. The total sample size needs to be increased to 450 patients to achieve around 80% power in the predefined subpopulation. As follows from Figure 4, under a more optimistic assumption of a 0.65

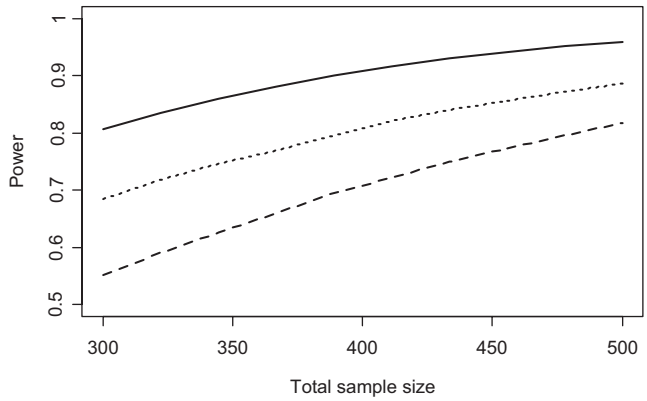

Figure 3. Probability of detecting a significant treatment effect in the overall population (dashed curve), predefined subpopulation (dotted curve) and either overall population or predefined subpopulation (solid curve). Hazard ratios in the target and complementary subpopulations are 0.7 and 0.8, respectively.

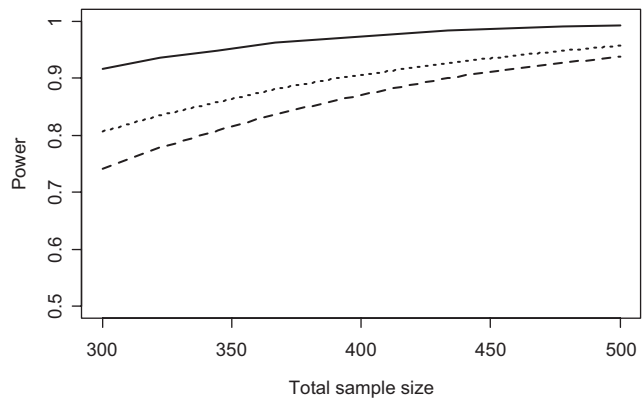

Figure 4. Probability of detecting a significant treatment effect in the overall population (dashed curve), predefined subpopulation (dotted curve) and either overall population or predefined subpopulation (solid curve). Hazard ratios in the target and complementary subpopulations are 0.65 and 0.8, respectively.

hazard ratio in the predefined subpopulation, the probability of a statistically significant effect in the overall population or predefined subpopulation exceeds 90% even with 300 patients. Furthermore, power of the PFS analysis in the predefined subpopulation analysis approaches 90% when the total sample size is 500 patients.

Sample size assessment for the influence and interaction conditions is presented in Figure 5. From a clinical perspective, assume an effect size (log-hazard ratio) of 0.1 is the minimum threshold for a clinically relevant benefit. Then the threshold in the influence condition, denoted by ε, may be set to 0.1. Furthermore, the interaction condition is based on a 1.2 threshold, that is, $\lambda = 1.2$. With this choice of the threshold for the interaction condition, the difference between the 2

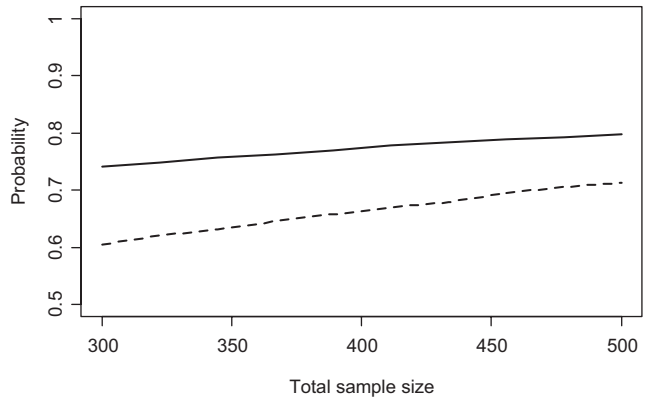

Figure 5. Probability of satisfying the influence condition (solid curve) and interaction condition (dashed curve). Hazard ratios in the target and complementary subpopulations are 0.65 and 0.8, respectively.

subpopulations is not considered clinically relevant unless the effect size in the predefined subpopulation exceeds the effect size in the complementary subpopulation by more than 20%. Assuming that the hazard ratio in the predefined subpopulation is 0.65, Figure 5 displays the probabilities of satisfying the influence and interaction conditions as a function of the total sample size. Figure 5 shows that the sponsor is guaranteed to have a sufficiently high probability of achieving the interaction condition (around 70%), if the total sample size is 500 patients. Furthermore, the probability of satisfying the influence condition is close to 75% over the entire range of sample sizes. Similar calculations can be performed for the case when the hazard ratio in the predefined subpopulation is 0.7.

To illustrate the decision-making process at the end of the phase III trial in patients with advanced non–small cell lung cancer, suppose that the total sample size in the trial is 400 patients and the observed hazard ratios in the predefined and complementary subpopulations are 0.69 and 0.82. The resulting hazard ratio in the overall population is 0.74. The treatment effects in the overall population and predefined subpopulation are statistically significant (with P values of .0017 and .0029, respectively). Suppose, furthermore, that the test of effect in the complementary subpopulation gives a P value of $P = .1128$. (It is important to note that this hypothesis test for the complementary subpopulation is not part of the preplanned analysis and is included here only for illustrative purposes.) Since a beneficial effect is present in the overall population as well as the predefined subpopulation, the labeling decisions can be based on the rules presented in Figure 1. First, the effect sizes in the predefined and complementary subpopulations (defined as the log hazard ratios) are given by 0.37 and 0.20, respectively. Note that the effect size in the complementary subpopulation exceeds $\varepsilon = 0.1$. The sponsor can rule out the possibility that the significant treatment effect in the overall population is

driven solely by the beneficial effect in the predefined subpopulation and thus the influence condition is satisfied. The next step involves assessment of the interaction condition based on the ratio of the effect sizes in the target and complementary subpopulations. This ratio is 1.9 and exceeds the prespecified threshold $\lambda = 1.2$, which implies that the interaction condition is also met. Since both conditions are met, it is recommended to consider a label with the broad indication based on a beneficial effect in the overall population and additional labeling to highlight the enhanced effect in the predefined subpopulation.

8. Discussion

In this article, we have focused on multipopulation tailoring trials that have as their primary objective assessment of effect in the overall population as well as in a predefined subpopulation. These trials offer significant benefits to patients, prescribers, trial sponsors, regulatory authorities, and general society because of their efficiency, the breadth of inference available, subsequent increased clarity of conclusions and relevant information, and potentially increased power to provide meaningful conclusions on treatment options for patients. Because of the additional complexity associated with multiple inferences, these clinical trials with tailoring objectives require additional consideration in the planning of design and analysis.

We have presented key considerations to support decision making based on a confirmatory multipopulation tailoring trial with continuous, binary or time-to-event endpoints. For analysis, we introduced 2 new concepts, the influence and interaction conditions, along with suggested methods for assessment. These assessments are recommended whenever there is (positive) joint inference of treatment effect in both the overall population and the predefined subpopulation. Similarly, assessment of the influence condition is recommended whenever there is (positive) inference of treatment effect in the overall population in a trial containing a predefined subpopulation. These assessments supplement the inference of the primary testing methodology, providing a decision framework for deeper understanding of the interplay between the population results.

Other authors[9,10] have presented a different analytical paradigm for this problem vis-à-vis consistency-ensured methods. With the consistency-ensured methods, positive inference for a predefined subpopulation is restricted to cases wherein there is *consistency* with the overall population result. While the testing paradigm is mathematically correct, this consistency requirement, in general, is in contrast to the objective of tailored medicine which is to determine appropriate populations—limited or broad—for which treatments are best tailored/suited. A new treatment may legitimately be appropriate for a subpopulation and not for the complementary subpopulation, that is, lacking consistency. Thus, the general restriction of consistency-ensured methods conflicts with the larger objective of tailored therapeutics. In contrast, while the influence condition may at first glance appear similar to a consistency requirement, this condition is relevant only when there is positive inference of overall population effect. In this case, the assessment is to ensure that the subpopulation result does not so overwhelm the data that a positive conclusion for the overall population is really due only to the subpopulation effect. As noted, this assessment, of course, is not relevant in the instance of positive inference for the subpopulation only.

We would like to point out that the requirement embodied in the consistency-ensured methods is relevant for a particular subset of tailoring trials. In particular, when the treatment involved presents sufficient risks to patients (eg, toxicity) and the probability of misclassification of patients to subpopulation versus its complement is relatively high, there is legitimate rationale to disallow labeling in a specific subpopulation without sufficient consistency with the overall population. The restrictions of consistency-ensured methods should not be assumed more broadly than such rare cases where the restriction is driven by clinical need.

It is important to note that the methods advocated herein are conceptually consistent with regulatory practice. That is, with any evaluation of the treatment effect in a population, the effects in subpopulations of interest are examined at least in an exploratory manner through additional analyses. In tailored therapeutics trials as considered in this article, there is opportunity to formally address key a priori questions that exist, thereby enabling more robust decision making. Within this framework, if there is sufficient evidence of lack of effect in a particular subpopulation while the overall population result is positive, consideration must be given to the best means to convey this information to patients and prescribers. This includes, but is not limited to, potentially restricting the indication. Similarly, when there is sufficient evidence of differential effects across a patient population, consideration must be given to conveying this information for the benefit of patients and prescribers, as opposed to simple broad population claims. The methods advocated in this article provide a framework to meet these needs.

The estimation framework presented herein, particularly the Bayesian estimation, has the advantage of helping facilitate cross-disciplinary (eg, statistics, medical) discussion and decision making; methods based solely on P values do not. Transparency of the potential need for assessment of the influence and interaction conditions may also facilitate dialogue between sponsor and regulatory agency regarding trial design and trial results, thereby resulting in clearer trial outcomes. This would, of course, be a win for patients, prescribers, trial sponsors, and regulators.

While all assessments presented in this article are at the trial level, it may be advantageous to consider these at the program level (across replicate trials). When trial designs permit, the increased sample size obtained at the larger program level would increase confidence in supplemental estimation-based assessments of the influence and interaction conditions and subsequent decision making.

In closing, the tailored therapeutics trials discussed here are more complex than traditional trials because of their objective of answering questions about multiple populations simultaneously. These trials can make far more efficient use of patients as clinical trial subjects and provide meaningful information on subpopulations of patients far more quickly than has been the historical norm. Statistical methods as discussed in this article provide appropriate machinery to support decision making based on these trials, and take full advantage of the immense possibilities from these novel designs.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Maitnourim A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine*. 2005;24:329-339.
2. Herceptin. U.S. prescribing information; 1998. http://www.gene.com/gene/products/information/pdf/herceptin-prescribing.pdf. Accessed July 12, 2012.
3. Xalkori. U.S. prescribing information; 2011. http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/202570s000lbl.pdf. Accessed July 12, 2012.
4. Zelboraf. U.S. prescribing information; 2011. http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/202429s000lbl.pdf. Accessed July 12, 2012.
5. Alimta. U.S. prescribing information; 2004. http://pi.lilly.com/us/alimta-pi.pdf. Accessed July 12, 2012.
6. Cappuzzo F, Ciuleanu T, Stelmakh L, et al. SATURN: a double-blind, randomized, phase III study of maintenance erlotinib versus placebo following nonprogression with first-line platinum-based chemotherapy in patients with advanced NSCLC. *J Clin Oncol*. 2009;27(15 S):8001.
7. Wang S, O'Neill R, Hung H. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm Stat*. 2007;6:227-244.
8. Wang S, Hung H, O'Neill R. Adaptive patient enrichment designs in therapeutic trials. *Biom J*. 2009;51:358-374.
9. Song Y, Chi GY. A method for testing a prespecified subgroup in clinical trials. *Stat Med*. 2007;26:3535-3549.
10. Alosh M, Huque M. A flexible strategy for testing subgroups and overall population. *Stat Med*. 2009;28:3-23.
11. Zhao YD, Dmitrienko A, Tamura R. On optimal designs of clinical trials with a sensitive subgroup. *Stat Biopharm Res*. 2010;2:72-83.
12. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res*. 2005;11:7872-7878.
13. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst*. 2007;99:1036-1043.
14. Dmitrienko A, Bretz F, Westfall PH, et al. Multiple testing methodology. In: Dmitrienko A, Tamhane AC, Bretz F, eds. *Multiple Testing Problems in Pharmaceutical Statistics*. New York, NY: Chapman and Hall/CRC Press; 2009:35-98.
15. Wiens B. A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharm Stat*. 2003;2:211-215.
16. Wiens B, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *J Biopharm Stat*. 2003;15:929-942.
17. Huque MF, Alosh M. A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *J Stat Plann Inference*. 2008;138:321-335.
18. Millen BA, Dmitrienko A. Chain procedures: a class of flexible closed testing procedures with clinical trial applications. *Stat Biopharm Res*. 2011;3:14-30.
19. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Stat Med*. 2009;28:586-604.
20. Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Stat Med*. 2009;28:739-761.
21. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*. 1985;41:361-372.
22. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interaction. *Stat Med*. 1993;12:1239-1248.