

Drug Information Journal

<http://dij.sagepub.com/>

Testing in a Prespecified Subgroup and the Intent-to-Treat Population

Mark D. Rothmann, Jenny J. Zhang, Laura Lu and Thomas R. Fleming

Drug Information Journal 2012 46: 175

DOI: 10.1177/0092861512436579

The online version of this article can be found at:

<http://dij.sagepub.com/content/46/2/175>

Published by:



<http://www.sagepublications.com>

On behalf of:



Drug Information Association

Additional services and information for *Drug Information Journal* can be found at:

Email Alerts: <http://dij.sagepub.com/cgi/alerts>

Subscriptions: <http://dij.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

>> [Version of Record](#) - Mar 1, 2012

[What is This?](#)

Testing in a Prespecified Subgroup and the Intent-to-Treat Population

Mark D. Rothmann, PhD¹, Jenny J. Zhang, PhD¹,
Laura Lu, PhD¹ and Thomas R. Fleming, PhD²

Drug Information Journal
46(2) 175-179
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0092861512436579
http://dij.sagepub.com

Abstract

In many settings, testing has been proposed to assess the effect of an experimental regimen within a biomarker-positive subgroup where it is biologically plausible that benefit is stronger in such patients, and in the overall population that also includes biomarker-negative subjects less likely to benefit from that regimen. A statistically favorable result in the biomarker-positive subgroup would lead to a claim for that subgroup, whereas a statistically favorable result for the overall population would lead to a claim that includes both biomarker subgroups. The latter setting is problematic when biomarker-negative patients truly do not benefit from the experimental regimen. When it is prespecified that biomarker-negative patients should not be included in the primary analysis of treatment effect in biomarker-positive patients because of the likelihood that treatment effects would differ between the 2 subgroups, it is logically inconsistent to include biomarker-positive patients in the primary analysis of treatment effect in biomarker-negative patients.

Keywords

biomarker, subgroup analysis, intent-to-treat

Introduction

The goal of personalized medicine is to optimize the treatment strategy on a patient-specific basis. Doing so requires an understanding about which drugs or regimens may or may not be beneficial for each patient. The motivation to pursue personalized medicine is enhanced by the recognition that molecularly targeted agents may benefit only a subset of patients. Although it is important for patients to have access to interventions that will provide benefit to them, patients belonging to a class that would not benefit from an agent should not receive it, particularly when such agents entail risks of toxicity. In a recent experience, despite yielding favorable statistically significant results on the primary end point for the entire intent-to-treat (ITT) population, a restriction for use was requested for 2 epidermal growth factor receptor antagonists when retrospective analyses revealed no benefit in patients whose tumors had KRAS (V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog) mutations.¹

Many authors have considered clinical trial designs that share the false positive error rate between a test conducted within only the biomarker-positive subjects and a test conducted in the entire ITT population.²⁻⁶ The underlying assumption of these approaches is the existence of a treatment by biomarker subgroup interaction effect that is strongly

quantitative. This assumption is in contrast to the standard design that tests solely in the entire ITT population and is generally used when there is prior belief that the treatment effect does not meaningfully vary across important subgroups.

These “ α -sharing” approaches are designed to retain the ability to assess the effect within the entire ITT population while providing adequate statistical power to evaluate effects within the biomarker-positive subgroup, even when there is minimal or no effect in the biomarker-negative group. The power of such approaches was assessed in settings in which

¹Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

²Department of Biostatistics, University of Washington, Seattle, WA, USA

This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

Submitted 12-Sep-2011; accepted 12-Dec-2011.

Corresponding Author:

Mark D. Rothmann, Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, 10903 New Hampshire Ave, Bldg 21, Room 3528, Silver Spring, MD 20993, USA (Email: mark.rothmann@fda.hhs.gov)

the intervention was assumed to have no effect in the biomarker-negative subgroup.^{2,3}

Wang et al considered approaches for allocation of false positive error that simultaneously test hypotheses for the overall ITT population and for the biomarker-positive subgroup, with an early futility analysis for the biomarker-negative subgroup.⁶ If futility is reached, then only biomarker-positive patients will be further accrued to increase the statistical power within that subgroup.

Many clinical trials have allocated (1-sided) false positive error rates of 0.02 and 0.005 for the comparison within the entire ITT population and the comparison within the biomarker-positive subgroup, respectively.² The sample sizes of such trials are usually based on having acceptable statistical power to detect a substantial effect in the biomarker-positive subgroup whether or not there is a beneficial effect in the biomarker-negative subgroup. A favorable, statistically significant result (eg, at 1-sided $\alpha = .02$) in the entire ITT population is intended to support a claim for all patients, whether biomarker-positive or biomarker-negative. This situation is problematic when there truly is little or no benefit in the biomarker-negative subgroup and statistical significance of the test within the entire ITT population is driven by a strong beneficial effect within the biomarker-positive subgroup. When favorable statistical significance is achieved in the testing of the biomarker-positive subgroup, testing within the ITT population simply checks whether statistical significance remains or is lost when adding the biomarker-negative patients.

There is a logical inconsistency with this α -sharing approach. If it is prespecified that biomarker-negative patients should not be included in the primary analysis of treatment effect in biomarker-positive patients, because of the likelihood that treatment effects would be different in these 2 subgroups, then it is logically inconsistent to include biomarker-positive patients in the primary analysis of treatment effect in biomarker-negative patients.

Let the parameters θ_+ and θ_- correspond to the treatment effects in the biomarker-positive and biomarker-negative subgroups, respectively. For settings in which there are clinically important interactions, these 2 parameters are quite different. Reaching favorable statistical significance in both tests means that θ_+ is positive and some average of θ_+ and θ_- is positive; however, it does *not* imply that θ_- must also necessarily be positive. Unfortunately, there is usually no formal testing of whether the experimental therapy is beneficial for the biomarker-negative patients.

In the remainder of this paper, in the second section, we provide an evaluation of the probability of a claim that includes the biomarker-negative subgroup even though the biomarker-negative patients in truth do not benefit. In the third section, we discuss some actual problematic designs involving positive and negative subgroups. Some concluding remarks are provided in the fourth section.

Error Rates

When an intervention does not provide a clinically meaningful beneficial effect in the biomarker-negative subgroup, those patients should not be included in any claims of benefit. For the α -sharing testing approaches discussed in the previous section, the error of including biomarker-negative patients in a claim of an effect could be large if the effect in the biomarker-positive subgroup is strong. In this section, we will examine the joint probabilities of the possible decisions based on the tests of the biomarker-positive subgroup and the entire ITT population.

When performing individual tests of the null hypothesis of no treatment effect within the entire ITT population and within the biomarker-positive subgroup, let $P(r-r)$ denote the probability of rejecting both null hypotheses, $P(r-nr)$ denote the probability of rejecting only the null hypothesis for the entire ITT population, $P(nr-r)$ denote the probability of rejecting only the null hypothesis for the biomarker-positive subgroup, and $P(nr-nr)$ denote the probability of not rejecting either null hypothesis. The probability of a claim that includes the biomarker-negative subjects is $P(r-r)+P(r-nr)$, and the probability of getting a claim only for the biomarker-positive subgroup is $P(nr-r)$. The overall power is $P(r-r)+P(nr-r)+P(r-nr)$. Ideally, when no benefit is provided to the biomarker-negative patients, $P(r-r)+P(r-nr)$ is small.

For our simulations, a clinical trial sample size of 200 patients was selected to provide a wide range in statistical power across the various scenarios for hazard ratios and proportions of subjects considered in the biomarker subgroup. Simulations (100,000 replications) were performed to determine the 4 probabilities of rejecting and failing to reject the null hypotheses in the entire ITT population and in the biomarker-positive subgroup. The simulations incorporate the conditional asymptotic distributions for the observed log-hazard ratio given the observed number of events. A 2-arm study was considered with (1) a total of 200 subjects randomized evenly between the experimental and control arms over a 21-month accrual period; (2) 10 months of additional follow-up; (3) exponential times to event for each treatment by biomarker subgroup combination; (4) median time to event of 6 months within each biomarker subgroup of the control arm; (5) proportions of patients in the biomarker-positive subgroup (w) of 0.75, 0.60, 0.40, or 0.25; (6) as in Jiang et al,³ experimental to control hazard ratios within the biomarker-positive subgroup of 0.57, 0.40, or 0.31, and within the biomarker-negative subgroup of 1; and (7) the analysis for the entire ITT population being stratified by biomarker subgroup. The 1-sided α level for testing within the entire ITT population (α_1) was set at .015 or .02, whereas the 1-sided α level for testing within only the biomarker-positive subgroup (α_2) was determined as $.025 - \alpha_1$.

Tables 1 and 2 provide the joint probabilities of the possible decisions for the cases studied. In all cases, as stated above, the

Table 1. Cell Probabilities when (1-sided) $\alpha_1 = .02$

w	HR _{sub}	P(r-r)	P(nr-r)	P(r-nr)	P(nr-nr)
0.75	0.57	0.6286	0.0626	0.0882	0.2206
	0.40	0.9723	0.0145	0.0050	0.0081
0.60	0.57	0.4276	0.1432	0.0860	0.3432
	0.40	0.8656	0.0921	0.0110	0.0313
0.40	0.57	0.1768	0.1955	0.0738	0.5538
	0.40	0.4954	0.3327	0.0276	0.1444
	0.31	0.6831	0.2781	0.0067	0.0322
0.25	0.57	0.0587	0.1531	0.0574	0.7308
	0.40	0.1863	0.3939	0.0408	0.3791
	0.31	0.2952	0.5099	0.0209	0.1741

Table 2. Cell Probabilities when (1-sided) $\alpha_1 = .015$

w	HR _{sub}	P(r-r)	P(nr-r)	P(r-nr)	P(nr-nr)
0.75	0.57	0.6418	0.1309	0.0340	0.1933
	0.40	0.9687	0.0243	0.0014	0.0056
0.60	0.57	0.4287	0.2357	0.0379	0.2978
	0.40	0.8468	0.1285	0.0037	0.0210
0.40	0.57	0.1759	0.2946	0.0390	0.4905
	0.40	0.4654	0.4177	0.0115	0.1054
	0.31	0.6470	0.3307	0.0022	0.0201
0.25	0.57	0.0603	0.2318	0.0349	0.6730
	0.40	0.1734	0.5007	0.0210	0.3050
	0.31	0.2673	0.5974	0.0091	0.1261

true hazard ratio is one within the biomarker-negative subgroup. For (1-sided) $\alpha_1 = .02$ and $.015$, the probabilities of a claim that includes biomarker-negative subjects ($P[r-r]+P[r-nr]$) ranged, respectively, from 0.116 to 0.977 and from 0.095 to 0.970. The greater the proportion of biomarker-positive subjects and/or the greater the true effect in that subgroup, the larger the probability of a claim that includes the biomarker-negative subjects. The probability of a claim that includes the biomarker-negative subjects was slightly less for (1-sided) $\alpha_1 = .015$ than for (1-sided) $\alpha_1 = .02$. For (1-sided) $\alpha_1 = .02$, the probability of a claim on the basis of a favorable result only for the entire ITT population was as high as 0.088. As there is a 0.02 probability of achieving a favorable result when all subjects are biomarker negative, the cases of “statistical significance” for the entire ITT population only are driven by the results in the biomarker-positive subgroup.

Problematic Formulations

Based on the authors’ experiences, problematic analyses are presented that involve testing in a selected subgroup and testing within the entire ITT population. Various proposed tests only check whether statistical significance of a previous test remains or is lost when a subgroup is added to the analysis population.

These tests do not evaluate whether there is efficacy within the added subgroup. The final case provided in this section involves the level of supportiveness on the results from subgroup analyses.

Case 1

In this scenario, if a favorable result (eg, 1-sided P value $< .025$) was obtained in a more restricted population defined by a prespecified cut point for a quantitative biomarker, then attempts would be made to extend to a favorable result in a broader population defined using a more inclusive cut point. Obtaining a 1-sided P value $< .025$ when the additional subjects are included does not mean that the investigational drug benefits subjects who have biomarker values between the prespecified cut point and this new cut point.

In a related case, regimens were investigated that involved the addition of an experimental drug, A, to background therapies, B or C. In this setting, a previous large study had compared C + A to C alone in a more advanced setting of the disease. No differences were observed when the primary end point was analyzed and C alone was favored for an important secondary end point. The new study used a hierarchical testing approach, in which an initial comparison of B + A with B alone was made on the primary end point at a (1-sided) $.025$ level; if statistically significant beneficial effects were achieved, then the integrated comparison of B + A with B and C + A with C was to be tested at a (1-sided) $.025$ level. It was intended to make a positive claim for adding A to B if a favorable result was obtained from the first test. A claim for adding A to C would also be made if the second test was favorable. This proposed hierarchical analysis does maintain a (1-sided) type I error rate of 0.025. However, the second test simply checks whether statistical significance remains or is lost when patients given C or C + A are added to the analysis; it does not test whether the effect of adding A to C is positive. This is a particular concern, given the unfavorable results in the comparison of C + A with C alone from the previous large trial in the more advanced setting of the disease.

Case 2

Consider an add-on or placebo-controlled study based on a time-to-event primary end point, in which the proposed analysis uses the α allocation of (1-sided) $.005$ for the biomarker-positive subgroup and (1-sided) $.02$ for the entire ITT population. For the studied indication, about 60% of potential subjects are biomarker-positive. Biomarker-positive subjects have shorter times to the event than biomarker-negative subjects. The timing of the study analysis is such that if 60% of the subjects entering the trial are biomarker-positive, about 64% of the events for the analysis will come from the biomarker-positive subjects. Hence, based on a stratified analysis, the parameter being evaluated for

the entire ITT population would be $0.64\theta_+ + 0.36\theta_-$, not $0.6\theta_+ + 0.4\theta_-$. There may also be a concern that such a testing approach or objective may encourage enrollment of biomarker-positive patients but discourage enrollment of biomarker-negative patients, leading to an even greater difference in the percentage of deaths between the 2 biomarker subgroups at the time of analysis, and to the assessment of a parameter quite different from that of the overall patient population that has the disease. Additionally, because of the α allocation, when 60% of events are within the biomarker-positive subgroup, favorable statistical significance can be achieved in the entire ITT population without statistical significance achieved in the biomarker-positive subgroup, even when no difference is observed within the biomarker-negative subgroup. Although the overall false positive error rate is preserved by the testing scheme, if the data confirm the biological expectation that treatment benefit is much more evident in the biomarker-positive subgroup, should there not be persuasive evidence of benefit in the complement subgroup before a global claim is made?

Case 3

External information may arise suggesting that biomarker-negative patients may not benefit or may be harmed by the experimental therapy or a similar product. This situation may prompt changes to an ongoing trial that involves both biomarker-positive and biomarker-negative subjects. If accrual is ongoing, all subsequent accrued patients may need to be biomarker positive.

Amending the primary analysis to hierarchically test first within the known biomarker-positive subgroup at (1-sided) $\alpha = .025$, and if favorable statistical significance is achieved, then test within the entire ITT population at (1-sided) $\alpha = .025$ has similar concerns as with earlier cases. The second test simply checks whether statistical significance remained or was lost when the patients in the biomarker-positive complement group (ie, biomarker-negative) were added to the analysis. The treatment effect within the biomarker-negative patients was not formally analyzed. This situation is particularly problematic when the study was modified because of concerns that biomarker-negative patients may not benefit or may be harmed by the experimental drug. In such cases, it would make sense that any conclusion regarding the efficacy and safety of the experimental drug for biomarker-negative patients be based only on biomarker-negative patients. Prior to the emergence of use of this "creative" α -sharing approach, a commonly considered strategy would have been to change the primary analysis population to just the biomarker-positive patients.

Case 4

Statisticians often test for an interaction when analyzing the same end point over 2 disjoint subgroups. When there is a demonstrated interaction, separate inferences are made in the subgroups. When an interaction is not demonstrated, a pooled analysis is performed. However, the failure to establish an interaction does not mean that the observed interaction is small or that substantial interaction effects have been ruled out.

This conditional approach to determining whether to pool across subgroups based on the results of a test for interactions is problematic for other reasons, as well. The results from one group may be supportive when evaluating whether an observed positive treatment effect in another group is real, even when an interaction has been shown. Consider the following 2 scenarios for a time-to-event end point where individual comparisons are done for each sex. In the first scenario, the experimental-to-placebo hazard ratio and 95% confidence interval are 0.84 and (0.70, 1.00), and 0.87 and (0.73, 1.04), respectively for the female and male subgroups. In the second scenario, the experimental-to-placebo hazard ratio and 95% confidence interval are 0.44 and (0.37, 0.52), and 0.87 and (0.73, 1.04), respectively for the female and male subgroups. Which scenario provides stronger evidence that the experimental therapy works for males? In the first scenario, usual practice would conclude that each subgroup analysis supports the other and that the results could be pooled across sexes, as there is no apparent interaction effect. In the second scenario, there is a clear interaction effect, and usual practice would be to separately analyze the females and the males. Therefore, under usual practice, the first scenario provides borderline positive results for each sex with a pooled result achieving statistical significance, whereas the second scenario provides a clear, real positive effect in females but an effect in males that fails to achieve the traditional standard for statistical significance. In contrast to the reasoning in this "usual approach," clinical common sense would recognize that the marginally significant result in males is more strongly reinforced by the large effect in females in the second scenario than by the much smaller effect in females in the first scenario.

Concluding Remarks

When it is thought that a clinically important positive effect is more likely in the biomarker-positive subgroup, it has been common to design Phase 2 trials to obtain additional evidence about effect modification, and then to design randomized clinical trials to confirm the effect of a product only in that patient population in which benefit is expected. Alternatively, a randomized clinical trial could be conducted with both biomarker-

positive and biomarker-negative subjects. Such a trial allows a confirmatory assessment of the favorable expectation in biomarker-positive subjects, while learning about potential effects in biomarker-negative subjects. In this setting, the analysis plan should be consistent with this reasoning. In particular, for conclusions of favorable efficacy to be extrapolated to include the biomarker-negative subjects, not only should there be care to address the multiplicity caused by subgroups when protecting the false positive error rate, but there also should be a requirement that there be an adequate amount of data within the biomarker-negative subgroup to reliably estimate the level of effect in that subgroup. Furthermore, the size of the estimated effect in the biomarker-negative subgroup should be clinically relevant and should be at least as large as what would be needed to achieve “statistical significance” in an analysis conducted in the entire ITT population.

When there is no expectation that treatment effects will be sharply different across subgroups, the primary analysis should be limited to the entire ITT population. Subgroups can be investigated in an exploratory manner. This approach is quite robust, even when treatment effects are larger in a biomarker-positive subgroup. More specifically, if an intervention provides clinically important benefits in a biomarker-positive subgroup and if that subgroup accounts for a substantial fraction of the population, then the analysis in the entire ITT population will be sensitive to this beneficial effect, except in rather extreme circumstances, such as when there is meaningful harm in the biomarker-negative subgroup.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This article reflects the views of the authors and should not be construed to represent FDA’s views or policies.

Funding

For Thomas R. Fleming, the source of financial support for research described in this article is NIH/NIAID grant entitled “Statistical Issues in AIDS Research” (R37 AI 29168).

References

1. FDA Oncologic Drugs Advisory Committee, Dec 16, 2008 transcript. <http://www.fda.gov/ohrms/dockets/ac/cder08.html#OncologicDrugs>. Accessed February 8, 2012.
2. 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.
3. 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.
4. Simon R. New challenges for 21st century clinical trials. *Clin Trials* 2007;4:167-169.
5. Simon R, Wang S-J. Use of genomic signatures in therapeutics development in oncology and other diseases. *Pharmacogen J*. 2006;6:166-173.
6. Wang S-J, O’Neill RT, and Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset, *Pharmaceut Statist*. 2007;6:227-244.