Published in final edited form as:

J Biopharm Stat. 2009 ; 19(3): 530–542. doi:10.1080/10543400902802458.

Clinical Trial Designs for Predictive Biomarker Validation: One Size Does Not Fit All

Sumithra J. Mandrekar and Daniel J. Sargent

Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA

Abstract

Traditionally, anatomic staging systems have been used to provide predictions of individual patient outcome and, to a lesser extent, guide the choice of treatment in cancer patients. With targeted therapies, biomarkers have the potential for providing added value through an integrated approach to prediction using the genetic makeup of the tumor and the genotype of the patient for treatment selection and patient management. Specifically, biomarkers can aid in patient stratification (risk assessment), treatment response identification (surrogate markers), or differential diagnosis (identifying individuals who are likely to respond to specific drugs). In this study, we explore two major topics in relation to the design of clinical trials for predictive marker validation. First, we discuss the appropriateness of an enrichment (i.e., targeted) vs. an unselected design through case studies focusing on the clinical question(s) at hand, the strength of the preliminary evidence, and assay reproducibility. Second, we evaluate the efficiency (total number of events and sample size) of two unselected predictive marker designs for validation of a marker under a wide range of clinically relevant scenarios, exploring the impact of the prevalence of the marker and the hazard ratios for the treatment comparisons. The review and evaluation of these designs represents an essential step toward the goal of personalized medicine because we explicitly seek to explore and evaluate the methodology for the clinical validation of biomarker guided therapy.

Keywords

Biomarker; Enrichments designs; Predictive marker designs; Unselected designs

1. Introduction

With the advent of targeted therapies, biomarkers provide an increasingly promising means of individualizing therapy. Biomarkers are classified into prognostic markers (associated with disease outcome) and/or predictive markers (associated with drug response). A prognostic marker is a single trait or signature of traits that separates a population with respect to the outcome of interest in the absence of treatment, or despite nontargeted “standard” treatment. Prognostic marker validation is relatively straightforward; it is associated with the disease or the patient and at least theoretically can be established using data from a series of patients

Address correspondence to Sumithra J. Mandrekar, PhD, Associate Professor of Biostatistics, Division of Biomedical Statistics and Informatics, Mayo Clinic, Harwick 8, 200 1st Street SW, Rochester, MN 55905, USA; mandrekar.sumithra@mayo.edu.

Publisher's Disclaimer: Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

treated with placebo or with standard treatment. A predictive marker, on the other hand, is a single trait or signature of traits that separates a population with respect to the outcome of interest in response to a particular (targeted) treatment. A validated predictive marker can prospectively identify individuals who are likely to have a favorable clinical outcome, such as improved survival and/or decreased toxicity. Predictive signatures are becoming increasingly common in cancer treatment to monitor disease severity and to predict the outcome to different treatments (Bonomi et al., 2007; Conley and Taube, 2004; Paik, 2003; Sequist et al., 2007; Slamon, 2000; Taube et al., 2005). The goal of a predictive biomarker is to select the optimal therapy from among multiple treatments, and thus the same standards of evidence required to adapt a new treatment are appropriate. If we accept this proposition, then by definition, a predictive marker validation must be prospective in nature, and the obvious strategy would be to conduct a prospectively designed randomized controlled trial (RCT) to test a marker by treatment interaction (Sargent and Allegra, 2002).

In the setting of a single or a multimarker predictive signature that can be distilled to a binary measure, at least two distinct RCT trial designs have been proposed for prospective validation of a proposed predictive marker. The first design, the marker by treatment interaction design, uses the marker status as a stratification factor when randomizing subjects to treatment. The second, the marker-based strategy design, randomizes subjects to have their treatment either based on or independent of the marker status. We will refer to these two designs collectively as unselected predictive marker designs, unselected because all patients of a specific disease type and stage are eligible for the clinical trial, regardless of their actual marker status. A focused assessment and discussion of these two approaches for specific clinical situations has been published previously (Mandrekar et al., 2005; Sargent and Allegra, 2002; Sargent et al., 2005). There is, however, continued discussion regarding the relative merits of these designs, some of which we investigate in this study.

In addition to the discussion of the relative merits of the two proposed unselected designs, there is ongoing debate regarding the overall utility of the unselected predictive marker designs vs. a more restrictive clinical trial design, the so-called enrichment design, to establish the predictive utility of a biomarker. An enrichment design screens patients for the presence or absence of a marker or a panel of markers and then only includes patients in the clinical trial who either have or do not have a certain marker characteristic or profile. The enrichment design results in a stratification of the study population with a goal of understanding the safety, tolerability, and clinical benefit of a treatment in the subgroup of the patient population defined by a specific marker status. This design is based on the paradigm that not all patients will benefit from the study treatment under consideration, but rather that the benefit will be restricted to a subgroup of patients who either express or do not express a specific molecular feature.

Simon and Maitournam (2004) and Maitournam and Simon (2005) evaluated the relative efficiency of a targeted (i.e., enrichment) vs. an untargeted (i.e., unselected) design for a randomized clinical trial comparing a new treatment to a control when there is strong preliminary data supporting the hypothesis that a therapy will only benefit a specific marker-defined subpopulation. In particular, they evaluated the enrichment vs. the unselected design based on the number of patients screened and the number of patients randomized. Under the assumption that the mechanism of action of the drug was well studied and that the assay for identifying the “target” patients was reliable, their simulations showed that the targeted design would require fewer randomized patients compared to the untargeted design. Under these same conditions, the targeted design also required fewer patients to be screened in comparison to the number of patients required for randomization with the untargeted design. However, the reduction in the sample size was dependent on the accuracy of the assay, as well as the prevalence of the markers in patients, in addition to other limitations (Maitournam and Simon, 2005; Simon and Maitournam, 2004). Clearly, such a marker-based selection of patients is

justified when there is compelling preliminary evidence to suggest that patients with or without that marker profile do not benefit from the treatments in question. However, when such preliminary evidence is less solid or unavailable, an unselected design might be more appropriate.

In this manuscript, we explore these two major topics in greater detail. In Section 2, we present three specific examples to illustrate when the unselected vs. the enrichment designs might be most appropriate in the context of a clinical situation. In Section 3, we discuss and compare the relative merits of the two unselected predictive marker designs for a range of clinically relevant scenarios. Section 4 includes concluding remarks and thoughts for further research.

2. Unselected Predictive Marker Designs vs. Enrichment Designs

The choice between performing a trial in a selected population (enrichment design) vs. conducting an unselected trial has major implications in terms of study sample size, complexity, and likelihood of success. Although there is no global solution for when a particular design should be employed, we discuss the appropriateness of unselected vs. enrichment designs in the context of three timely examples in the treatment of cancer patients.

2.1. Example 1—HER2 as a Marker for Herceptin in Breast Cancer

Trastuzumab (Herceptin), a human monoclonal antibody, is currently approved for treatment of HER2 positive breast cancer patients in the adjuvant setting (Romond et al., 2005). Careful assay validation and substantial preclinical evidence suggested that only HER2 positive breast cancer patients would benefit from treatment with Herceptin (Pauletti et al., 2000; Pegram et al., 2004; Pietras et al., 1998; Shepard et al., 1991), where HER2 positivity was defined as either over expression of HER2 protein by immunohistochemistry (3+) or HER2 gene amplification by fluorescence in situ hybridization (FISH-HER2:CEP17 ratio of >2.0). This definition for HER2 positivity was based on data from the advanced disease setting, but was nevertheless applied to the adjuvant setting, and led to an enrichment design strategy of enrolling only HER2 positive patients in two national intergroup adjuvant breast cancer trials: 1) the National Surgical Adjuvant Breast and Bowel Project trial (NSABP B-31) comparing doxorubicin and cyclophosphamide followed by paclitaxel every 3 weeks with the same regimen plus 52 weeks of trastuzumab beginning with the first dose of paclitaxel and 2) the North Central Cancer Treatment Group trial (NCCTG N9831) comparing three regimens, doxorubicin and cyclophosphamide followed by weekly paclitaxel, the same regimen followed by 52 weeks of trastuzumab after paclitaxel, and the same regimen plus 52 weeks of trastuzumab initiated concomitantly with paclitaxel. A combined analysis of these two trials demonstrated that Trastuzumab combined with paclitaxel after doxorubicin and cyclophosphamide significantly improves disease-free survival outcomes among women with surgically removed HER2-positive breast cancer (Romond et al., 2005). In this case, the enrichment strategy seems to have been successful: Only approximately 20% of women are HER2 positive, and if there truly were no benefit of Herceptin in 80% of women that are deemed HER2 negative, a much larger sample size would have been required to establish statistically significant results in an unselected study.

Subsequent analyses of the data from NCCTG N9831 and NSABP B-31, however, have raised the possibility of a beneficial effect of Herceptin in a more broadly defined patient population. Specifically, post-hoc central testing for HER2 expression from the available tumor tissue blocks from the NSABP B-31 trial has demonstrated that patients with tumors that were negative for FISH and had less than IHC 3+ staining by central testing also derived benefit from Herceptin, thus suggesting that the definition for HER2 positivity based on FISH or IHC for the adjuvant disease setting may need to be refined (Paik et al., 2007). Similarly, in the cohort of patients entered on the NCCTG N9831 trial based on local HER2 positivity but found

to be HER2 negative by central testing, the observed benefit of herceptin was similar to the HER2 positive patients by central testing (Perez et al., 2007). Additionally, early in the NCCTG N9831 study, there was a high degree of discordance (approximately 25%) in the HER2 results between central and local testing for IHC and FISH (Perez et al., 2006). Because patients deemed HER2 negative based on the local evaluation were not enrolled onto the NSABP B31 and NCCTG N9831 trials, the validity of HER2 as a predictive marker could not be fully established, and the question remains open whether Herceptin therapy may benefit a potentially much larger group than the approximately 20% of patients defined as HER2 positive in these two trials.

Thus, in this situation, although the enrichment strategy did clearly and quickly define an effective treatment for a subset of patients, several other questions regarding the predictive utility of HER2 were left unanswered due to the issues of assay reproducibility and inclusion of only biomarker defined subgroups. An unselected design, allowing for both HER2 positive and negative patients, may have helped provide these answers in a definitive and ultimately more timely manner.

2.2. Example 2—EGFR as a Marker for Erlotinib in Lung Cancer

Randomized trials in unselected patients with advanced non-small-cell lung cancer (NSCLC) comparing erlotinib or gefitinib (two epidermal growth factor receptor (EGFR) tyrosine kinase (TK) inhibitors) to placebo have demonstrated a small survival advantage for erlotinib-treated patients and a trend toward improved survival for gefitinib-treated patients (Cappuzzo et al., 2005; Shepherd et al., 2005). There is growing evidence to indicate that these EGFR inhibitors benefit a minority of advanced NSCLC patients, likely as a result of the fact that only a subset of tumors are dependent on this signaling pathway. Several retrospective analyses have suggested that patients with EGFR+ tumors by IHC, FISH, and/or mutation status appear to derive greater benefit from erlotinib than patients with EGFR– tumors; however a survival benefit due to erlotinib in the EGFR– subset cannot be excluded because the confidence intervals for the EGFR+, EGFR–, and EGFR unknown subsets are wide and overlap (Cappuzzo et al., 2007; Eberhard et al., 2005; Hirsch et al., 2003a,b). There are also uncertainties regarding the benefit of other currently available therapies (for example, chemotherapy) for these patients and regarding the optimal methodology for EGFR determination (e.g., technique, antibody, scoring system, etc.). Recent reports from controlled trials of chemotherapy have not clearly and consistently established the predictive role of EGFR FISH in the setting of treatment with chemotherapy, likely due to the small numbers of patients that had tissue available for post-hoc FISH analysis. Moreover, EGFR FISH represents a poor prognostic factor in untreated NSCLC patients (Hirsch et al., 2003b), and thus the increased EGFR gene copy number by FISH might be a good predictive marker for chemotherapy as well, which could explain the lack of EGFR TK inhibitor superiority in the chemotherapy trials. Given this background, a prospective treatment by marker interaction design to validate EGFR FISH as a predictive and/or prognostic marker for clinical benefit from EGFR inhibitor erlotinib is essential.

A national cooperative group trial, MARVEL (Marker Validation for Erlotinib in Lung Cancer) was recently activated that evaluates the clinical predictive utility of EGFR copy number as measured by FISH. This trial employs the predictive treatment by marker interaction design, where patients with advanced NSCLC will be randomized to Pemetrexed or Erlotinib within both the FISH positive and the FISH negative groups. Tumor samples will be classified as FISH positive if the EGFR gene copy number is high, which is defined as high polysomy (four copies in 40% of cells) or gene amplification (presence of tight gene clusters, a gene-to-chromosome ratio per cell of 2, or 15 copies of EGFR per cell in 10% of cells analyzed). Prior to beginning study therapy, patients will submit tumor tissue for FISH analysis along with IHC testing. Analysis of survival outcomes among patients with or without the prespecified

molecular markers will ascertain the presence or absence of true differences in objective clinical endpoints due to treatment and not due to inherent tumor behavior conferred by the molecular marker. If there is a true difference, the randomization of treatment will reveal superiority of one treatment arm over the other based on the prespecified molecular marker. Given that the primary goal of this trial is to establish the predictive value of EGFR FISH for selection of an EGFR inhibitor therapy (i.e., Erlotinib), an enrichment design strategy utilizing only the FISH positive subgroup would fail to identify patients that do not benefit from the therapy. Specifically, patients that are deemed FISH negative will never be studied in an enrichment design, thus the benefit or no benefit of the therapy in that subgroup can never be established. An unselected treatment by marker interaction design is more appropriate in this setting.

2.3. Example 3—KRAS as a Marker for Panitumumab in Colorectal Cancer

Kirsten rat sarcoma (KRAS) mutations are among the most common oncogenic alterations in cancer. Retrospective data from at least three monotherapy Phase II trials, as well as a Phase III RCT of Panitumumab vs. best supportive care in metastatic colorectal cancer patients, indicate that only patients with wild type KRAS derive benefit from Panitumumab monotherapy. In the randomized Phase III trial, KRAS mutation information was available on 92% (427/463) of the patients enrolled, and it was available on a majority of patients from the Phase II trials, thus making the retrospective analyses statistically sound (Amado et al., 2008a,b; Freeman et al., 2008; Hecht et al., 2007; Van Cutsem et al., 2007). The hazard ratio for treatment effect (panitumumab vs. best supportive care) on progression-free survival (PFS) in the wild type and mutant subgroups was 0.45 and 0.99, respectively, in the Phase III trial (Amado et al., 2008b). Based on this strong evidence, there is growing consensus for prospective KRAS genotyping for all metastatic colorectal cancer patients who are considering receiving Panitumumab monotherapy (Amado et al., 2008a,b; Freeman et al., 2008). In fact, the label for Panitumumab monotherapy has been restricted to KRAS wild type patients in Europe. Based on the strong preliminary data from one RCT and multiple Phase II trials, an enrichment design strategy to further test the predictive utility of KRAS in this setting is ethically and scientifically valid. Importantly, this data establishing the need for an enrichment design for future trials in this disease setting with this agent was made possible through thorough analyses of previously completed clinical trials in an unselected population, thus illustrating the complementary nature of the two design approaches.

2.4. Enrichment vs. Unselected Designs Summary

As is clear from these three case studies, an unselected vs. an enrichment design strategy depends on the strength of the preliminary evidence, assay reproducibility, and the clinical question at hand. If there is strong evidence from well-designed retrospective analyses of RCT to indicate that only certain subgroups may benefit from a treatment, such as in the case of KRAS and panitumumab, then an enrichment design strategy to establish benefit in the subgroup of patients is appropriate and ethical. In cases such as EGFR and NSCLC, where the evidence from preliminary data regarding treatment benefit are conflicting and where assay validation and reproducibility are questionable, a prospectively designed trial for predictive marker validation is optimal. The smaller sample size of the enrichment-based clinical trials is clearly appealing; however, as demonstrated in the case of Herceptin in adjuvant breast cancer, this design may leave important clinical and scientific questions unanswered.

Stepping back from the specific clinical situations presented in this section, in general, an enrichment design is appropriate and ethical when 1) therapies have modest absolute benefit in the unselected population but cause significant toxicity; 2) in the absence of selection, therapeutic results are similar whereby a selection design (even if incorrect) would not hurt; 3) an unselected design is ethically impossible based on previous studies; 4) there is compelling

preliminary evidence to suggest that patients with or without that marker profile do not benefit from the treatments in question; and 5) assay reproducibility and accuracy is well established. When one or more of the preceding does not hold, an unselected predictive marker design is more appropriate.

3. Comparison of the Predictive Marker Designs

We now turn to the topic of the choice of the two specific prospective randomization strategies in the setting of an unselected clinical trial. Preliminary work suggests that the treatment by marker interaction design may be superior to the marker-based strategy design in terms of the overall number of events (and hence the total sample size) required (while keeping all the parameters the same for both designs) under specific clinical settings, whereas the opposite may be true in other settings (Sargent et al., 2005). To explore this more thoroughly, we conducted a head to head comparison of the two designs over a wide spectrum of clinically relevant scenarios. We varied the prevalence rates for the marker from 10% to 90% and the hazard ratios for the different comparisons from 1.1 to 3.0. Details of the scenarios considered and results follow.

Let k be the prevalence of the marker with two levels (positive, +; negative, -). Let m_{A+} , m_{A-} , m_{B+} , m_{B-} indicate the median overall survival for marker + and marker - patients receiving treatments A and B, respectively. A schematic representation of the marker by treatment interaction design (design 1) and the marker-based strategy design (design 2) for a two-level marker with prevalence k for marker + subgroup and where $m_{B+} > m_{A+}$ (median overall survival for marker + patients receiving treatment B is greater than the median overall survival for marker + patients receiving treatment A) and $m_{B-} < m_{A-}$ (median overall survival for marker - patients receiving treatment B is lower than the median overall survival for marker - patients receiving treatment A) is given in Figs. 1 and 2. Under a proportional hazards assumption and an exponential survival model, the hazard ratio for the treatment comparison is given by

$$\frac{\text{median survival in treatment arm A}}{\text{median survival in treatment arm B}} = \frac{m_B}{m_A}$$

The total number of events (i.e., deaths) required for a specific type I error rate α and type II error rate β is therefore given by (Collett, 2003)

$$\frac{\text{constant}}{[\log(\text{HR})]^2} = \frac{\text{constant}}{\left[\log\left(\frac{m_A}{m_B}\right)\right]^2}, \text{ where constant (const)} = 4\left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 \quad (1)$$

The total number of events required for design 1 is the sum of the number of events required for the marker + arm comparison and the marker - arm comparison, which is

$$\frac{\text{const}}{\left\{\log\frac{m_{A+}}{m_{B+}}\right\}^2} + \frac{\text{const}}{\left\{\log\frac{m_{A-}}{m_{B-}}\right\}^2} = \text{const} \left(\frac{\left\{\log\frac{m_{A+}}{m_{B+}}\right\}^2 + \left\{\log\frac{m_{A-}}{m_{B-}}\right\}^2}{\left\{\log\frac{m_{A+}}{m_{B+}}\right\}^2 \left\{\log\frac{m_{A-}}{m_{B-}}\right\}^2} \right) \quad (2)$$

The total number of events for design 2 is given by the number of events required for the comparison of the outcome in the marker-based strategy arm vs. the non-marker-based arm. Since the hypothesis in the marker-based strategy design, without loss of generality, is that $m_{B+} > m_{A+}$, and $m_{B-} < m_{A-}$, all marker + patients will receive treatment B and marker – patients will receive treatment A in the marker-based arm. In the non-marker-based arm, $k/4$ and $(1 - k)/4$ proportion of patients will be randomized to treatments A and B, respectively. Therefore, the total number of events required for design 2 is given by

$$\frac{\text{const}}{\left\{ \log \left[\frac{\frac{k}{2}m_{B+} + \frac{1-k}{2}m_{A-}}{\frac{k}{4}(m_{A+} + m_{B+}) + \frac{1-k}{4}(m_{A-} + m_{B-})} \right] \right\}^2} = \frac{\text{const}}{\left\{ \log \left[\frac{2km_{B+} + 2(1-k)m_{A-}}{k(m_{A+} + m_{B+}) + (1-k)(m_{A-} + m_{B-})} \right] \right\}^2} \quad (3)$$

We let the marker prevalence, k , vary from 0.1 to 0.9 by 0.1 and the median overall survival for treatments A and B for the marker+ and marker – patients range from 1 month to 60 months in increments of 1 month. We then calculated, based on Eqs. (2) and (3), the sample size required for all possible clinical trials with these design parameters. We stress that this is not a simulation but rather a calculation and comparison of the sample size required for each cell of our evaluation, where a cell is a specific combination of k , m_{A+} , m_{A-} , m_{B+} , and m_{B-} . For this possible set of trials, Table 1 presents the proportion of times the number of events required for design 1 (treatment by marker interaction design) is larger than the number of events required for design 2 (marker based strategy design). For practical purposes, scenarios 1 and 4 listed in Table 1 would only entail evaluating the treatments in the marker positive patients (i.e., these scenarios would call for an enrichment as opposed to an unselected design) because there is no expected difference in treatments for the marker negative patients. We include these designs here for the sake of completeness.

Overall, design 1 has better efficiency (i.e., lesser number of events and therefore smaller sample size) than design 2, except in scenarios 1 and 4, where both treatments A and B are equally effective on the marker negative patients. As expected, for these 2 scenarios, the number of events required is greater for design 1 as compared to design 2 for most prevalence rates because the sample size approaches infinity when $m_{A-} \approx m_{B-}$ in the marker negative arm. In scenarios 2 and 6, design 1 requires fewer events than design 2 under all marker prevalence rates. In scenarios 3 and 5, design 1 requires fewer events than design 2 in most cases. Thus, the marker by treatment interaction design performs better than the marker-based strategy design in terms of sample size requirements for the scenarios considered.

The reason for the improved efficiency of the treatment by marker interaction design is because the marker-based strategy design fundamentally includes patients treated with the same regimen on both the marker-based and the non-marker-based arms. This overlap results in a significant number of patients receiving the same treatment regimen in both arms, thus reducing the overall hazard ratio between the two arms. The overlap increases as the prevalence of the marker defined subgroups increases. Based on these considerations, in the setting of a binary marker designed to decide between the treatments, the marker by treatment interaction design has greater efficiency than the marker-based strategy design. In the setting of a complex marker with many levels or a marker designed to dictate a choice between more than two treatments, the marker-based strategy design may be preferred, although this needs to be explored in greater detail as done in the preceding for a binary marker (Sargent et al., 2005).

4. Discussion and Future Work

Advancing new discoveries from the bench to the bedside is the ultimate goal of clinical and translational research. The concept of predictive biomarkers moves the field a step closer toward individualized medicine, whereby individuals who are likely to have a favorable clinical outcome such as improved survival and/or decreased toxicity from a treatment can be prospectively identified. In this study, we evaluated the efficiency (in terms of overall sample size and events) of two prospective unselected predictive marker designs in the setting of a single or multimarker signature that can be distilled to a binary measure, under the assumption that the accuracy (sensitivity and specificity) of the assay is perfect. The treatment by marker interaction design performed better in terms of sample size requirements for the scenarios considered, however it remains to be studied if the marker-based strategy design would outperform the treatment by marker interaction design in a multimarker situation or a marker designed to make a choice between multiple treatments. Although the impact of the error in measurement of the biomarker on the efficiency of these designs needs to be explored further, it is likely that it will have a similar effect on either design, inflating the required sample size in both cases due to patient misclassification.

In addition to these two unselected designs, in the setting of a biomarker measured on a continuous or a graded scale, the newly proposed biomarker-adaptive threshold design could be considered. This design allows for the incorporation of a statistically valid hypothesis for identification and validation of a cut point for a prespecified biomarker into an RCT (Jiang et al., 2007). This design preserves the ability to detect an overall treatment effect in the unselected population while allowing for the flexibility to prospectively validate a biomarker without a predefined cutoff for identifying sensitive patients, all within a single RCT. This new design may provide a middle ground between the unselected and enrichment clinical trial designs, at the cost of a somewhat larger sample size and/or redundant power dictated by the strategy of partitioning the overall type I error rate.

Whereas prospectively designed clinical trials to validate a predictive marker may be the ideal approach, testing for a marker by treatment interaction effect utilizing data collected from previously conducted RCT comparing therapies for which a marker is proposed to be predictive is clearly a much more feasible and timely option. In our opinion, such a strategy may be a reasonable alternative when 1) a prospective RCT is ethically impossible based on results from previous trials, and/or 2) a prospective RCT is not logistically feasible (large trial and long time to complete). For such a retrospective analysis to be valid, samples must be available on a large majority of patients to avoid any selection bias in the patients that have or do not have the samples. The hypotheses, analyses techniques, patient population, and precise algorithm for assay techniques must also be stated prospectively using patients already enrolled on the previous RCT. All marker subgroup analyses have to be stated up front, with appropriate sample-size justification. This approach can aid in bringing forward effective treatments to a marker defined subgroup of patients in a timely manner that might otherwise be impossible due to ethical and logistical considerations. If two or more findings from large well-designed retrospective analyses (that meet all of the above mentioned criteria) of data from prospective RCT trials demonstrate consistent findings regarding the predictive validity of a marker, this may be sufficient to establish the predictive utility of the marker and possibly move it into routine clinical practice.

Although there is no one size fits all solution for clinical validation of a biomarker guided therapy, the choice of a clinical trial design must be driven by a combination of scientific, clinical, statistical, and ethical considerations to help answer the fundamental questions: “Is the new treatment superior in all patients or just in the marker-defined subgroup?” and “What is the added value of marker assessment in every patient keeping in mind the prevalence of the

marker, the incremental benefit of treatment selection based on the marker, and the added costs of the Assay?” Further exploration of these designs and the strengths and drawbacks of each is necessary as we strive toward the goal of personalized medicine.

Acknowledgments

Supported in part by the National Cancer Institute Grants: Mayo Clinic Cancer Center (CA-15083), and the North Central Cancer Treatment Group (CA-25224).

References

- Amado, RG.; Wolf, M.; Freeman, D.; Peeters, M.; Van Cutsem, E.; Siena, S.; Suggs, S.; Devercelli, G.; Woolley, M.; Chang, D. Panitumumab (pmab) efficacy and patient-reported outcomes (PRO) in metastatic colorectal cancer (mCRC) patients (pts) with wild-type (WT) KRAS tumor status. Gastrointestinal Cancers Symposium; 2008a. Abstract 278
- Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson SD, Chang DD. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol* 2008b;26(10):1626–1634. [PubMed: 18316791]
- Bonomi PD, Buckingham L, Coon J. Selecting patients for treatment with epidermal growth factor tyrosine kinase inhibitors. *Clin Cancer Res* 2007;13(15 Pt 2):s4606–4612. [PubMed: 17671150]
- Cappuzzo F, Hirsch FR, Rossi E, Bartolini S, Ceresoli GL, Bemis L, Haney J, Witta S, Danenberg K, Domenichini I, Ludovini V, Magrini E, Gregore V, Doglioni C, Sidoni A, Tonato M, Franklin WA, Crino L, Bunn PA Jr, Varella-Garcia M. Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. *J Natl Cancer Inst* 2005;97:643–655. [PubMed: 15870435]
- Cappuzzo F, Ligatorio C, Toschi L, Rossi E, Trisolini R, Paioli D, Magrini E, Finocchiaro G, Bartolini S, Cancellieri A, Hirsch FR, Crino L, Varella-Garcia M. EGFR and HER2 gene copy number and response to first-line chemotherapy in patients with advanced non-small cell lung cancer (NSCLC). *J Thorac Oncol* 2007;2(5):423–429. [PubMed: 17473658]
- Collett, D. Modeling Survival Data in Medical Research. 2nd. Chapman & Hall/CRC; 2003. Chapter 10: Sample size requirements for a survival study; p. 299–311.
- Conley BA, Taube SE. Prognostic and predictive markers in cancer. *Dis Markers* 2004;20(2):35–43. [PubMed: 15322312]
- Eberhard DA, Johnson BE, Amler LC, Goddard AD, Heldens SL, Herbst RS, Ince WL, Jänne PA, Januario T, Johnson DH, Klein P, Miller VA, Ostland MA, Ramies DA, Sebisanoovic D, Stinson JA, Zhang YR, Seshagiri S, Hillan KJ. Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *J Clin Oncol* 2005;23(25):5900–5909. [PubMed: 16043828]
- Freeman, D.; Juan, T.; Meropol, NJ.; Hecht, JR.; Berlin, J.; Van Cutsem, E.; Reiner, M.; Radinsky, R.; Amado, RG.; Peeters, M. Association of Somatic KRAS Gene Mutations and Clinical Outcome in Patients With Metastatic Colorectal Cancer Receiving Panitumumab Monotherapy. 14th European Cancer Conference; 2008. Abstract 3014
- Hecht JR, Patnaik A, Berlin J, Venook A, Malik I, Tchekmedyian S, Navale L, Amado RG, Meropol NJ. Panitumumab monotherapy in patients with previously treated metastatic colorectal cancer. *Cancer* 2007;110(5):980–988. [PubMed: 17671985]
- Hirsch FR, Scagliotti GV, Langer CJ, Varella-Garcia M, Franklin WA. Epidermal growth factor family of receptors in preneoplasia and lung cancer: perspectives for targeted therapies. *Lung Cancer* 2003a; 41(Suppl 1):S29–42. [PubMed: 12867060]
- Hirsch FR, Varella-Garcia M, Bunn PA Jr, Di Maria MV, Veve R, Bremmes RM, Baron AE, Zeng C, Franklin WA. Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis. *J Clin Oncol* 2003b; 21:3798–3807. [PubMed: 12953099]

- 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(13):1036–1043. [PubMed: 17596577]
- Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Stat Med* 2005;24(3):329–339. [PubMed: 15551403]
- Mandrekar SJ, Grothey A, Goetz MP, Sargent DJ. Clinical trial designs for prospective validation of biomarkers. *Am J Pharmacogenomics* 2005;5(5):317–325. [PubMed: 16196501]
- Paik S. Clinical trial methods to discover and validate predictive markers for treatment response in cancer. *Biotechnol Annu Rev* 2003;9:259–267. [PubMed: 14650930]
- Paik, S.; Kim, C.; Jeong, J.; Geyer, CE.; Romond, EH.; Mejia-Mejia, O.; Mamounas, EP.; Wickerham, D.; Costantino, JP.; Wolmark, N. Benefit from adjuvant trastuzumab may not be confined to patients with IHC 3+ and/or FISH-positive tumors: Central testing results from NSABP B-31. *J Clin Oncol; ASCO Annual Meeting Proceedings Part I*; 2007. p. 511
- Pauletti G, Dandekar S, Rong H, Ramos L, Peng H, Seshadri R, Slamon DJ. Assessment of methods for tissue-based detection of the HER-2/neu alteration in human breast cancer: a direct comparison of fluorescence in situ hybridization and immunohistochemistry. *J Clin Oncol* 2000;18(21):3651–3664. [PubMed: 11054438]
- Pegram MD, Konecny GE, O'Callaghan C, Beryt M, Pietras R, Slamon DJ. Rational combinations of trastuzumab with chemotherapeutic drugs used in the treatment of breast cancer. *J Natl Cancer Inst* 2004;96(10):739–749. [PubMed: 15150302]
- Perez EA, Suman VJ, Davidson NE, Martino S, Kaufman PA, Lingle WL, Flynn PJ, Ingle JN, Visscher D, Jenkins RB. HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. *J Clin Oncol* 2006;24(19):3032–3038. [PubMed: 16809727]
- Perez, EA.; Romond, EH.; Suman, VJ.; Jeong, J.; Davidson, NE.; Geyer, CE.; Martino, M.; Mamounas, EP.; Kauffman, PA.; Wolmark, N. Updated results of the combined analysis of NCCTG N9831 and NSABP B-31 adjuvant chemotherapy with/without trastuzumab in patients with HER2-positive breast cancer. *J Clin Oncol; ASCO Annual Meeting Proceedings Part I*; 2007. p. 512
- Pietras RJ, Pegram MD, Finn RS, Maneval DA, Slamon DJ. Remission of human breast cancer xenografts on therapy with humanized monoclonal antibody to HER-2 receptor and DNA-reactive drugs. *Oncogene* 1998;17(17):2235–2249. [PubMed: 9811454]
- Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE Jr, Davidson NE, Tan-Chiu E, Martino S, Paik S, Kaufman PA, Swain SM, Pisansky TM, Fehrenbacher L, Kutteh LA, Vogel VG, Visscher DW, Yothers G, Jenkins RB, Brown AM, Dakhil SR, Mamounas EP, Lingle WL, Klein PM, Ingle JN, Wolmark N. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med* 2005;353(16):1673–1684. [PubMed: 16236738]
- Sargent D, Allegra C. Issues in clinical trial design for tumor marker studies. *Semin Oncol* 2002;29(3):222–230. [PubMed: 12063675]
- Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 2005;23(9):2020–2027. [PubMed: 15774793]
- Sequist LV, Bell DW, Lynch TJ, Haber DA. Molecular predictors of response to epidermal growth factor receptor antagonists in non-small-cell lung cancer. *J Clin Oncol* 2007;25(5):587–595. [PubMed: 17290067]
- Shepard HM, Lewis GD, Sarup JC, Fendly BM, Maneval D, Mordenti J, Figari I, Kotts CE, Palladino MA Jr, Ullrich A, et al. Monoclonal antibody therapy of human cancer: Taking the HER2 protooncogene to the clinic. *J Clin Immunol* 1991;11(3):117–127. [PubMed: 1679763]
- Shepherd FA, Pereira JR, Ciuleanu T, et al. Erlotinib in previously treated non-small-cell lung cancer. *N Engl J Med* 2005;353:123–132. [PubMed: 16014882]
- Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res* 2004;10(20):6759–6763. [PubMed: 15501951]
- Slamon D. Herceptin: increasing survival in metastatic breast cancer. *Eur J Oncol Nurs* 2000;4(Sa):24–29. [PubMed: 12849614]
- Taube SE, Jacobson JW, Lively TG. Cancer diagnostics: decision criteria for marker utilization in the clinic. *Am J Pharmacogenomics* 2005;5(6):357–364. [PubMed: 16336001]

Van Cutsem E, Peeters M, Siena S, Humblet Y, Hendlisz A, Neyns B, Canon JL, Van Laethem JL, Maurel J, Richardson G, Wolf M, Amado RG. Open-label Phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. *J Clin Oncol* 2007;25(13):1658–1664. [PubMed: 17470858]

Figure 1.
Marker by treatment interaction design (design 1).

Figure 2.
Marker based strategy design (design 2).

Proportion of times the number of events required to provide identical power to detect a statistically significant result is greater for a treatment by marker interaction design (design 1) than for a marker-based strategy design (design 2)

Table 1

Scenarios		Proportion of times number of events for design 1 is greater than for design 2									
		Marker prevalence									
Median overall survival	Hazard ratios	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
$m_{B+} > m_{A+};$ $m_{A-} \approx m_{B-}$	[1.2, 1.5]; [0.9, 1.1]	0.21	0.33	0.47	0.59	0.69	0.79	0.87	0.93	0.98	
$m_{B+} > m_{A+};$ $m_{A-} > m_{B-}$	[1.2, 1.5]; [0.67, 0.83]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$m_{B+} > m_{A+};$ $m_{A-} \gg m_{B-}$	[1.2, 1.5]; [0.33, 0.67]	0.50	0.40	0.31	0.22	0.14	0.07	0.03	0.01	0.00	
$m_{B+} \gg m_{A+};$ $m_{A-} \approx m_{B-}$	[1.5, 3.0]; [0.9, 1.1]	0.35	0.61	0.77	0.87	0.93	0.97	0.99	1.00	1.00	
$m_{B+} \gg m_{A+};$ $m_{A-} > m_{B-}$	[1.5, 3.0]; [0.67, 0.83]	0.00	0.01	0.03	0.06	0.13	0.21	0.30	0.41	0.51	
$m_{B+} \gg m_{A+};$ $m_{A-} \gg m_{B-}$	[1.5, 3.0]; [0.33, 0.67]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	