ISSUE BRIEF

Conference on Clinical Cancer Research

October 2010

PANEL 1
Adaptive Clinical Trials Designs for Simultaneous Testing of Matched Diagnostics and Therapeutics

Howard I. Scher, Memorial Sloan-Kettering Cancer Center
Richard Simon, National Cancer Institute
Rajeshwari Sridhara, US Food and Drug Administration
Eric Rubin, Merck
Shelley Fuld Nasso, Susan G. Komen for the Cure

The need for adaptive clinical trials designs

Nearly all cancer drugs being developed today are designed to inhibit molecular targets that have been identified as deregulated in human tumors. Genomics has established, however, that the tumors of a given primary site are generally quite heterogeneous with regard to their mutated genes and deregulated pathways. Consequently, it is important that most new cancer drugs be developed in conjunction with diagnostics that identify tumors that are likely to be most sensitive to the anti-proliferative effects of a particular drug or drug combination.

The reality of co-developing a matched diagnostic and therapeutic has profound implications for the clinical trial designs used in the development of current drugs. Trials of cytotoxic drugs typically enroll unselected patients at a particular point in the continuum of a disease in the hope that the response of tumors that are sensitive to the treatment will be sufficient to show benefit for the population as a whole. While this approach may lead to broad labeling indications, it also results in the treatment of many patients who do not benefit and the possibility of discarding a drug that may dramatically benefit a subset of patients. Consequently, this strategy is not viable for most molecularly targeted drugs for which the activity is likely to be more restricted and often determined more by genetic makeup or molecular profiles than site of origin, or point in the progression of a specific cancer. Indeed, the use of anatomically based (primary site of disease), “all comers” approaches to targeted drug development has typically led to failure in phase III studies, or demonstration of “success” based on statistically significant, but clinically questionable benefit.

Although developing the right drug for a specific patient has great value to the individual, and for controlling societal medical costs, the complexities associated with identifying a predictive biomarker to use as the basis of a diagnostic, developing the analytically valid diagnostic test for clinical testing, and designing and executing the series of prospective clinical trials needed to generate the evidence to qualify the biomarker as a predictor of sensitivity in the target population defined by the diagnostic are grossly underestimated. This is particularly the case when the effectiveness of the drug in this population is uncertain. Developing the right drug for the right subset of patients also requires new clinical trial designs and new paradigms of data analysis.

Efforts to co-develop a matched diagnostic and therapeutic face other challenges as well. It is often difficult, for example, to identify a predictive biomarker based on preclinical studies or phase I trials of a given drug. This necessitates or highlights the need to evaluate candidate markers during phase II studies so that properly focused phase III trials can be conducted. Adaptive phase II designs, such as that recently used in the BATTLE clinical trial in NSCLC[1] and the I-SPY2 trial in breast cancer[2] are useful in this regard. The FDA has also issued a draft Guidance to Industry on adaptive design clinical trials for drugs and biologics.[3] Recently, many adaptive clinical trial designs have been published, including oncology clinical trials using different adaptive design approaches.[4,5]

Because of the complexity of cancer biology, it is in some cases not possible to firmly establish the biomarker(s) most likely to predict benefit to a particular drug or class of drug by the time pivotal phase III trials are set to begin. It is possible, however, to design the pivotal trial(s) so that the most suitable target population of patients is adaptively identified during the trial and the effectiveness of the drug evaluated in that population in a rigorously defined and statistically valid manner. For example, when the biomarker assay has been validated for measurement of a specific tumor characteristic with well-established assay performance characteristics, standardized and performance characteristics are known, adaptive signature design[6] and cross-validated adaptive signature design[7] are carefully crafted frequentist adaptive phase III design approaches that preserve the overall chance of any false positive conclusion while identifying an optimal target population. Neither design results in a change in randomization weights or eligibility criteria, making them better suited for phase III use than the Bayesian methods used in the phase II BATTLE trials. However these are complex designs that have not been tested in practice. Challenges to the use of these designs are that the treatment comparisons can only be conducted after completion of the study, that the developed predictive signature may be based on a combination of factors with unclear biologic meaning, and that it may be difficult (challenging) to interpret results if there are imbalances in other baseline prognostic factors between treatment arms in the marker positive subgroup.

Such designs, however, although in some ways conservative, are dramatically different than the kinds of designs used for the vast majority of clinical trials being conducted today. To illustrate how an adaptive trial design could be used in clinical practice, we propose a design to test four potential biomarkers (B1, B2, B3 and B4) in conjunction with a candidate targeted drug therapy (compound X), for which the biomarker assay has been validated and there is a strong scientific rationale that the biomarker is potentially predictive. For practical considerations, predictive biomarkers that can be analyzed in formalin fixed paraffin embedded material are preferred.

A Potential Phase III Adaptive Trial Design

Eligible patients consist of individuals with progressive castration resistant prostate cancer for whom a targeted therapeutic approach is being developed, and for whom tumor material is

[1] Printz C. BATTLE to Personalize Lung Cancer Treatment. Novel Clinical Trial Design and Tissue Gathering Procedures Drive Biomarker Discovery. Cancer. 116, 3307-3308 (2010).
[2] Barker A, Sigman C, Kelloff G, Hylton N, Berry D, Esserman L. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clinical Pharmacology and Therapeutics. 86(1), 97-100 (2009).
[3] http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm064981.htm accessed 10/4/10
[4] Sargent DJ, et.al. Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of Clinical Oncology 23, 2020-2027 (2005)
[5] Freidlin et.al. Randomized clinical trials with biomarkers: Design issues. Journal of the National Cancer Institute 102, 152-160 (2010)
[6] Freidlin B and Simon R. Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients. Clinical Cancer Research. 11, 7872-7878 (2005).
[7] Freidlin B, Jiang W, and Simon R. The Cross-Validated Adaptive Signature Design. Clinical Cancer Research. 16, 691-698 (2010).

available. For some biomarkers, primary tumor samples representing the diagnostic prostate biopsy or radical prostatectomy specimen may be sufficient. In other cases, a new metastatic tumor sample may be required. Biopsy specimens (flash frozen or paraffin embedded) are stored for later assay. After eligibility is confirmed and the availability of sufficient tumor for analysis is confirmed, a patient is randomized to treatment with compound X or placebo.

The randomization will be pre-stratified by institution. Pre-stratification by any of the biomarker values is not necessary for statistical validity of the analyses to be described and analytically validated tests for some of the biomarkers may not be available until time of final analysis of the trial. The biomarkers will be assayed prior to the final analysis with tests that are analytically validated for use with the specimens available. The requirement of sufficient tumor material for analysis at entry ensures almost complete ascertainment of biomarker values.

The primary endpoint for the study is survival. The final analysis will take place after 700 total deaths are observed. This will provide approximately 90% statistical power for detecting a 25% reduction in hazard of death for compound X relative to control at a 1% two-sided statistical significance level. The remaining 4% of type I error will be used for evaluating the statistical significance of treatment effect on survival in the adaptively defined biomarker subset as described below. A total of 935 total patients will be accrued and the final analysis performed when there are 700 total deaths. The sample size results from the requirement for the high statistical power needed to detect modest improvements in median survival either overall or for an adaptively defined subset that might include only 33% of the patients. The number of patients required is strongly dependent on both factors and can be substantially reduced if larger treatment effects in more highly prevalent subsets are targeted. The expected number of patients can also potentially be reduced by interim analysis leading to early termination if the overall treatment effect is greater than anticipated. Such modifications can be made to the basic design described below.

The final analysis will be conducted in the following manner. The two treatment arms will be compared using survival times for all randomized patients using a log-rank test. If the two sided significance level is less than 0.01 and favors compound X, then compound X will be considered effective for the randomized population as a whole. If not, then the following analysis will be performed using the approach developed by Freidlin and Simon.[3]

A classifier C(B1,B2,B3,B4) will be developed that identifies whether a patient with biomarker values B1, B2, B3, and B4 is likely to benefit from drug X compared to control C. This classifier will be developed using a randomly selected training set of patients consisting of 33% of the cases. A training set consisting of approximately 233 events should be adequate for developing a classifier whose accuracy is close to that of the infinite sample classifier.[8] The algorithm for developing the classifier is described in the appendix below. The value of the classifier function C(B1,B2,B3,B4) equals 1 if the patient with those biomarker values is likely to benefit from X, and equals 0 otherwise. The set IND of combinations of biomarker values (B1,B2,B3,B4) for which the classifier equals 1 is the indication for treatment X should the subset analysis be statistically significant. As part of the final analysis, this indication will be described graphically, analytically, by decision tree, and as a classification function.

The estimated improvement in survival for X versus C in the indicated population IND will be estimated by classifying each patient in the trial who was not included in the training set used to

[8] Dobbin K, and Simon R. Sample Size Planning for Developing Classifiers Using High Dimensional DNA expression data. Biostatistics. 8, 101-1117 (2007).

develop the classifier. Let S denote the set of patients in this “test set” classified as likely to benefit from X using C(B1,B2,B3,B4).

Kaplan-Meier survival curves will be computed for the patients in S who received X and for the patients in S who received C. The difference between these two survival curves will be summarized using a log-rank statistic LR and a log hazard ratio (LHR) and a 96% confidence interval for LHR. If the log-rank statistic LR is significant at the 4% level of the chi-squared distribution with one degree of freedom and if the hazard ratio of X versus C is less than 1, then the treatment X will be considered effective in improving survival of patients with an indication specified by the set IND defined based on the classifier C(B1,B2,B3,B4) as described above.

The statistical power of the biomarker specified subset analysis depends on the proportion of patients who are included in the adaptively defined subset S. In order to have 80% power for detecting a 37% reduction in the hazard of death for X versus C, approximately 157 deaths are required in the classifier positive subset of the test set of patients (i.e. patients not used for developing the classifier). If one-third of patients are classifier positive, then 471 total deaths are required in the test set. The test set will contain about two thirds of the patients and events. The total number of deaths at the time of final analysis will be 700 and hence this power target should be achieved. As noted above, if the proportion of patients who are classifier positive is targeted to be larger than 33% or the size of the treatment benefit in the adaptively defined subset is targeted to be greater than 37%, then the number of patients required can be substantially reduced.

A single interim futility analysis will be performed during the clinical trial. The analysis will be based on freedom from progression at 6 months as an intermediate endpoint. This endpoint is not claimed to be a surrogate for survival, but is a conditional surrogate in that a drug that fails to prolong time till disease progression is unlikely to prolong survival. Using a 6-month time landmark ensures that the intermediate endpoint can be assessed without bias. Freedom from progression at 6 months is more suitable than survival for an interim futility analysis because it is more rapidly observed and hence can better protect patients from further exposure to a drug which may be unlikely to help them.

The interim analysis will be conducted when approximately 340 total patients have been followed for 6 months after randomization. This will provide 90% statistical power for detecting a 12 percentage point increase in the proportion of patients free of progression at 6 months from a baseline of 40% for the control regimen at a one-sided significance level of 0.20. If this is not achieved, then accrual to the clinical trial will be terminated and no claims of efficacy of compound X will be claimed; otherwise, accrual will continue as planned to the final analysis. Because the futility analysis will not result in early termination with a claim of effectiveness, it does not consume any of the overall 5% type I error of the study. We do not get into the issues of designing interim analyses for superiority of X over C leading to early termination, but such analyses would generally be included in the design. Such analyses would use the regulatory endpoint, in this case survival, not 6-month progression-free survival. If analytically validated tests for the biomarkers were available at the start of the trial, then the interim analyses could potentially utilize those markers. The additional complexities that this would introduce are not addressed here, however.

Conclusion

If the goal of developing the right drug for the right patient is to become more than a cliché, sponsors, investigators, and regulators must recognize that some of the conventional wisdom used to guide clinical trial design and analysis in the era of broadly targeted cytotoxic agents is no longer appropriate. Indeed, the continued use of traditional clinical trials designs is likely to hamper the development of new drugs that are highly effective for molecularly well defined subsets of patients.

Using conventional, primary site-based approaches to develop targeted cancer therapeutics is in many cases not consistent with knowledge of tumor biology, exposes patients to toxic drugs to which they are not expected to benefit, and may result in long delays for the approval and availability of drugs which offer substantial benefit to molecularly characterized subsets of patients. However, new clinical trial design and analysis methods must be no less rigorous than conventional designs in their use of randomized controls, clinically meaningful endpoints and protection of type I error. Clearly in this new era, previously 'standard' issues such as the role of subset analysis, the role of stratification, the need to have broad eligibility criteria, and the use of adaptive methods must be critically re-examined.

Appendix 1

The classifier will be developed using the following algorithm.

A proportional hazards model will be fit to the data for the combined treatment X and control group. Denote this model by

$$\log(\lambda(t,B1,B2,B3,B4,v)/\lambda_0(t)) = \delta v + \beta_1 B1 + \beta_2 B2 + \beta_3 B3 + \beta_4 B4 + v(\gamma_1 B1 + \gamma_2 B2 + \gamma_3 B3 + \gamma_4 B4)$$

where v is a binary treatment indicator (v=1 for X, v=0 for C), δ is the regression coefficient that represents the main effect of treatment on survival, the β's reflect the prognostic effects of the biomarkers, and the γ's are the interaction effects that represent the predictive effects of the biomarkers. The left hand side of the equation represents the log hazard relative to the baseline hazard. The markers will only be binary if a cut-point is pre-defined based on preliminary data. Otherwise, no cut-point will be imposed on the modeled values.

For a patient with biomarker values (B1,B2,B3,B4), the log hazard ratio if the patient receives treatment X minus the log hazard ratio if the patient receives the control C is

$$\Delta(B1,B2,B3,B4) = \delta + \gamma_1 B1 + \gamma_2 B2 + \gamma_3 B3 + \gamma_4 B4$$

By fitting the model to the data, we obtain estimates of the regression coefficients and a covariance matrix for these estimates. Hence for any vector of biomarker values, we can compute $\hat{\Delta}(B1,B2,B3,B4)$ in which the regression coefficients are replaced by their estimates, and we can compute the variance $V[\hat{\Delta}(B1,B2,B3,B4)]$. A binary classifier will be defined by

$$C(B1,B2,B3,B4) = 1 \text{ if } \hat{\Delta}(B1,B2,B3,B4)/\sqrt{V[\hat{\Delta}(B1,B2,B3,B4)} \le c$$

The patient is classified as likely to benefit from X if the standardized log hazard ratio of X relative to C is less than or equal to a constant c. The constant will be determined by 10-fold cross validation within the training set to maximize the log-rank statistic for treatment effect within the training set patients classified as likely to benefit from X.