Chinese Journal of Cancer

Perspective

New evidence-based adaptive clinical trial methods for optimally integrating predictive biomarkers into oncology clinical development programs

Robert A. Beckman[1,2] and Cong Chen[3]

Abstract

Predictive biomarkers are important to the future of oncology; they can be used to identify patient populations who will benefit from therapy, increase the value of cancer medicines, and decrease the size and cost of clinical trials while increasing their chance of success. But predictive biomarkers do not always work. When unsuccessful, they add cost, complexity, and time to drug development. This perspective describes phases 2 and 3 development methods that efficiently and adaptively check the ability of a biomarker to predict clinical outcomes. In the end, the biomarker is emphasized to the extent that it can actually predict.

Key words Predictive biomarkers, adaptive clinical trials, evidence-based approach

The future of oncology drug development lies in using predictive biomarkers to identify subsets of patients who will benefit from particular therapies. Increasingly, national health authorities and insurers are demanding value from medicines. For example, the United Kingdom's National Institute for Clinical Excellence demands a cost of less than or equal to 30,000 British pounds per quality-adjusted life-year (QALY). Most cancer medicines, however, are far more expensive. The low value of cancer medicines is largely driven by two factors: (1) the low average benefit of cancer medicines because they benefit only a subset of the population, and (2) the high cost of oncology drug development due to its high failure rate and the need for large pivotal trials to detect small average benefits.

Predictive biomarkers, or responder identification biomarkers, are molecules or other characteristics of a patient or a patient's malignancy that predict increased benefit from a particular drug. Predictive classifiers, which may be constructed from one biomarker or a composite of biomarkers, identify patients more likely to benefit. With increasing knowledge of the molecular biology of cancer, the number and potential of these predictive biomarkers and classifiers are increasing.

Examples of predictive biomarkers of importance in oncology include HER2/neu expression for trastuzumab therapy, sensitizing mutations in the epidermal growth factor receptor (*EGFR*) gene for gefitinib and erlotinib therapy, and wild-type *KRAS* for therapy with cetuximab or panitumumab[1-8]. However, failures of predictive biomarkers have also been reported. In these cases, biomarker use added cost, complexity, and time, and narrowed the treated population unnecessarily. A notable setback was the failure of EGFR expression to predict the efficacy of EGFR-directed antibodies. This anomaly may have been due to insufficient sensitivity, biased sampling, loss of antigen expression with storage, or tumor evolution between the time the biopsy was obtained and when the therapy was applied[9,10], but these issues affect any real world attempt to test a predictive biomarker hypothesis. Thus, predictive biomarker classifiers and the assays used to test them must be robust to these pitfalls.

Authors' Affiliations: [1]Oncology Clinical Research, Daiichi Sankyo Pharmaceutical Development, Edison, New Jersey 08837, USA; [2]Center for Evolution and Cancer, Helen Diller Family Cancer Center, University of California at San Francisco, San Francisco, California 94115, USA; [3]Biostatistics and Research Decision Sciences, Merck & Co., North Wales, Pennsylvania 19454, USA.

Corresponding Author: Robert A. Beckman, Center for Evolution and Cancer, Helen Diller Family Cancer Center, University of California at San Francisco, San Francisco, California 94115, USA. Tel: +1-610-304-5919; Fax: +1-732-906-5690; Email: eniac1@snip.net.

doi: 10.5732/cjc.012.10248

The great promise of predictive biomarkers, together with inconsistent results and the significant investment of time and money required, have led to variable attitudes ranging from uncritical enthusiasm to harsh skepticism[11-13]. The skepticism is well expressed by Ratain and Glassman: "Whereas 'wins' have occurred here··· most attempts to identify such biomarkers have been nothing more than expensive fishing expeditions. Drug response is multifactorial; patient populations are heterogeneous; potential markers are innumerable; and scientific underpinnings to marker development are imperfect."[13]

These issues and legitimate concerns may hinder the development of a field that increases in promise as the molecular understanding of cancer increases. Interdisciplinary drug development teams often fail to reach a consensus on if, when, and how to apply predictive classifiers, and this manifests in present-day clinical trials that lack a meaningful use for these classifiers. The approach presented below was developed after extensive discussions among discovery scientists, translational medicine experts, clinicians, statisticians, regulatory affairs experts, and commercial experts from several pharmaceutical and biotechnology firms, and is inspired by a broad consensus from these discussions (but does not represent the official position of the firms). In this approach, predictive classifier use is actively considered in each case, and early investments are made in preclinical and clinical programs to determine if they predict clinical benefit. Classifiers are then applied in phase 3 trials only to the extent that they can be shown to predict clinical benefit.

This report will first discuss four fundamental strategic principles underlying this approach, and then follow with four tactical innovations for efficient implementation of the strategic principles. Biomarker development can be divided into an exploration phase and a confirmation phase[14]. This report focuses on the confirmation phase.

Strategic Principles

The central goal is to apply predictive biomarker classifiers in exact proportion to the evidence supporting their clinical value. Four strategic principles are fundamental to the approach: (1) maximum efficiency of development based on objective mathematical functions that quantify benefit or utility (such as the number of approved therapeutic indications) per resource unit expended (such as money spent or patients enrolled), (2) adaptive decision making, (3) continuous integration of biomarker and clinical information, and (4) validation of predictive biomarker hypotheses against clinical benefit. Adaptive decision making and continuous integration of biomarker and clinical information are inherent in the tactics discussed later in this and subsequent sections. In particular, adaptive decision making that integrates biomarker and clinical information will be evident in the decision analysis-based transition from phase 2 to phase 3, and in the adaptation within the phase 3 study based on both interim results from phase 3 and maturing results from phase 2 (*the phase 2+ method*). In the strategy sections immediately below, we further explain the principles of maximum efficiency of development and validation of predictive biomarker hypotheses against clinical benefit.

Departing from tradition: development efficiency and type III error

Testing whether predictive biomarkers work requires both biomarker-positive and biomarker-negative patients in late development. This could require more patients in some cases. To conserve resources, we would like to do the most efficient development we can—that is, getting the answer with the smallest number of patients in phases 2 and 3 trials.

A critical step in drug development is the proof of concept (PoC) trial, a phase 2 trial designed to provide an initial test of a particular therapy or combination of therapies in a defined population or indication. Given that there are typically nearly 1,000 approved and experimental therapies for cancer active in clinical trials at any one time, that they can be combined in twos and threes, that different schedules may be used, and that many clinical indications and lines of therapy are available, the number of potential PoC trials that could potentially be performed is enormous. The number only increases when considering possible subsets defined by biomarker classifiers, which must take into account approximately 30,000 genes in the human genome and the genetic instability and consequent heterogeneity of cancer [15-17]. Although preclinical information offers prioritization of these possibilities, there still remain a very large number of potentially useful PoC trials that far exceeds the availability of patients and funding from either public or private sources.

Statisticians traditionally design randomized PoC trials with the concepts of type I and type II error in mind. These refer, respectively, to false positive and false negative rates due to chance as a result of the finite sample size in a PoC trial. False positive results lead to phase 3 trials undertaken in error and will likely produce negative results at great expense. False negative results lead to the wrong conclusion that the drug is ineffective for the indication, resulting in a loss of opportunity. Traditionally, phase 2 PoC trials are designed to have a type I error of 10% and a type II error of 20% (the *power* is 100% minus the type II error, thus 80% *power* in this case). So engrained is this tradition that PoC trials with less than 80% power are

often termed *underpowered*, even though the traditional powering still allows significant room for error, and "perfect" PoC trials would require infinite sample sizes. However, there is no absolute scientific basis for selecting particular types I and II error rates in PoC trials; these are simply a function of risk tolerance, which is in turn a function of strategy. Indeed, we observe an alternative style of smaller underpowered trials being executed in many cases.

Chen *et al.*[18,19] investigated whether optimal type I and type II error rates could be objectively defined by requiring that the efficiency of phase 2 and phase 3 development be maximized. Given the fact that the possible expenditure on PoC trials of interest in oncology usually exceeds available patient and financial resources, efficiency was defined as the risk-adjusted number of truly effective drug/indication combinations identified by PoC trials and developed to approval (benefit) divided by the risk-adjusted number of patients enrolled in phases 2 and 3 trials (cost). The risk-adjusted benefit is diminished by the risk because of missing a truly effective drug that was not recognized as such by the PoC trial (false negative: type II error). The risk-adjusted cost includes the risk of enrolling patients in a costly phase 3 trial involving an ineffective experimental drug presumed to be effective based on the PoC trial (false positive: type I error). In this benefit/cost ratio, cost is defined as number of patients in that patients are the ultimate limiting resource in drug development. Indeed, we wish to find optimal therapies while exposing the fewest patients to the failures that are common in drug development. It should be noted that financial cost is not exactly proportional to the patient number, as there is a fixed start-up cost of even the smallest clinical trial.

The scenario assumed that there was a fixed total patient budget for PoC trials and that there were many more PoC hypotheses of equal merit that could be tested than that could be funded within the budget, if all trials were designed using traditional type I and type II errors. It was further assumed that every positive PoC trial would result in a phase 3 pivotal trial with type I and type II error rates fixed according to health authority requirements. Finally, it was assumed that the PoC trial type I and type II error rates, which (along with the minimum benefit of clinical interest) dictate the PoC trial sample size, can be selected to optimize the efficiency function. The question was essentially: is it better to perform larger PoC trials to minimize the adverse consequences of type I and II statistical errors, or to perform smaller PoC trials so that more valuable hypotheses can be tested under a fixed PoC trial budget?

Surprisingly, it is up to 30% more efficient to reduce the power of PoC studies from 80% to 60%, while maintaining the traditional type I error and correspondingly performing approximately twice as many PoC trials (because of the reduced sample size required compared to traditional PoC trials). This is called *Chen-Beckman power or Chen-Beckman powering*. Further reductions in sample size do not lead to further increases in efficiency. The reason for this surprising result is readily apparent: doing trials with traditional powering, which needs more patient or financial resources per PoC trial, may cause us to not perform some valuable PoC trials due to budgetary restrictions. Not performing a PoC trial may cause us to miss an opportunity to test a hypothesis that might have been successful. This lost opportunity is an invisible mistake but is severely detrimental to the efficiency of development. The resulting opportunity cost has been termed *type III error*[12]. Larger, traditional PoC trials have lower type I and type II errors but higher type III error when performed as part of an overall development program under a fixed PoC budget.

Because the development efficiency function is "flat" as we go to smaller and smaller PoC trials, it is not helpful to look for the actual optimum. Rather, the recommended procedure is to reduce the power and size of the trial until most of the potential efficiency increase is realized. This is generally at 60% power, with a type I error of approximately 10%, assuming that the effect size (degree of clinical benefit) used for power calculations is the minimum clinically significant effect size, such as 33% or 2 months improvement in progression-free survival (PFS), whichever is greater. This is in contrast to the practice of powering on an aspirational or desired larger effect size, so that the traditional power may be achieved with a small, practical sample size. The latter practice uses optimism about the effect size, not always justified, to mask the reality of the phase 2 trial power statistical properties.

The above results assume a variety of equally meritorious hypotheses to be tested. However, based on preclinical information or other considerations, the interdisciplinary drug development team may judge some hypotheses to have greater value and/or probability of success than others. The same mathematics can be used to find the corresponding optimal type I and type II error rates for each hypothesis and the optimal allocation of resources in the case of hypotheses of unequal merit, such that PoC trials corresponding to the strongest or most valuable hypotheses get more than their share of resources, others less resources, and PoC trials corresponding to the weakest or least valuable hypotheses may not be done. For example, if two hypotheses are of equal merit, it is more efficient to do two trials at *Chen-Beckman power* rather than to test only one hypothesis at traditional power. But if the hypotheses are of unequal merit, the mathematics

suggests devoting more resources to testing the better hypothesis. If the difference in relative merit of the hypotheses is very great, the mathematics may suggest devoting all the resources to the better hypothesis, mirroring the traditional paradigm. The mathematical approach can further incorporate the correlation between the endpoint for phase 2 (for example, PFS) and the primary variable of interest in phase 3 [for example, overall survival (OS)][20].

Using this optimal sizing of phase 2 trials has certain advantages and disadvantages, in addition to the efficiency considerations outlined above. Smaller phase 2 trials can lead to faster enrollment, accelerating the appearance of successful drugs and greatly benefiting patients who need additional treatment options as soon as possible. More rapid drug development is also beneficial for pharmaceutical and biotechnology companies. The effect of smaller phase 2 trials on motivating phase 3 investigators in the subsequent trial is another important issue. Optimal cost-effectiveness with less power also corresponds to a higher empirical bar for "go–no go" decisions that determine whether a drug will advance to phase 3 development [18,19]. Thus, phase 3 investigators may have greater motivation to enroll phase 3, which will be supported by a larger observed phase 2 effect size than usually required for a "go." Conversely, these same phase 3 investigators may have less motivation to enroll phase 3 if they find a larger phase 2 trial more convincing. Finally, smaller phase 2 trials decrease the precision of estimating the probability of phase 3 success and the clinical benefit effect size. We do not recommend designing phase 3 around the estimated effect size based on phase 2 results; rather, we recommend designing phase 3 around the minimum clinically significant effect size (such as 25% improvement in OS or 2 months, whichever is greater) for this reason. Exceptions to this recommendation may be considered in the case of particularly compelling phase 2 results.

When a predictive biomarker classifier is validated in a PoC study, we are really testing two PoC hypotheses: a PoC hypothesis concerning the drug and a PoC hypothesis concerning the biomarker classifier. This would, in principle, double the size of the PoC study because one needs to perform statistical tests on both biomarker-positive and biomarker-negative patient subgroups. However, by powering each of the subgroups at the Chen-Beckman power, the sample size again becomes manageable.

Validation of predictive biomarker classifiers: a moratorium on fishing and the importance of iteration

To demonstrate the value of a predictive biomarker classifier, the classifier must be validated against patient clinical benefit. In the program described herein, this validation is formal statistical validation, beginning with the randomized PoC trial, and continuing, if indicated by the data, through to a phase 3 pivotal trial that definitively validates the biomarker classifier as a predictor of clinical benefit. This would result in simultaneous health authority approval for both the drug and the associated diagnostic test for patient selection [co-diagnostic or *in vitro* diagnostic (IVD)]. By incorporating formal statistics into the predictive classifier evaluation in a randomized phase 2 study, we are able to optimally manage risk across a portfolio of drugs and putative biomarker classifiers (see "decision analysis guided phases 2–3 predictive biomarker transition" below).

Formal statistical analysis precludes "fishing expeditions" as described by Ratain and Glassman, in which a large number of biomarkers are informally tested in an exploratory fashion. In this exploratory approach, many possible biomarker hypotheses are simultaneously tested, and the possibility that one of these hypotheses will appear to be true based on chance alone is extremely high. This multiple hypothesis problem leads to an unacceptably high type I error rate in typical biomarker studies. Moreover, if the hypotheses are not specified prospectively, but rather after the study data is available, the chance of crafting an artifactually successful predictive biomarker hypothesis is even higher.

Therefore, we require that one primary predictive biomarker hypothesis be specified in advance. This hypothesis is termed the primary predictive biomarker clinical benefit identification (ID) hypothesis. The randomized PoC study is formally powered around the subgroups defined by this biomarker classifier, at powering designed for optimal efficiency. Ideally, the primary predictive biomarker hypothesis will be specified prior to the start of the randomized PoC study, as it is helpful in the study design. If necessary, however, its specification can be delayed until just before the samples are analyzed at the end of the study, the prospective retrospective approach[21]. Clinical benefit is stated in general terms and may refer to tumor shrinkage (response), increased PFS or OS, or even enhanced health-related quality of life.

The time just prior to the end of phase 2 sample analysis is also the latest time by which assays to determine biomarker status must be available. Ideally, these assays will be analytically validated in terms of sensitivity, specificity, reproducibility, accuracy, and linearity[22] at this time. Such analytically validated assays require only validation against clinical data to be approved as IVDs, and are termed *IVD candidates*. If absolutely necessary, the phase 2 samples can be analyzed with a non-analytically validated "research use

only" assay and the IVD candidate assay delivered at the time of analysis of phase 3 samples (interim or final analysis), but this entails the risk of misleading results from the phase 2 biomarker evaluation.

Clinical validation of a predictive biomarker classifier involves the use of at least some patients who are biomarker-negative to verify that the diagnostic test can distinguish between those who will benefit and those who will not; this is necessary because the test will be used in the future to guide patient selection, including advising biomarker-negative patients not to use a therapy. Testing biomarker negative patients raises ethical issues if the confidence in the proposed biomarker classifier is believed to be high. Accordingly, we recommend that predictive biomarker classifier testing be performed as an add-on design, employing standard-of-care therapy with and without experimental therapy on the two respective study arms, so that all patients receive standard-of-care therapy. In some cases, add-on to standard-of-care is not recommended due to possible pharmacologic antagonism between the standard-of-care and the proposed experimental agent. In these cases, an add-on design with a second experimental agent predicted not to antagonize the first experimental agent may be indicated. The objective is to design a rigorous, randomized test of the first experimental agent without compromising the rights of patients who are biomarker-negative for this agent to receive therapy that is thought to have a chance of being effective.

In some cases, the inclusion of biomarker-negative patients may dilute the efficacy signal from a drug that is effective in only a subpopulation. Thus, trastuzumab would have likely failed to achieve PoC in a mixed population. Because oncology drug development usually entails several indications being studied in parallel, a parallel study may look at a different indication in biomarker-positive patients only, to optimize the chance of PoC for the drug.

The argument to include bio−marker-negative patients assumes a certain degree of *equipoise*—that is, uncertainty about the truth or falsity of the biomarker-based clinical benefit ID hypothesis. Equipoise is often underestimated in real situations (e.g., in cases where the predictive biomarker clinical benefit ID hypothesis is invented by people on the development team who may not objectively recognize the inconsistent translation of preclinical results to the clinic). Moreover, publication bias leads to more frequent and prominent publication of biomarker success stories than cautionary tales. The surprising failure of EGFR expression to clearly segregate patients who would benefit from anti-EGFR antibody therapy from those who would not argues that even the most "obvious" biomarker-based clinical benefit ID hypotheses require validation. We must remember that even the most well supported clinical benefit ID hypothesis will be limited by assay performance under real world clinical conditions.

The primary predictive biomarker clinical benefit ID hypothesis is chosen based on data from preclinical studies, phase 1 clinical studies; phase 2a exploratory, non-randomized studies; neoadjuvant studies where tissue for exploratory biomarker work can be readily obtained; and where applicable, experimental medicine studies in patients or volunteers. The selected hypothesis should be the one best supported by the scientific rationale and data to that point. Moreover, studies of tissue banks should have determined the expected prevalence of biomarker-positive and -negative subgroups in the proposed PoC indications. If the biomarker-positive subgroup is too small, it may be difficult to enroll a suitable trial, and if the biomarker-negative subgroup is too small, it may not be cost effective to screen when the error rate of the assay is considered. Ideally, the biomarker classifier should be designed such that biomarker-positive and -negative subgroups are both sufficiently large in an unselected population. For this reason, it is preferable for study design if the primary clinical benefit ID hypothesis can be specified prior to the start of the phase 2 study rather than before the assay of samples at the end.

Given the intricacies of cancer biology, choosing one primary clinical benefit ID hypothesis will not be a foolproof exercise. As a backup, additional predictive biomarker hypotheses may be tested in an exploratory fashion in the randomized PoC study. If the primary clinical benefit ID hypothesis is false, perhaps one of these exploratory hypotheses will generate promising data. Such new findings represent a lower level of evidence and should be validated in a second randomized PoC study designed with the new clinical benefit ID hypothesis as primary. This represents another iteration through phase 2 development. However, such iteration should not be viewed as failure. Successfully unraveling the complexities of biomarker-directed cancer therapy will require persistence. Just as backup compounds are available in the event of failure of a lead compound in development, backup predictive biomarker-based clinical benefit ID hypotheses, which need to be validated with a subsequent iteration of phase 2, should be planned for and expected.

Tactics

Efficiency-optimized, biomarker-stratified, randomized phase 2 PoC study

Biomarker-directed randomized PoC studies fall into three categories—enrichment, biomarker stratified, and biomarker strategy—and their efficiency has been eva-

luated[23]. In the enrichment study, the randomized study is performed in biomarker-positive patients only. This practice is most efficient if there is very high confidence in the clinical benefit ID hypothesis. As discussed, we would recommend it only rarely as equipoise around the clinical benefit ID hypothesis tends to be underestimated. In the biomarker-stratified design, enrollment is stratified by the biomarker status and the results evaluated in each independent stratum. This is the most efficient design when there is equipoise concerning the predictive biomarker based clinical benefit ID hypothesis. Finally, the biomarker strategy design randomizes patients between two patient allocation strategies: biomarker-directed allocation based on the clinical benefit ID hypothesis, or standard randomization. This latter strategy has been shown to be less efficient because the patient can be randomized to the same therapy on either arm, diluting comparisons. Moreover, a difference between arms may occur due to efficacy of the experimental therapy, even when the clinical benefit ID hypothesis is false.

The tactic utilized herein is a biomarker-stratified design, with both biomarker-positive and -negative strata powered at the Chen-Beckman level, thus optimizing the efficiency of simultaneously testing two strata (or, equivalently, two hypotheses—one concerning drug efficacy and the other, the clinical benefit ID hypothesis)[18,19]. The strata are defined based on a single prospectively-specified, primary predictive biomarker based clinical benefit ID hypothesis. The efficiency of the study is greatest when neither the biomarker-positive nor biomarker-negative subgroups are too small, with optimal efficiency at a 50−50 split. This should be borne in mind when choosing the indication and designing the clinical benefit ID hypothesis and classifier. Additional backup biomarkers may be assayed in the usual exploratory fashion. The arms should be standard-of-care with or without experimental therapy (an add-on design), so that biomarker-negative patients are guaranteed standard-of-care. In instances where there is concern about possible pharmacologic antagonism with standard-of-care, one might consider an add-on design with a second experimental therapy that is expected to be additive or synergistic with the agent in question.

We recommend PFS as the primary endpoint in most cases. In contrast to response, it is a continuous variable and is informative in all patients, which is particularly helpful with targeted therapies that may not influence response rate. PFS may also correlate more closely than response with the primary variable of interest, OS[24].

Because the availability of fresh tissue specimens cannot be guaranteed, and because circulating tumor cells[25], while promising, may not be recoverable in sufficient quantity, we currently recommend that the focus be on developing assays that can work in archival formalin-fixed, paraffin-embedded specimens. Nonetheless, such specimens have several disadvantages: (1) some biomarkers may not be assayable under these conditions or may degrade in storage, and (2) the tumor may have evolved to a different biomarker status between the time the specimen was obtained and the present. Hence, the wider availability and applicability of fresh frozen specimens and/or circulating tumor cells is eagerly anticipated. A new fixative that allows better preservation of phosphoprotein biomarkers and tissue morphology may offer an advantage over formalin fixation in this regard[26].

Waiting for the delivery of an archival tumor specimen, which may have been obtained at another institution, and for biomarker tests to be performed on that specimen may be intolerable for patients with rapidly progressing disease, and they may choose another, simpler study. In this regard, it is better to require only documented tissue availability upfront, not assay results for pre-selection. The indication and biomarker classifier must then be designed such that the subgroups will be relatively close to 50−50 without pre-selection. A blinded interim analysis to assure balance between strata and adequate numbers in each subgroup may be done, and a serious imbalance may have to be corrected through pre-selection of patients in the remainder of the study.

Decision analysis-guided phase 2−phase 3 predictive biomarker transition

In a decision analysis-guided phase 2−phase 3 predictive biomarker transition [12], evidence supporting clinical benefit for biomarker-positive and biomarker-negative patients contributes to a two-dimensional decision rule, plotted on a two-dimensional graph of *P* values for efficacy in biomarker-positive and biomarker-negative patients, respectively. Regions of the graph correspond to four possible choices: (1) no go, when the drug is clearly ineffective in both biomarker-positive and biomarker-negative patients; (2) go to biomarker-positive only enriched phase 3, when the drug is effective and the clinical benefit ID hypothesis is clearly true (e.g., when the drug is clearly and statistically effective in biomarker-positive patients and clearly ineffective in biomarker-negative patients); (3) go to traditional, unselected phase 3, when the drug is effective but the clinical benefit ID hypothesis is false (e.g., when the drug equally demonstrates clear, statistically significant efficacy in both biomarker-positive and biomarker-negative patients); or (4) go to a biomarker adaptive phase 3, when the drug is effective but the results concerning the clinical benefit ID hypothesis are equivocal (e.g., when the drug is clearly and statistically effective in the biomarker-positive group and shows a non-statistically significant trend towards efficacy in the

biomarker-negative group).

To draw the regions in the graph, the team assigns utility values to possible outcomes resulting from these choices. Regions are drawn to maximize risk-adjusted utility while maintaining the type I error at the level chosen with the original study powering. For example, possible outcomes include: (1) approval of the drug in the full population; (2) approval of the drug in the biomarker-positive population when the clinical benefit ID hypothesis is true; (3) approval of the drug in the biomarker-positive population only when the clinical benefit ID hypothesis is false (opportunity cost due to overemphasis on the clinical benefit ID hypothesis); or (4) failure to approve the drug in the full population when the clinical benefit ID hypothesis is true (due to inappropriate emphasis on the full population diluting the clinical benefit in the biomarker-positive population).

We note that in practical terms, it is the drug development team (usually from a biotechnology or pharmaceutical company) that defines the utilities of possible outcomes as an input to the decision analysis. However, patients, insurers, or society as a whole may view the utilities differently compared with the drug development team. Research into these differences could lead to objective criteria for utility that may be broadly applied as a best practice.

In line with the strategic principles, the decision analysis-guided phase 2–phase 3 predictive biomarker transition is the first point where the clinical development program adapts based on integration of clinical and biomarker data.

Adaptive predictive performance-based hypothesis prioritization in phase 3

The efficiency-optimized, biomarker-stratified, randomized phase 2 study attempts to definitively test two hypotheses simultaneously: the hypothesis concerning drug efficacy and that concerning the truth of the clinical benefit ID hypothesis. However, given that phase 2 studies are not definitive, the truth or falsehood of the clinical benefit ID hypothesis may still be unknown at the end of phase 2. In this case, the program proceeds to a further adaptive design in phase 3 to resolve this question.

In this adaptive phase 3 study, two hypotheses are being tested simultaneously: (1) the drug is effective in the full population, or (2) the drug is effective in the biomarker-positive subset only. The study enrolls the full population so that both hypotheses can be tested.

The total type I error rate (false positive rate) for both hypotheses combined is set at 5%, as required by health authorities. This raises the question of how to divide the rate between the two hypotheses. The hypothesis to which more of the type I error is assigned is effectively prioritized or emphasized in the final statistical analysis. Previous approaches have either arbitrarily assigned 4% to the full population hypothesis (fixed allocation)[27] or assigned all 5% to the hypothesis that is better supported at an interim analysis (all or none adaptive allocation)[28]. In the approach described below, the split of the type I error is neither fixed nor all or none. At an interim analysis point, the data are applied to determine the optimal allocation based on maximizing an objective efficiency function, which can be phase 3 study power or phase 3 study power weighted by indication size (e.g., more weight to the full population hypothesis, which has the potential to benefit more people). Thus, the clinical benefit ID hypothesis is prioritized or emphasized to the exact degree that is justified by its predictive performance to that point in development. Another tactical innovation, the *phase 2+* method, is deployed so that the maturing phase 2 data may be combined with the phase 3 data to the interim point and therefore be considered in determining the allocation of type I error (see below). The term *phase 2+* indicates that the phase 2 data is given added significance by allowing it to influence an adaptation within the phase 3 trial.

The allocation is determined at the interim analysis by an independent review board according to pre-specified rules that must be approved by health authorities in advance. The adaptation is only to the analysis strategy and does not affect patient selection or management. The results from the interim analysis will not be used to claim efficacy. Type I error is strictly controlled. Based on these safeguards, we believe this approach will be acceptable to ethics committees, institutional review boards, and national health authorities.

Adaptive predictive performance-based hypothesis prioritization in phase 3 is a second example of the strategic principle of adaptation by integration of biomarker and clinical information.

The phase 2+ method for allowing maturing phase 2 data to influence adaptation within phase 3

Continuous adaptation in response to data is a cornerstone of our proposed strategy. Nonetheless, frequent adaptation in oncology is hampered because the primary endpoint of greatest interest, OS, takes significant time to collect. More rapid endpoints are of interest only to the degree that they have some predictive ability for survival.

In the traditional approach, a decision is made to continue to phase 3 based on PFS in phase 2. Then, if there is an adaptation within phase 3, it is at an interim analysis where minimal OS data are available, and therefore is based on PFS. However, PFS is an imperfect surrogate for OS. The phase 2 OS data, which

may have matured substantially by that point, does not contribute to the phase 3 adaptation.

In the approach recommended here, maturing PFS and OS data from the phase 2 study are combined with the phase 3 data at the time of the interim analysis in phase 3, so that all clinical information to that point contributes to the phase 3 adaptation [29,30]. Alternatively, given that PFS is an imperfect surrogate for OS, one may choose to adapt in phase 3 based solely on phase 2 OS. In this way, one is making no assumptions about the correlation between PFS and OS in the analysis. Because of the safeguards described above, we do not believe there will be objections from ethics committees or institutional review boards on the use of data outside the phase 3 study they are regulating.

The phase 2 study may thus have up to three analysis points: (1) a primary analysis point for PFS that results in a go–no go decision to phase 3, (2) a final analysis when OS data are also mature, and (3) an analysis to support interim decision making in phase 3, if the final phase 2 OS analysis has not occurred by that point. The possibility of these three analyses must be pre-specified to avoid concerns about "cherry picking" phase 2 analysis points. The phase 3 interim analysis using phase 2 and phase 3 data will not be used to support efficacy claims.

Discussion

We have devised methods to adaptively integrate predictive biomarkers into oncology clinical development programs in a data-driven manner. These biomarkers are emphasized in exact proportion to the evidence supporting their clinical predictive value. The program is built on four strategic principles: (1) maximum objective efficiency functions based on utility per resource unit expended, (2) adaptive decision making, (3) continuous integration of biomarker and clinical information, and (4) validation of predictive biomarker hypotheses against clinical benefit.

Central to this paradigm is the selection of a single primary predictive biomarker clinical benefit ID hypothesis that will be subjected to statistical validation as a predictor of clinical benefit in a randomized phase 2 PoC study. This hypothesis may be based on a single biomarker or a defined composite of biomarkers. The primary hypothesis and related assays to determine biomarker status should ideally be available at the beginning of this phase 2 study, but it is possible to wait until the sample analysis at the end. Ideally, an analytically validated IVD candidate assay will be available for this purpose. This is required by the time of interim analysis in phase 3.

The tactical innovations described herein maximize the use of integrated biomarker and clinical data for adaptation of the development program. These innovations must be pre-specified, and the phase 3 design pre-approved by health authorities. The goal of co-registering the drug with a diagnostic assay that determines the optimal patient population is approached in an efficient manner.

This approach is exacting and demands value from predictive biomarkers, steering a middle course between uncritical enthusiasm and harsh skepticism. The approach could lead to rejection of putative predictive biomarkers and/or the need to iterate through development a second time with a new predictive biomarker hypothesis, but in the end it is expected to lead to more enduring value.

Acknowledgments

The authors wish to thank Jason Clark for contributing the decision analysis-guided phase 2–phase 3 predictive biomarker transition, and Donald Bergstrom, Robert Phillips, and Richard M. Simon for helpful discussions.

Received: 2012-10-09; revised: 2013-01-01;
accepted: 2013-01-28.

References

[1] Cobleigh MA, Vogel CL, Tripathy D, et al. Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease. J Clin Oncol, 1999,17:2639–2648.

[2] Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med, 2001,344:783–792.

[3] Paez JG, Janne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science, 2004,30:1497–1500.

[4] Pao W, Miller V, Zakowski M, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci USA, 2004,101:13306–13311.

[5] Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol, 2008,26:1626–1634.

[6] Lievre A, Bachet JB, Le Corre D, et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res, 2006,66:3992–3995.

[7] Bokemeyer C, Bondarenko I, Hartmann JT, et al. KRAS status and efficacy of first-line treatment of patients with metastatic colorectal cancer (mCRC) with FOLFOX with or without cetuximab: the OPUS experience. J Clin Oncol, 2008,26 Suppl 15:4000.

[8] Van Cutsem E, Lang I, D'haens G, et al. KRAS status and

efficacy in the first-line treatment of patients with metastatic colorectal cancer (mCRC) treated with FOLFIRI with or without cetuximab: The CRYSTAL experience. J Clin Oncol 2008: 26 Suppl 15:2.
[9] Yan L, Beckman RA. Pharmacogenetics and pharmacogenomics in oncology therapeutic antibody development. Biotechniques, 2005,39:565–568.
[10] Ransohoff DF, Gourlay ML Sources of bias in specimens for research about molecular markers for cancer. J Clin Oncol, 2010,28:698–704.
[11] Dalton WS, Friend, SH. Cancer biomarkers—an invitation to the table. Science, 2006,312:1165–1168.
[12] Beckman RA, Clark J, Chen C. Integrating predictive biomarkers and classifiers into oncology clinical development programmes. Nat Rev Drug Discov, 2011,10: 735–749.
[13] Ratain MJ, Glassman RH. Biomarkers in phase I oncology trials: signal, noise, or expensive distraction? Clin Cancer Res, 2007,13:6545–6548.
[14] Orloff J, Douglas F, Pineiro J. et al. The future of drug development: advancing clinical trial design. Nat Rev Drug Discov 2009; 8: 949–957.
[15] Beckman RA. Efficiency of carcinogenesis: in the mutator phenotype inevitable? Semin Cancer Biol, 2010,20:340–352.
[16] Loeb LA, Bielas JH, Beckman RA. Cancers exhibit a mutator phenotype: clinical implications. Cancer Res, 2008,68:3551–3557.
[17] Beckman RA, Loeb LA. Genetic instability in cancer: theory and experiment. Semin Cancer Biol, 2005,15:423–435.
[18] Chen C, Beckman RA. Optimal cost-effective Go–No Go decisions in late stage oncology drug development. Stat Biopharm Res, 2009,1:159–169.
[19] Chen C, Beckman RA. Optimal cost-effective Phase II proof of concept and associated Go–No Go decisions. J Biopharm Stat, 2009,1:424–436.
[20] Chen C, Sun L, Chih C. Evaluation of early efficacy endpoints for proof-of-concept Trials. J Biopharm Stat, 2013,23:413–424.
[21] Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst, 2009,101:1446–1452.
[22] Dahle-Smith A, Petty RD. Developing predictive biomarkers in oncology: how can we achieve consistent success? In: Treatment Strategies-Oncology. London, UK: The Cambridge Research Centre, 2010:47–54.
[23] Freidlin B, McShane L, Korn EL. Randomized clinical trials with biomarkers: design issues. J Natl Cancer Inst, 2010,102:152–160.
[24] Tang PA, Bentsen SM, Chen EX, et al. Surrogate endpoints for median overall survival in metastatic colorectal cancer: literature-based analysis from 39 randomized controlled trials of first-line chemotherapy. J Clin Oncol, 2007,25: 4562–4568.
[25] Maheswaren S, Sequist LV, Nagrath S, et al. Detection of mutations in EGFR in circulating lung-cancer cells. N Engl J Med, 2008,359:366–377.
[26] Mueller C, Edmiston KH, Carpenter C et al. One-step preservation of phosphoproteins and tissue morphology at room temperature for diagnostic and research specimens. PLoS One, 2011,6:e23780.
[27] Freidlin B, Simon RM. 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.
[28] Simon RM. The use of genomics in clinical trial design. Clin Cancer Res, 2008,14:5984–5993.
[29] Chen C, Beckman RA. Hypothesis testing in a confirmatory Phase III trial with a possible subset effect. Stat Biopharm Res, 2009,1:431–440.
[30] Chen C, Sun L. On quantification of PFS effect for accelerated approval of oncology drugs, Stat Biopharm Res, 2011,3:434–444.