The Pharmacogenomics Journal (2006) 6, 166–173

www.nature.com/tpj

ORIGINAL ARTICLE

Use of genomic signatures in therapeutics development in oncology and other diseases

R Simon[1,3] and S-J Wang[2,3]

[1]*Biometric Research Branch, Division of Cancer Treatment & Diagnosis, National Cancer Institute, Bethesda, MD, USA and* [2]*Office of Biostatistics, Office of Pharmacoepidemiology and Statistical Science, Center for Drug Evaluation & Research, U.S. Food and Drug Administration, Rockville, MD, USA*

Correspondence:
Dr R Simon, Biometric Research Branch, Division of Cancer Treatment & Diagnosis, National Cancer Institute, NIH, MSC 7434, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA.
E-mail: rsimon@nih.gov

Pharmacogenomics is the science of determining how the benefits and adverse effects of a drug vary among a target population of patients based on genomic features of the patient's germ line and diseased tissue. By identifying those patients who are most likely to respond while eliminating serious adverse effects, the therapeutic index of a drug can be substantially increased. This may facilitate demonstrating the effectiveness of the drug and may avoid subsequent problems due to serious adverse events. Our objective here is to provide clinical trial designs and analysis strategies for the utilization of genomic signatures as classifiers for patient stratification or patient selection in therapeutics development. We review methods for the development of genomic signature classifiers of treatment outcome in high-dimensional settings, where the number of variables available for prediction far exceeds the number of cases. The split-sample and cross-validation methods for obtaining estimates of prediction accuracy in developmental studies are described. We present clinical trial designs for utilizing genomic signature classifiers in therapeutics development. The purpose of the classifier is to facilitate the identification of groups of patients with a high probability of benefiting from it and avoiding serious adverse events. We distinguish exploratory analysis during the development of the genomic classifier from prospective planning and rigorous testing of therapeutic hypotheses in studies that utilize the genomic classifier in therapeutics development. We discuss a variety of clinical trial designs including those utilizing specimen collection and assay prospectively for newly accrued patients and those involving a prospectively planned analysis of archived specimens from a previously conducted clinical trial. Our discussion of the development and use of classifiers of efficacy is mostly focused on applications in oncology using classifiers based on biomarkers measured in tumors. Some of the same considerations apply, however, to development of efficacy and safety classifiers in nononcologic diseases based on single-nucleotide germline polymorphisms.
The Pharmacogenomics Journal (2006) **6,** 166–173. doi:10.1038/sj.tpj.6500349; published online 17 January 2006

Keywords: clinical trial; gene expression; genomic signature; oncology; study design; validation

Introduction

Our objective here is to provide clinical trial designs and analysis strategies for the utilization of genomic signatures for patient selection and stratification in therapeutics development. Effective use of such signatures can lead to substantial improvements in the efficiency of drug development, and to major increases in

[3]Both authors contributed equally to this manuscript.

Received 4 May 2005; revised 20 September 2005; accepted 6 October 2005; published online 17 January 2006

the proportion of treated patients who benefit from the drug, including increased efficacy at an acceptable toxicity level. Conventional approaches to therapeutics development are effective for identifying treatments that work 'on average' for a population of patients similar to those on whom it was tested. In many cases, however, for example with cancer treatments, this involves treating the many for the benefit of the few. It has been estimated that only about 60% of prescriptions written produce the desired therapeutic benefits, and 7% of patients sustain serious drug-related injuries or death.[1,2] Increasing economic stress on national healthcare systems makes the inefficiency of conventional shotgun targeting of therapeutics increasingly nonsustainable.

Rational targeting of new therapeutics using genomic biomarkers to identify the patients who are most likely to benefit and avoid serious adverse effects is increasingly possible. Technology such as high throughput genotyping or transcript expression profiling has resulted in a vast increase in the data available for biomarker development. There appears to be, however, substantial confusion about how to validate biomarkers and to utilize biomarkers for therapeutics development in a validated manner. We will attempt to clarify this issue so that the new technology can be efficiently applied in a manner that leads to improvements in human health.

Biomarkers have been previously defined in very general terms as characteristics that are objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention, and criteria have been defined for the validation of biomarkers.[3] In this paper, instead of addressing the validation of disease biomarkers in such an absolute sense, we will focus attention on the use of genomic classifiers for selecting or stratifying patients in the clinical development of a therapeutic. In the next section, we discuss methods for developing genomic classifiers that can be used to guide treatment selection, and give several examples of such classifiers. We also describe the methods that can be used for 'internal validation', that is to estimate the prediction accuracy of the classifiers. We then discuss the use of genomic classifiers in therapeutics development. We emphasize the importance of prospective planning and testing of prespecified hypotheses in pharmacogenomic trials. We present specific clinical trial designs for using genomic classifiers and describe their relative merits. Prospective planning is essential to all effective trial designs, but in some cases an effective design may utilize archived samples from clinical trials in which the patients have already been accrued. We illustrate prospective and retrospective designs using literature case examples. Concluding remarks follow.

Developing genomic signature classifiers

Multiple steps are generally required to develop a genomic signature to be used for treatment selection. The signature may consist of the protein expression level of a single receptor target of the drug. In many cases, the drug may have multiple targets and there may be no obvious way to measure whether the disease is driven by disregulation of a pathway downstream of a drug target. Consequently, in many cases, it is best to let the data define the genomic signature to distinguish which patients respond to the drug and which do not. For instance, one could power the study to identify at least 80% of truly differentially expressed genes, viz., 80% sensitivity, 90% true discovery rate, or 95% prediction accuracy while accounting for correlation among genes,[4] and then define the genomic signature based on the empirical data.

With the empirical approach to genomic signature development, one measures a large number of 'features' on each treated patient on a body of 'training data' and then selects the features that are most significantly correlated with patient response. The features could, for example, be gene expression levels as determined by a whole genome microarray expression profile of the patient's tumor, or single-nucleotide polymorphisms as measured by genotyping the patients' lymphocytes for a panel of candidate genes. For expression profiling, a statistical significance level is generally computed for the comparison of the logarithm of expression in the responders to the nonresponders, and the most significant genes are selected for inclusion in the classifier. In settings without a binary measure of response, the computed significance levels correspond to univariate tests of association of the logarithm of expression for the genes with some measure of outcome, such as change in blood pressure.

Having selected the features that are most correlated with clinical response to treatment, those features are next combined into a multivariate signature classifier.[5,6] It is important that the genomic signature classifier be reproducibly measurable and be accurate in predicting outcome. It is not essential that each feature selected to be part of the classifier be informative. This is an important distinction, which is often misunderstood. Most statistical methods were developed and traditionally employed for inference problems, not for prediction problems. However, development of a genomic signature classifier is a prediction problem. The statistical significance of the degree of correlation of the features with outcome do not matter in themselves except to the extent that they influence the predictive performance of the signature classifier. For example, with whole genome expression profiling, the features may be selected as those significantly correlated with outcome at the 0.001 significance level. If there are 30 000 genes represented on the array, then the expected number of false positive genes selected is 30. However, there may be 100 genes that are significantly correlated with outcome at the target level. In this case, 30% of the selected features are false positives, but the genomic signature classifier may nevertheless be very accurate. For prediction problems, the omission of informative features generally has a much more serious influence on predictive accuracy than the inclusion of noninformative features. It is preferable, of course, to have all of the selected

features to be informative and for the classifier to be biologically interpretable. In many cases, however, obtaining a simple and biologically interpretable classifier is a much more difficult objective than just accurate prediction of drug response.[7]

There are, of course, numerous algorithms for selecting informative features and for combining them in a signature classifier.[5,6] These range from simple linear classifiers to complex nonlinear ones like neural networks. Dudoit *et al*.[8] found that simple classifiers often perform as well as or better than more complex types since the latter ones tend to overfit the typical data set in which the number of cases is much smaller than the number of candidate features. With linear classifiers, the logarithm of the expression levels of the genes selected for inclusion in the classifier is multiplied by weighting coefficients and added together. One or more cut-points are then used to convert the weighted sum into a risk group specification. For example, Rosenwald *et al*.[9] developed a linear classifier to predict those patients with advanced large B-cell lymphoma, who had good responses to anthracycline-based chemotherapy. They used DNA microarray expression profiling and selected genes for inclusion in the classifier based on correlation of gene expression with disease-free survival in a training set of 160 treated patients. They categorized the functions of the genes that were significantly correlated with patient outcome and selected genes representing each of the functional categories represented. The classifier was based on 16 genes, three characteristic of germinal center B cells, three related to cell proliferation, six reflective of lymph node reactivity due to stromal response and immune cell infiltration, and four characteristic of MHC class II expression. They built a multivariate model relating disease-free survival to a linear risk index consisting of the average of gene expression for each of the four groups of genes. The relative weights for each of the gene groups were determined by optimizing the fit of the model to the data for patients in the training set.

Internal validation

Before using a genomic classifier as the basis for clinical development or for the design of a large phase III trial, one needs some measure of accuracy of the classifier. When the signature is based on a set of features that have been selected from a large number of candidate features, one must exercise special care in measuring accuracy. In many situations, the number of candidate features (p) is greater than the number of patients (n) available for developing the signature. In such a case, 'prediction' accuracy for the same set of patients used to select the features gives an invalid and highly biased estimate of true prediction accuracy for future patients. 'Prediction' for the patients whose data were used to select the features used in the signature is not prediction at all.

The most straightforward way to obtain a valid measure of prediction accuracy for a genomic signature classifier is to apply the classifier to a separate set of patients. For example, the classifier might be developed using patients treated during the initial phase IIA studies of the treatment. The usual initial studies of the drug using short-term end points of efficacy would be modified to include pretreatment tissue sampling, either tumor sampling for oncology studies or normal tissue sampling for genotyping for other diseases. By assaying the sampled tissue and determining which features distinguish responders from nonresponders, or patients experiencing subclinical signs of adverse events from those who do not show such changes, classifiers of response or risk of adverse events may be developed. The prediction accuracy of the classifier developed in the phase IIA studies might be evaluated on patients treated in the expanded phase IIB studies. A completely specified signature classifier should be developed with the first set of patients, and simply used to predict outcome for the second set. The number of prediction errors , the number of responses among the patients predicted to be nonresponders, plus the number of nonresponses among patients predicted to be responders are counted. One cannot expect the confidence limits to be narrow unless the number of patients in the test sample is large. This split-sample method was used by Rosenwald *et al*.[9] for predicting outcome in patients with advanced diffuse large B-cell lymphoma. Their data on 240 patients were divided into a training set of 160 patients and a test set of 80 patients. The expression profiles for the 80 patients in the test set were not used at all until a single completely specified genomic classifier of outcome was developed on the 160 patient training set. They used a linear classifier of 3-year disease-free survival and also evaluated whether their classifier performed better than the International Prognostic Index.

There are also more complicated methods for validly estimating prediction accuracy. Those methods utilize resampling of the total database to repeatedly develop signature classifiers on a training set and evaluate it on a separate test set, and then average the procedure over the resamplings. They require that the feature selection and classifier development process be completely objective and specified as an algorithm. These methods are often applied improperly, however, with authors first selecting features using the entire data set and then only cross-validating the classifier specification for the selected set of features. This 'partial crossvalidation' provides very biased[10] estimates of prediction accuracy. If the resampling methods are used properly, however, they can provide more accurate estimates of predictive error than the simpler split-sample method described above. A study of breast cancer also illustrates the point: van't Veer *et al*.[11] predicted clinical outcome of patients with axillary node-negative breast cancer (metastatic disease within 5 years versus disease free at 5 years) from gene expression profiles, first based on preselected genes and then using a fully crossvalidated approach. The improperly crossvalidated and the properly crossvalidated methods resulted in estimated error rates of 27% (12 out of 44) and 41% (18 out of 44), respectively. While van't Veer *et al*.[11] report both estimates of the error rate, the properly crossvalidated estimate was reported only in the supplemental results section on the website and the invalid estimate received more attention.

We call both the split-sample method and the resampling methods 'internal' methods of validation, because they are internal to the study developing the classifier. If the patients in this study do not adequately reflect the variability that might be seen for subsequent patients, then the estimates of prediction accuracy may be similarly limited. Also, if the assay used for reading the genomic signature is performed at a single center in phase II but at a large number of centers in phase III, then there may be some deterioration in predictive accuracy. In general, the phase II experience cannot be expected to mimic the phase III experience in all respects, but the internal measures of predictive accuracy, if obtained in the ways described above, should provide useful information for the design of the phase III trials.

We emphasize here the validation of the predictive accuracy of a genomic classifier. This is quite different than attempting to validate the contribution of the individual components of the classifier. For single gene or single protein classifiers, the two concepts are the same, but this is not the case for multivariate classifiers. In selecting predictive features based on whole genome expression profiling, for example, one cannot expect robustness in the set of genes selected for inclusion in the classifier. This is because the expression of genes is correlated by coregulation and because the statistical power for selection of each informative gene is limited by the stringency of the significance level used in gene selection. Robustness of the individual gene components, however, is not essential. It is robustness of classification accuracy with independent data that matters. Consequently, classifier validation should not consist of re-examination of the significance of the correlation of individual gene components with outcome in independent data, but rather on whether the classifier as a whole predicts accurately for independent data.

Clinical trial designs using genomic signatures in therapeutics development

A phase III clinical trial traditionally tests whether a specified new treatment is effective in a specified class of patients relative to a control treatment. Patient classifiers based on genomic signatures serve to restrict and focus on the set of patients in which treatment evaluations are made. The hypothesis testing character of phase III trials should not change as a result of the availability of new genomic technologies, however. In oncology, for example, most therapeutics now developed are molecularly targeted to the products of disregulated genes. An increasing body of evidence, however, indicates that some epithelial tumors of common primary sites are heterogeneous mixtures of tumors with different genomic pathogeneses. Therapeutics for such heterogeneous diseases may only be effective for a small subset of the cases.[12,13] Phase III clinical trials are often most efficiently conducted in the subset of patients who are considered most likely to benefit from the therapeutic.[14] Conducting the phase III trial in the conventional manner in a broad group of patients may produce a false negative result and may expose many patients, who are unlikely to benefit, to the potential side effects of the treatment. The target set of patients in whom the treatment is tested can sometimes be identified prior to starting the phase III clinical trial so that the trial can be designed to get a reliable answer to a clearly defined hypothesis. A classifier distinguishing those patients most likely to respond to the new therapeutic from those less likely may be developed using genomic profiling of responders and nonresponders in phase II development.

The paradigm of using expression profiles of diseased tissue to guide clinical development is not appropriate for many diseases outside of oncology. Profiles of genetic polymorphisms in candidate genes or genome wide can potentially serve a similar function, however. Polymorphisms may reflect the metabolic processing, pharmacokinetics and pharmacodynamics of the drug in its relationship to normal tissue. Such polymorphisms can potentially be used to identify patients at risk for serious adverse reactions to the drug. It is possible, however, that germline polymorphisms may reflect genetic differences in predisposition to different variants of a symptomatically defined disease. As drug efficacy may vary dramatically among disease variants, genotyping of germline DNA could potentially yield useful classifiers of patients likely to benefit from the drug. The likelihood of developing a classifier for efficacy based on germline polymorphisms is less promising than that based on expression profiling of tumor tissue in oncology.

As in the case of gene expression profiling of tumors, it is often not *a priori* evident which polymorphisms are informative for predicting serious adverse events (SAEs) in nononcologic diseases. Consequently, large numbers of polymorphisms must be screened in order to develop a classifier of those patients at increased risk. Whereas oncology studies may utilize patients treated in phase II studies to develop classifiers of efficacy by comparing responders to nonresponders, this is much less feasible for developing SAE classifiers. If many SAEs are observed during phase II development of nononcology drugs, it is not likely that the drug will be further developed. It is quite possible, however, that SAEs will be observed in phase III trials that were not seen above baseline levels during phase II development. Generally, the number of patients treated during phase II development is too small to use SAE as an end point for identifying patients at higher risk, and incompletely validated biochemical surrogate end points will have to be used to develop risk classifiers during phase II. There are, unfortunately, many examples of drugs that fail during phase III trials because of safety problems. If archived lymphocytes are available for such patients, then one phase III trial with a substantial number of SAEs can be used to develop the SAE classifier of the patients at risk, and the classifier can be tested on patients from other phase III trials of the drug.

In order to use a genomic signature classifier effectively in a phase III clinical trial, the signature must take the form of a completely specified classifier that can be objectively and prospectively used to select and stratify patients for the

clinical trial. It is not sufficient to have identified a set of marker genes or polymorphisms to be investigated in the phase III trial, because that does not provide a setting for testing a focused hypothesis about effectiveness of the new treatment for a prespecified target set of patients. Instead, the marker genes or polymorphisms should be combined into a completely specified classifier which can be used to select or stratify patients.

Even in oncology, development of a classifier and assay for identifying the patients most likely to respond to the drug before conducting phase III trials will often be difficult and require a larger phase II database with more extensive biological characterization of patient specimens. Even where the mechanism of action of the drug is thought to be known, there may be no obvious assay for identifying the cases where the target is important in the pathogenesis of the disease. For example, in the development of gefitinib (Iressa) for non-small-cell lung cancer (NSCLC), it was only after the conduct of large negative phase III trials that it was discovered that the drug efficacy appeared enhanced among patients whose tumors contained mutations in the phosphyrolation loop of the epidermoid growth factor receptor (EGFR) gene.[12,13] In spite of the fact that EGFR was the molecular target of gefitinib, during phase II development, there was no attempt to look for EGFR mutations. How to best identify the NSCLC patients likely to respond to small molecule EGFR inhibitors remains unclear as a phase III trial of a similar drug, erlotinib (Tarceva), was positive in unselected patients, but retrospective analysis indicated that the survival effectiveness of the drug was better predicted based on expression of the EGFR protein or amplification of the EGFR gene than by the presence of an EGFR mutation.[15]

Prospective designs

Screening enrichment designs. In an enrichment design, the genomic classifier is used to select patients and to study only the selected patients. An enrichment study design may be attractive when prior hypothesis strongly supports a genomically targeted patient population for concentrated therapeutic efficacy. Roses[16] gave a genomic efficacy enrichment example used for proof of concept in the drug development program of a GSK molecule to treat obesity. Roses indicated that a three SNP genomic signature gave GSK confidence to plan a phase IIB trial to study the treatment effect on weight loss in an enriched subset.

Using a genomic classifier to exclude patients for study, however, generally presupposes a substantial level of confidence in the classifier. This may exist, for example, for an antibody like Herceptin with a known therapeutic target. In that case, it may not be attractive, or even ethically justifiable to include patients whose tumors do not express the target. With classifiers developed empirically to discriminate responders from nonresponders in early phase clinical trials, however, there will often not be sufficient confidence to use the classifier as an exclusion criterion. The enrichment requires an available assay that is reasonably sensitive and specific. Sensitivity is the probability of identifying as classifier positive patients who are responsive to the new treatment and specificity is the probability of identifying as classifier negative those who are not responsive.

Simon and Maitournam[14] showed that the enrichment design can substantially reduce the number of patients needed, but that the efficiency depends on the operating characteristics of the assay and on the prevalence of patients who are preferentially responsive to the new treatment. For the development of drugs such as cetuximab (Erbitux) for colon carcinoma or head and neck carcinomas where the target molecule is expressed in the majority of the cases, selecting patients based on expression of the target or a genomic signature may not be necessary. For example, cetuximab appears to prolong survival substantially in unselected head and neck cancer patients.[17] The situation for cetuximab and colon cancer is less clear. The drug appears to cause tumor shrinkage in only a small proportion of patients, but randomized trials to evaluate its impact on survival have not been reported.

Genomic signature stratified designs. In many cases there will not be sufficient confidence in a genomic signature classifier to use it for excluding patients, but there will be interest in using it to stratify patients for phase III trials. Although it is possible that the objective is simply to compare the treatment to the control group overall, and stratification is performed in order to ensure balance with regard to the genomic signature, which is thought to be prognostic, it is unlikely that genomic classifiers will be developed for this limited objective. We will consider two alternatives.

The objective of the trial may be to evaluate the new treatment separately in the two subgroups determined by the genomic classifier. Hence, the study must be sized to have adequate statistical power separately in each of the two strata. This design is called the separate test marker by treatment interaction design by Sargent *et al*.[18] Since the stratified trial described here could have been conducted formally as two separate clinical trials, testing each hypothesis at the conventional 5% level seems justified. The total number of patients required will often be large because essentially two separate clinical trials are being conducted.[18] Owing to the large sample size required, this design has not been widely used. For example, in the comparison of trastuzmab (Herceptin) plus chemotherapy to chemotherapy alone in naïve and refractory metastatic breast cancer patients,[19,20] cases with less than a 2+ level of expression of the Her2/neu protein were excluded. The trial was sized for overall analysis and showed statistically significant benefits for Herceptin with regard to several end points. *Post hoc* analyses of the subsets of patients with 2+ and 3+ levels of HER2/neu expression were performed and suggested that the benefit of Herceptin was restricted to patients with 3+ levels of expression. The drug was approved for patients with 2+ and 3+ levels of expression.

A second alternative is to size the trial for testing the treatment overall for all patients, but to include a contingency for a single preplanned subset analysis in those

patients predicted to be responsive by the genomic classifier in case the overall analysis is negative. In this case, the two hypotheses should be tested at reduced significance levels in order to ensure that the chance of a false positive finding in the trial is limited to 5%. With this design, the statistical power for the subset analysis may be inadequate if the trial is sized for the overall analysis and if the proportion of patients in the responsive subset is very small. In this case, the planned sample size can be increased, but the sample size should be specified in the protocol and not based on subsequent findings. Since the treatment effect is expected to be greater in the subset predicted responsive by the classifier, a relatively small sample size for that subset may be adequate, however. This design strategy has not, to our knowledge, been previously proposed, and consequently there are no examples of its use that can be cited. It provides sponsors with an incentive to develop genomic classifiers without undue risk of limiting their labeling indications if results indicate broad effectiveness. If the trial is sized for evaluating overall treatment effectiveness at a somewhat reduced significance level, the number of patients required should be substantially less than for the separate analysis strategy described above.

The key issue is not the act of stratification for balancing the randomization process, but rather the clear specification of the hypotheses to be tested in the protocol of the study before patients are accrued. These objectives can be pursued even if stratified randomization is not performed, possibly because the assay for reading the genomic signature is not ready at the start of the trial. What is essential, however, is that the trial be sized properly and the classifier and treatment efficacy hypotheses to be tested be clearly defined in advance. If the phase III trial data are used to refine the genomic classifier, then an additional phase III trial will generally be required to test the treatment efficacy hypotheses in the patient subset(s) determined by the classifier.

Sargent *et al*.[18] and Pusztai and Hess[7] describe other designs for evaluating the clinical utility of a predictive marker. That objective is important for determining whether a new diagnostic is useful in conjunction with an existing widely available treatment.

Retrospective designs for efficacy. In some cases, it will not be clear at the start of human testing whether or not a genomic signature will be useful in development of the drug. If in the phase I/II trials the drug shows broad activity for the target patient population, then development of a genomic signature for efficacy may not be warranted. In some cases, phase III studies will have been conducted without the development of genomic signatures. If such studies are negative, then there may be interest in attempting to (i) salvage the drug by identifying a genomic signature for efficacy based on data from the phase III trials or (ii) understand the mechanism of the action of the drug in a genetically heterogeneous patient population for future clinical development of the drugs.

Attempting to salvage a drug that has failed in conventional phase III trials is problematic and must be approached carefully. This approach is only possible if suitable biological samples are available from patients in the phase III trials. For signatures based on genotypes, this would only require archived blood samples. The genomic signatures developed in a failed phase III trial would have to be evaluated in separate phase III trials. Often this will involve the design of one or more new phase III trials based on the completely specified genomic signature. In some cases, however, if multiple large negative phase III trials were conducted with archived specimens, one of the trials could be used to develop a genomic signature, and the other previously conducted phase III trials could be used to evaluate the completely specified genomic signature.

Retrospective designs for safety. Previously conducted phase III trials can also be used to develop genomic signatures for safety rather than efficacy. For many diseases, there are not enough adverse events observed during the pre-phase III clinical tests in order to develop a genomic signature for safety. If there are numerous adverse events observed during the phase I/II period, in many cases, the drug will not be further developed. The phase III trials may demonstrate significant efficacy, but an incidence of serious adverse events could be troubling for the treatment indication. In such cases, it may be useful to attempt to develop a genomic signature of which patients experience the adverse events based on the phase III data. Such a genomic signature would have to be validated in data from other phase III trials.

When a serious adverse event is observed within the treated group, the observational case–control design can be considered when the adverse event is rare.[21] The design is common in epidemiologic research and is retrospective in nature. In such designs, the cases are those who manifest a treatment-related adverse event and the controls (non-cases) are those who also receive drug treatment but do not develop adverse events. The hypothesis to be tested is that a predefined genomic classifier distinguishes those treated patients who experience the adverse event from those who do not. The genomic classifier should be developed using data external to the case–control study employed for the validation. Case–control sampling can also be used for the previous study that developed the genomic classifier. Case–control sampling assumes that archived pretreatment specimens are available, but it avoids assaying all specimens for treated patients. It is desirable to assay pretreatment specimens for all treated patients who experience the adverse event and for matched treated patients who do not experience the event. Some suggest matching k controls to each case for investigation of a rare event. One should consider matching based on ethnicity or on use of multiple unlinked markers (known as genomic control) to minimize potential bias from population stratification.[22] Such matching is critical in multinational clinical trials involving multiethnic patients.

Unlike the traditional case–control study in which exposure status relies on the subject's recall, the pertinent retrospective data in pharmacogenomic trials are the genomic signature presumably obtained from archived

samples. The classical recall bias is generally not an issue in a retrospective treated case–control study. If the samples are available on only a subset of patients, however, then there may be concern about whether that subset is representative of the entire cohort.

The case–control design can be effective for testing whether a genomic signature is associated with an adverse event, but it is less effective for evaluating the positive and negative predictive accuracy of the signature classifier. A cohort design in which specimens for all treated patients are assayed is more effective for that purpose. The validation study must also establish that excluding treatment patients who are at high risk of adverse events does not eliminate the effectiveness of the treatment. If the case–control study is used to develop the genomic signature, rather than to test a completely specified signature, another clinical trial is generally required to validate its clinical significance.

Following completion of the abacavir clinical trial in patients with HIV-1 infection, Hetherington *et al*.[23] reported that serious hypersensitivity reaction (HSR) to abacavir was associated with HLA-B*5701 polymorphisms. The association was discovered using a matched case–control design. Mallal *et al*.[24] reported finding the same HSR-HLA association as well as an association with the linked B-DR-DQ locus. Mallal utilized a case–control design for both abacavir-treated and non-abacavir-treated patients. Hughes *et al*.[25] confirmed the HLA-B*5701 association with HSR to abacavir in white males, white females, and Hispanics. However, no significant association was found in blacks in this retrospective matched case–control design. The results indicate the importance of matching on ethnicity.

In retrospective safety evaluations of a placebo-controlled or untreated-controlled clinical trial, a case–control study of treated patients is a sensible approach if the adverse event is sufficiently rare in the placebo group. If the adverse event is not that rare, however, then an analysis of treated patients alone will not be satisfactory because it may identify patients at risk for the adverse event but not due to the treatment. In this case, the genomic signature to be evaluated should be assayed in both treated and untreated patients, and the hypothesis to be tested is that the signature identifies the patients for whom the treatment-related component of risk of the adverse event is increased.

Conclusion

Many research efforts are taking place to explore possible causes of a low success rate of phase III clinical trials. One likely cause is disease heterogeneity and limitation of treatment effectiveness to an unidentified subset of patients. This results in the overall treatment effect in many cases being too small to be detected with the sample sizes employed. It is also the case that many treatments that are approved for medical use are effective only for a limited proportion of the patients to whom they are applied. This results in serious safety issues for patients and economic issues for our healthcare system. It is imperative that we develop effective tools for delivering the right medications to the right patients.

Genomic and bioinformatic technologies have made it increasingly feasible to develop signature classifiers for identifying patients likely to respond to a given therapeutic. The rate-limiting step in delivering effective pharmacogenomic tools to treating physicians, however, is likely to be the design and analysis of proper clinical validation studies. The research literature on prognostic markers that never made it to the bedside is voluminous. It is relatively easy to perform and publish retrospective analyses of prognostic markers using specimens from a heterogeneous group of patients. Unfortunately, such studies are frequently unreliable or not clinically relevant because of the lack of focus and structure of the analysis. The situation is compounded when using genomic technologies that provide tens of thousands of candidate predictors. Consequently, we have tried to focus on the design of clinical validation studies for genomic classifiers completely specified in developmental studies. We have described a variety of study designs for utilizing pharmacogenomic signature classifiers in the development of new drugs and for the identification of patients who experience adverse events from marketed drugs. We have described designs based on prospective accrual of new patients in phase III trials and designs based on use of archived specimens from previously conducted clinical trials. All the designs we propose, however, are characterized by prospective specification of clear objectives, hypotheses based on completely specified classifiers, and explicit analysis plans. We hope that these recommendations facilitate the conduct of efficient and reliable clinical trials and the delivery of safe and effective therapeutics to the right patients.

Acknowledgments

We thank the reviewers for numerous important suggestions for improving the manuscript. This research work was supported by the CDER RSR #03-12 funds awarded to Dr Sue-Jane Wang by the Center for Drug Evaluation and Research, US Food and Drug Administration.

Duality of interest

None declared.

References

1 Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *J Am Med Assoc* 1998; **279**: 1200–1205.
2 Evans BJ, Flockhart DA, Meslin EM. Creating incentives for genomic research to improve targeting of therapies. *Nat Med* 2004; **10**(12): 1289–1291.
3 Group BDW. Commentary: biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Therapeut* 2001; **69**(3): 89–95.
4 Tsai CA, Wang SJ, Chen DT, Chen JJ, Tsai CA. Sample size for gene expression microarray experiments. *Bioinformatics* 2004; **21**: 1502–1508.

5 Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. *Br J Cancer* 2003; **89**: 1599–1604.
6 Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. *Design and Analysis of DNA Microarray Investigations*. Springer: New York, 2003.
7 Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. *Ann Oncol* 2004; **15**: 1731–1737.
8 Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for classification of tumors using gene expression data. *J Am Stat Assoc* 2002; **97**: 77–87.
9 Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002; **346**: 1937–1947.
10 Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data: class prediction methods. *J Natl Cancer Instit* 2003; **95**: 14–18.
11 van'tVeer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**: 530–536.
12 Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Branigan BW. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 2004; **350**: 2129–2139.
13 Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 2004; **304**: 1497–1500.
14 Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res* 2004; **10**: 6759–6763.
15 Tsao MS, Sakurada A, Cutz JC *et al.* Erlotinib in lung cancer – molecular and clinical predictors of outcome. *N Engl J Med* 2005; **353**(2): 133–144.
16 Roses AD. Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nat Rev Genet* 2004; **5**: 645–656.
17 Bonner J, Giralt J, Harari PM, Cohen C, Jones C, Sur RK *et al.* Phase III study of high dose radiation with or without cetuximab in the treatment of locoregionally advanced squamous cell cancer of the head and neck. *Proc Am Soc Clin Oncol* 2004; **22**: 5507.
18 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.
19 Baselga J. Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: pivotal trials. *Oncology* 2001; **61**(Suppl 2): 14–21.
20 Eiermann W. Trastuzumab combined with chemotherapy for the treatment of HER2-positive metastatic breast cancer: pivotal trial data. *Ann Oncol* 2001; **12**(Suppl 1): S57–S62.
21 Weiss ST, Silverman EK, Palmer LJ. Case–control association studies in pharmacogenetics. *Pharmacogenomics J* 2001; **1**: 157–158.
22 Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–228.
23 Hetherington S, Hughes AR, Mosteller M, Shortino D, Baker KL, Spreen W. Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet* 2002; **359**: 1121–1122.
24 Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse transcriptase inhibitor avacavir. *Lancet* 2002; **359**: 727–732.
25 Hughes AR, Mosteller M, Bansal AT, Davies K, Haneline SA, Lai EH *et al.* Association of genetic variations in HLA-B region with hypersensitivity to abacavir in some but not all populations. *Pharmacogenomics* 2004; **5**(2): 203–211.