Pharmacogenomic Predictor Discovery in Phase II Clinical Trials for Breast Cancer

Lajos Pusztai,[1] Keith Anderson,[2] and Kenneth R. Hess[2]

Abstract

Purpose: We examined if supervised analysis of gene expression data from phase II studies could identify HER-2 overexpression as a predictor of response to trastuzumab.

Experimental Design: Gene expression data from 132 newly diagnosed breast cancers were used to simulate 50,000 single-agent phase II trastuzumab studies. True HER-2 amplification was assessed by fluorescence *in situ* hybridization.

Results: Only 3.67% of the simulated studies yielded HER-2 as the top predictor, >96% of the individual "studies" picked a different gene as the most predictive of trastuzumab response. HER-2 was included in the top 10 gene list 9.73% of the time. When HER-2 was a priori defined as a potential predictor, 99.6% of the simulated studies confirmed overexpression among responders. Candidate marker testing may be more efficient than *de novo* predictor discovery in phase II trials. We describe a tandem, two-step phase II trial design for rapid marker assessment that combines two optimal two-stage phase II trials into a single study. In the first stage, unselected patients are treated, and if insufficient responses are seen, the trial remains open for marker-positive patients only and a second two-stage trial commences.

Conclusions: The probability of successful discovery of drug-specific pharmacogenomic response markers in a typical phase II study is small. The evaluation of predefined predictors using tandem two-step phase II design has the advantages of estimating response rates in both unselected and marker-selected patient populations and allows for simultaneous screening of multiple different predictors for the same drug and several distinct predictor-drug pairs in a single, parallel multiarm trial.

The advent of high-throughput gene expression profiling raised the hope that one can discover powerful new predictors of response to therapy by examining thousands of genes in pretreatment tumor specimens. Several studies reported gene signatures that were predictive of response to chemotherapy in internal cross-validation and some also showed promising predictive values in small independent validation cohorts (1–4). Therefore, it has become increasingly common to employ gene expression profiling as a predictive marker discovery tool in phase II clinical trials. The rationale is that a semiquantitative and unbiased look at thousands of transcripts in cancers will reveal a strong association between the expression of at least some genes and response to therapy. However, there are equally compelling caveats why this unbiased screening approach to predictor discovery may not yield reliable predictors under many circumstances. The multiple comparison problem inherent to microarray analysis is well known (5). It stems from the large number of variables (i.e., genes) that are compared between two usually small (relative to the number of variables) data sets. This could lead to a large number of very small *P* values from these comparisons, most of which are due to chance (6). It is possible to adjust for this phenomenon by calculating false discovery rates for particular *P* values. Another confounder relates to the coordinated expression of thousands of genes. Individual transcripts do not represent independent variables, but rather, their expression is highly correlated with one another. Many phenotypic characteristics of cancer are associated with, and likely caused by, the coordinated expression of thousands of genes. For example, estrogen receptor (ER)–positive breast cancers differ from ER-negative cancers in the expression of thousands of genes (7, 8). The gene expression pattern of high-grade tumors is also different from low-grade breast cancers (9). These structured, large-scale gene expression differences associated with strong phenotypic features can have profound influence on the predicative marker discovery process.

ER-negative, high-grade breast cancers are more sensitive to many different types of chemotherapies compared with ER-positive and low-grade tumors. Comparison of transcriptional profiles of tumors that responded to chemotherapy with those that did not could reveal many differentially expressed genes. However, most of these genes will reflect the gene expression

Authors' Affiliations: Departments of [1]Breast Medical Oncology and [2]Biostatistics and Applied Mathematics, the University of Texas M. D. Anderson Cancer Center, Houston, Texas

Received 4/6/07; revised 6/13/07; accepted 7/9/07.

Grant support: NCI grant RO1-CA106290 (L. Pusztai), the Breast Cancer Research Foundation, and the Goodwin Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Lajos Pusztai, Department of Breast Medical Oncology, The University of Texas M. D. Anderson Cancer Center, Unit 1354, P.O. Box 301439, Houston, TX 77230-1439. Phone: 713-792-2817; Fax: 713-794-4385; E-mail: lpusztai@mdanderson.org.

doi:10.1158/1078-0432.CCR-07-0809

differences that underlie the phenotypic differences between the two response groups. The resulting pharmacogenomic response predictor may represent, to a large extent, a predictor of phenotype (i.e., high-grade, ER-negative cancers versus low-grade ER-positive cancers). Several predictive gene signatures have been reported, but currently, it remains unknown to what extent these signatures include genes that are predictive to a particular drug as opposed to being dominated by phenotype-associated genes that are predictive of general chemotherapy sensitivity. It is possible to adjust for or stratify cases by known clinical variables during the marker selection process but the limited sample size often makes this difficult to do.

The probability that the supervised pharmacogenomic discovery approach, when responders are compared with nonresponders, could lead to regimen-specific predictors depends on (*a*) to what extent the response groups are balanced for strong phenotypic markers, (*b*) and the extent of molecular differences that determine drug-specific response. If drug sensitivity is influenced by the 2- to 3-fold higher or lower expression of a few dozen genes, these differences might not be readily discovered through supervised pharmacogenomic analysis of data from a typical phase II trial including 30 to 60 patients. These modest gene expression differences between responders and nonresponders can easily be masked by the larger scale molecular differences due to any phenotypic imbalance between the response groups. The technical noise of microarray experiments can also obscure small-scale gene expression differences. A typical gene chip includes 15,000 to 25,000 probe sets, and even with a high level of reproducibility, hundreds of genes could show several fold differences in expression values in simple replicate experiments (i.e., the same RNA profiled twice). For example, when 24,000 measurements are done (e.g., Affymetrix U133A gene chip) and the overall concordance is 97.98% in a technical replicate; 1.31% of all measurements could have ≥2-fold variation. This means that 314 genes could be ≥2-fold decreased or increased from one experiment to another due to technical noise alone (10).

In this article, we examined if we could have discovered HER-2 mRNA overexpression as single gene predictor of response to trastuzumab through supervised analysis of pharmacogenomic data from simulated phase II trials using real breast cancer gene expression data. Our results suggest that the probability of successful drug-specific pharmacogenomic response marker discovery from a typical phase II study can be small. We suggest that prospective testing of a priori defined candidate markers may be more efficient than *de novo* predictor discovery in phase II trials. We describe a tandem, two-step phase II clinical trial design for rapid assessment of candidate response predictors. The design combines two classic optimal two-stage phase II trials into a single study and can estimate response rates in both unselected and marker-selected patient populations. It also allows for simultaneous screening of multiple different predictors for the same drug and several distinct predictor-drug pairs in a single, parallel multiarm trial.

Patients and Methods

We used real gene expression data obtained with Affymetrix U133A gene chips from 132 newly diagnosed breast cancers and real HER-2 amplification results determined by fluorescence *in situ* hybridization analysis to simulate single-agent phase II trastuzumab studies. The patient population and data processing were described in detail previously and all clinical and microarray data is available at the M. D. Anderson Bioinformatics web site[3] (4). None of these patients received trastuzumab therapy; therefore, their response to this drug cannot be directly ascertained. No gene expression data is publicly available from patients who received single-agent trastuzumab. However, it is known from the literature that ~30% to 35% of HER-2–amplified cases respond to single-agent trastuzumab therapy and HER-2–nonamplified breast cancers do not respond to this treatment (11, 12). Currently, there are two clinically routine methods to select patients for trastuzumab therapy; these include detection of HER-2 protein overexpression with immunohistochemistry and HER-2 gene amplification assessed by *in situ* hybridization methods. We randomly selected 45 HER-2 normal (non–gene-amplified) and 15 HER-2 gene–amplified cases based on real fluorescence *in situ* hybridization results to simulate a 60-patient phase II study population. Five (33%) of the 15 HER-2–amplified cases were randomly assigned "responder" category because true response was unavailable. The remaining 10 HER-2–amplified cases, together with the 45 HER-2 normal cases were considered "nonresponders", this corresponds to an overall response rate of 8.3% (5 of 60) for the whole study population and 33% for the HER-2–amplified population. The gene expression profiles of the two groups were compared using unequal variance *t* test to identify differentially expressed genes. This is one of the most commonly used approaches in the literature to identify informative genes for predictive marker discovery. We did this analysis 50,000 times, randomly picking different sets of cases from the larger patient pool of 132, and randomly assigning 1/3 of the HER-2–amplified cases to the "responder" category. The goal was to examine how often HER-2 was ranked by its *P* value as the most differentially expressed gene in these 50,000 iterations. Each iteration could be considered as a single 60-patient clinical trial and the analysis follows the commonly used supervised analysis to discover molecular predictors of response from pharmacogenomic data. A more complex clinical trial modeling process could have been designed that allows for variable response rate in each simulated study and where the response rates follow normal distribution with a mean at 35%. However, this more realistic modeling would have made it even less likely for HER-2 to be identified as a predictor of response in more than a small percentage of individual studies. Our fixed response rate biases the simulations towards higher power to detect HER-2 overexpression as a marker of response.

We also examined a complementary scenario. Preclinical data is usually available to propose potential predictors; the amount of the drug target itself or measures of its functional activity are among the two most obvious candidates. Therefore, we tested the hypothesis that increased HER-2 mRNA expression is a marker of trastuzumab response based on results from preclinical studies. We did the same unequal variance two-sample *t* test on the transcriptional profile data to test the hypothesis that HER-2 mRNA expression is higher in responders than in nonresponders. The probe closest to the 3′ end (216836_s_at) was selected to represent HER-2 mRNA expression (13). Because in this analysis, we test a single hypothesis and a single gene, we considered $P < 0.05$ to be statistically significant. We plotted how often HER-2 expression was significantly higher in responders in the 50,000 iterations.

Results

HER-2 gene overexpression as predictor of response to trastuzumab may not have been discovered through supervised analysis of pharmacogenomic data from a single phase II clinical study. There is a strong correlation between HER-2 gene amplification and increased HER-2 mRNA level (Fig. 1; ref. 13). We examined, in a series of simulated phase II studies, if we could have discovered HER-2 mRNA overexpression as a

[3] http://bioinformatics.mdanderson.org/pubdata.html

predictor of trastuzumab response. The results are presented in Table 1. HER-2 was indeed ranked most often as the number one differentially expressed gene in these simulations. However, in only 3.67% of all iterations was HER-2 ranked as the top gene, in 96% of the individual "studies", a different gene was picked as the most predictive of trastuzumab response. HER-2 was included in the top 10 gene list in 9.73% of the simulations, and in the top 50 list 20% of the time. Other genes frequently listed among the top 100 differentially expressed genes included GRB7, which is frequently coamplified with HER-2, as well as ER and several ER-associated transcripts. This is not surprising because HER-2 amplification is more common among ER-negative cancers and illustrates the confounding effect of large-scale phenotype-related gene expression differences. In the current data set, 36% of ER-negative cancers were HER-2–amplified compared with 18% of the ER-positives (χ^2, $P = 0.022$). These results suggest that it is rather unlikely that HER-2 could have been discovered as a single gene predictor of trastuzumab response from supervised pharmacogenomic analysis of a single phase II clinical trial. In fact, there is an ~80% chance that it would not have been included in the top 50 gene list of any individual study.

HER-2 mRNA overexpression as an* a priori *defined candidate predictor of response. One could examine HER-2 mRNA expression as an a priori defined potential predictor of response to trastuzumab based on the known mechanism of action of this drug. We plotted how often HER-2 expression was significantly higher in responders compared with nonresponders in the 50,000 iterations. Figure 2 shows histograms of observed P values and indicates that 99.6% were $P < 0.05$. Therefore, any single study among the 50,000 simulations could have easily identified HER-2 overexpression as a predictor of response with a very high probability when this single hypothesis (i.e., single gene) was tested.

The reason why supervised analysis of gene expression data from a single study is likely to miss HER-2 as the correct predictor is because there are many genes with lower P values than HER-2 in any randomly selected (or accrued) data set. Most of these genes are not directly related to trastuzumab response but represent random chance and/or phenotypic confounders; nevertheless, they are ranked higher on the differentially expressed gene list ordered by P values.

Tandem, two-step phase II trial design for predictive marker evaluation. These observations have important consequences for the design of clinical trials. The results show that the use of

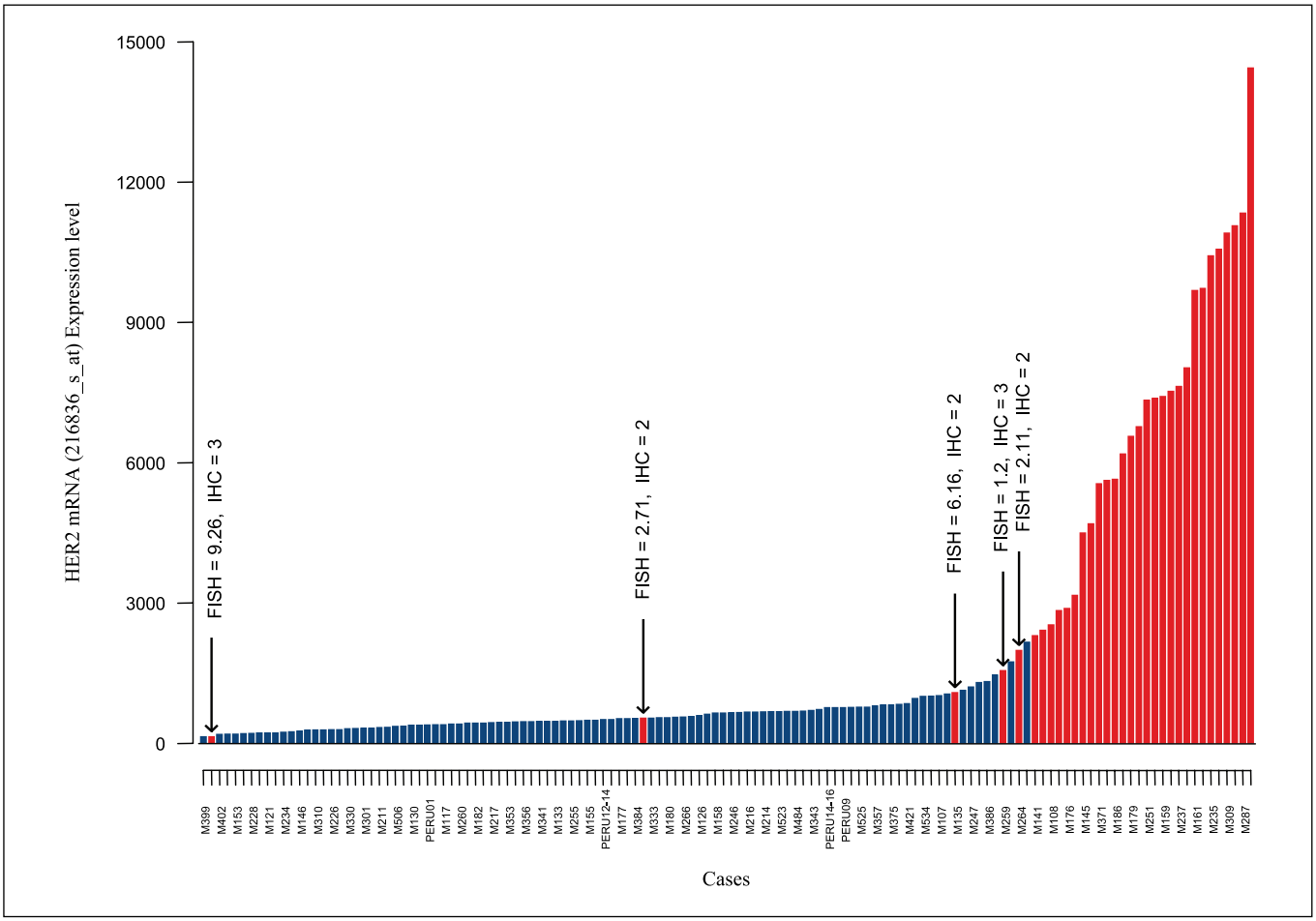

Fig. 1. HER-2 mRNA expression and clinical HER-2 status. Rank-ordered HER-2 mRNA expression values for the 132 cases are presented. Clinically HER-2–positive cases (*red, n* = 33) were defined as fluorescence *in situ* hybridization amplification or 3+ immunohistochemistry results. Individual fluorescence *in situ* hybridization and immunohistochemistry results are shown for the five outlier cases with low HER-2 mRNA expression but clinically positive HER-2 status.

Table 1. Top 49 probe sets sorted by frequency of how often they are ranked as no. 1 in 50,000 iterations

Probe set ID	Ranking	Top 1	Top 10%	Top 50%	Top 100%	Top 500%	Top 1,000%	Gene symbol
216836_s_at	1	3.67%	9.73%	20.06%	31.23%	59.96%	69.08%	*ERBB2*
205440_s_at	2	1.20%	9.75%	20.21%	24.74%	43.51%	53.57%	*NPY1R*
221811_at	3	0.81%	4.53%	13.45%	20.72%	42.55%	56.41%	*PERLD1*
203569_s_at	4	0.58%	2.89%	7.77%	11.76%	22.76%	26.32%	*OFD1*
212070_at	5	0.58%	2.20%	3.20%	4.36%	11.94%	16.61%	*GPR56*
200029_at	6	0.57%	1.65%	5.14%	8.21%	16.67%	23.76%	*RPL19*
210828_s_at	7	0.53%	3.67%	11.32%	15.53%	24.72%	28.47%	*ARNT*
55616_at	8	0.47%	3.04%	11.17%	18.34%	40.70%	54.78%	*PERLD1*
205590_at	9	0.47%	2.67%	7.42%	10.92%	27.39%	42.29%	*RASGRP1*
212196_at	10	0.46%	3.21%	5.76%	7.66%	18.91%	26.87%	*IL6ST*
216835_s_at	11	0.45%	3.22%	7.84%	11.35%	21.79%	28.06%	*DOK1*
218094_s_at	12	0.40%	1.19%	3.03%	4.96%	13.92%	21.62%	*C20orf35*
203685_at	13	0.39%	2.19%	5.13%	6.53%	12.27%	21.11%	*BCL2*
201011_at	14	0.37%	1.62%	3.80%	5.35%	13.60%	19.57%	*RPN1*
213712_at	15	0.36%	2.50%	7.40%	11.30%	24.76%	33.20%	*ELOVL2*
203439_s_at	16	0.34%	2.25%	9.79%	18.73%	52.53%	63.92%	*STC2*
219359_at	17	0.34%	3.94%	7.43%	8.14%	11.51%	14.99%	*FLJ22635*
205354_at	18	0.29%	0.95%	2.67%	4.15%	11.68%	19.46%	*GAMT*
218153_at	19	0.28%	0.93%	1.44%	1.82%	4.66%	7.92%	*FLJ12118*
201255_x_at	20	0.27%	1.29%	2.98%	4.23%	10.20%	14.67%	*BAT3*
205225_at	21	0.27%	1.19%	1.72%	2.35%	5.06%	9.18%	*ESR1*
204699_s_at	22	0.27%	0.62%	1.21%	1.59%	3.83%	6.86%	*C1orf107*
209140_x_at	23	0.26%	1.20%	2.51%	3.34%	7.24%	10.99%	*HLA-B*
208715_at	24	0.24%	0.77%	1.58%	2.35%	7.55%	14.75%	*TMCO1*
208962_s_at	25	0.23%	0.81%	1.96%	2.97%	7.47%	11.53%	*FADS1*
203929_s_at	26	0.23%	3.59%	14.67%	22.21%	42.84%	51.38%	*MAPT*
200070_at	27	0.23%	0.40%	0.65%	1.04%	3.42%	6.30%	*C2orf24*
207332_s_at	28	0.22%	1.31%	3.27%	4.88%	11.40%	17.25%	*TFRC*
36994_at	29	0.22%	1.37%	3.95%	6.10%	15.37%	22.04%	*ATP6V0C*
202411_at	30	0.22%	1.03%	2.27%	3.50%	11.67%	18.72%	*IFI27*
218276_s_at	31	0.22%	1.16%	2.87%	4.28%	11.51%	18.13%	*SAV1*
210761_s_at	32	0.21%	0.99%	3.64%	6.65%	31.35%	50.79%	*GRB7*
220414_at	33	0.20%	0.69%	1.10%	1.39%	3.03%	4.46%	*CALML5*
200837_at	34	0.19%	0.50%	0.82%	1.35%	5.08%	8.91%	*BCAP31*
219872_at	35	0.19%	2.91%	8.83%	13.03%	30.64%	40.52%	*DKFZp434L142*
209016_s_at	36	0.19%	1.63%	5.75%	9.19%	21.98%	25.92%	*KRT7*
210930_s_at	37	0.18%	0.40%	0.96%	1.92%	11.01%	20.81%	*ERBB2*
220038_at	38	0.18%	1.73%	5.43%	8.72%	24.36%	36.14%	*SGK3*
204029_at	39	0.18%	1.82%	5.95%	8.95%	21.28%	29.60%	*CELSR2*
203627_at	40	0.17%	1.31%	3.39%	5.37%	16.33%	24.20%	*IGF1R*
219197_s_at	41	0.17%	2.00%	6.37%	9.78%	20.93%	26.66%	*SCUBE2*
221727_at	42	0.17%	1.10%	3.42%	6.01%	20.81%	30.72%	*SUB1*
205186_at	43	0.17%	0.79%	2.03%	3.09%	9.41%	15.47%	*DNALI1*
219155_at	44	0.17%	1.30%	3.15%	4.39%	10.17%	16.23%	*PITPNC1*
209007_s_at	45	0.16%	0.64%	1.17%	1.49%	3.38%	6.32%	*C1orf63*
204686_at	46	0.16%	0.71%	1.63%	2.72%	10.66%	19.20%	*IRS1*
209485_s_at	47	0.16%	0.54%	1.85%	3.01%	9.11%	15.53%	*OSBPL1A*
211176_s_at	48	0.16%	2.36%	8.27%	11.89%	24.40%	30.64%	*PAX4*
216580_at	49	0.16%	1.03%	2.72%	4.06%	10.16%	15.70%	*RPL7*

NOTE: There are two probe sets that target the HER-2 transcript. Probe set "216836_s_at" is ranked no. 1 most often, in 3.67% of all iterations. However, the second probe set "210930_s_at" makes it to the top only 0.18% of the time (ranked 37th). This is most likely because it targets the HER-2 transcripts further 5′ and its median overall expression intensity is only 80.2 compared with 670.6 for probe 216836_s_at.

broad pharmacogenomic screening to identify molecular predictors for drugs that show low overall response rate (8-10%) is a high-risk strategy for marker discovery in a typical phase II study. Using this approach, we could not identify the only currently known single gene predictor of response to trastuzumab, HER-2 overexpression. Because there is no known gene signature that predicts response to this drug, we could not directly test whether a multigene signature could have been detected by our simulation studies. However, because we could not identify the most informative single gene, it is rather unlikely that a statistically robust multigene signature could have been identified.

Our results also show that it is a more productive strategy to prospectively test an a priori defined predictor in pharmacogenomic data obtained during a phase II study. Enough is known about the mechanism of action of most drugs that one could rationally propose at least one or more potential response predictors. These could include the expression levels of a single gene, complex gene signatures, or any other molecular measurement (14–16). One promising strategy is to identify predictors in cell line models (17). How to best define the response predictor based on the preclinical data or using results from archived specimens will vary from case-to-case and is not

the subject of this article. The predictor could be a single gene or a complex gene signature and could be measured at the mRNA or protein levels. However, the predictor must be fully defined, including cutoff values for positivity and negativity prior to evaluating its value in a prospective clinical trial. Conceptually, testing a rationally designed response predictor in a prospective clinical trial is no different from testing a candidate drug in a therapeutic study.

The two-stage phase II trial design has been used for several decades to identify drugs with promising clinical activity and quickly discard those with low activity. The goal of the phase II clinical trials is to determine whether a new drug has enough clinical activity to warrant larger scale evaluation. During the first stage of a classic two-stage phase II study, "n_1" number of patients are entered into the trial and if fewer than "r_1" number of responses are observed, the accrual terminates for lack of activity. Otherwise, accrual continues to a total of "n" evaluable patients. At the end of this second stage, the drug is recommended for further evaluation if the final response rate is "$\geq r$" (18). In order to calculate sample size, investigators must first specify a drug activity level of interest and probability variables for early stopping. The design can be easily modified to include interim efficacy monitoring using a Bayesian approach (19).

Similar phase II trial designs could be applied to prospectively evaluate putative response markers. Assume that a drug has completed phase I evaluation and a dose was selected for phase II testing, and also at least one, but preferably more, putative predictive markers are available but the response rate in unselected patients is still unknown. A tandem, two-stage, phase II clinical trial design could be applied to test the drug and the predictors simultaneously. The goal of the study is to determine if the drug is likely to have a certain level of activity in unselected patients, and if it is below the level of interest, can a particular patient selection method enrich the responding population to meet the targeted level of activity in the molecularly selected group.

The concept of this design is illustrated in Fig. 3. The study starts out as a two-stage phase II trial for unselected patients

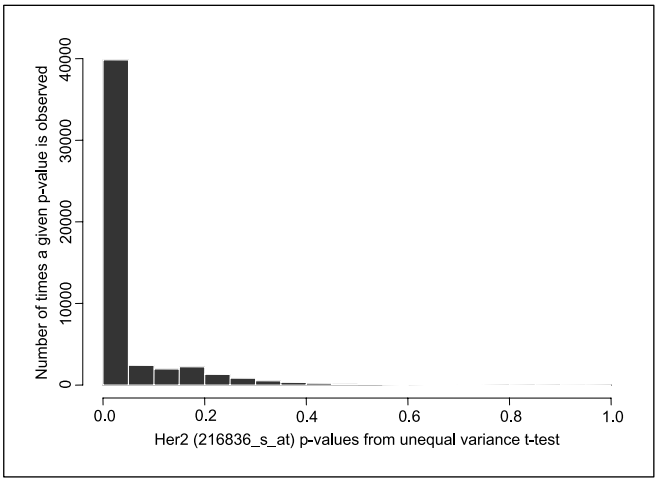

Fig. 2. Histogram of *P* values of HER-2 mRNA differential expression in data from 50,000 simulated phase II trials. HER-2 mRNA levels measured by probe set 216836_s_at were compared between five randomly designated responders in the HER-2 – amplified cases and the remaining 55 cases including 10 HER-2 – amplified "nonresponders".

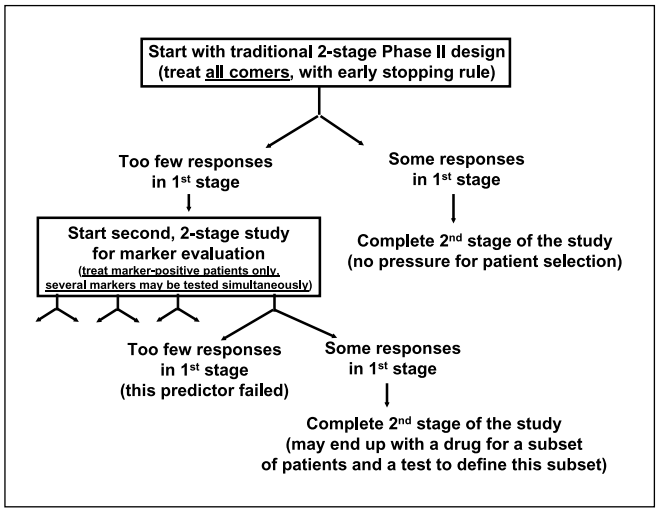

Fig. 3. Schema of the tandem two-step phase II predictor marker evaluation trial design.

with an early stopping rule for futility. If sufficient numbers of events (e.g., clinical benefit rate or response rate) are observed during the first stage, the study proceeds to the second stage to establish the benefit rate more precisely in unselected patients. However, if an insufficient number of events are seen during the initial stage, instead of stopping accrual for futility, the trial remains open for response marker – positive patients only and a second optimal two-stage trial commences. This second stage is introduced because it is very unlikely that the small group of patients who participated in the first phase (typically $n_1 < 25$) included a sufficient number of marker-positive cases to draw a conclusion about the activity of the drug in this molecularly defined subset. If an insufficient number of events occur after accruing "n_2" number of marker-positive cases during the second step of the study, the trial is discontinued following the early sopping rules and the marker is rejected. Otherwise, the study proceeds to complete accrual of additional marker-positive patients in order to estimate the benefit rate more precisely.

Sample size calculations for the tandem two-step design follow the same rules as for a classic two-stage or Bayesian phase II design (18, 19). For example, let's assume that the targeted level of activity is 25% clinical benefit (CB) rate and we feel comfortable stopping the trial early if it becomes apparent during the interim efficacy monitoring that there is <7.5% chance that this level of activity will be achieved. Using a noninformative prior distribution of β (1, 1) for benefit rate, the stopping boundaries are listed in Table 2.

The same stopping boundaries will apply to the first and second steps of the study. The probability of early termination with the above design is 80% if the true CB rate is 10%, and it is 7.5% if the CB rate is 30%. Maximum sample size, if the study does not terminate early, can be set by defining the minimum CB rate of interest and the acceptable posterior credible intervals. The estimated β distribution for minimum CB = 25% with 90% credible intervals is shown in Table 3.

For example, if the true CB rate is 25% with a maximum sample size of $N = 50$, the observed CB rates would fall between 17% and 36%, 90% of the time. These numbers can be adjusted

Table 2. Stopping boundaries for noninformative prior distribution of β (1, 1)

Number of evaluated patients (*N*)	Recommend stopping if CB is ≤
9	0
15	1
20	2

to suit the particular needs of an investigator. What benefit rate is considered acceptable in the marker-positive population, which corresponds to the positive predictive value of the test, can vary from one clinical situation to the other.

If there are multiple, nonoverlapping candidate response markers for the same drug, all of these could be tested simultaneously in a single, parallel, multiarm study. Accrual to each marker arm could occur simultaneously but results are analyzed separately for each arm. If the predictors capture overlapping patient populations (i.e., the same patient is positive for several of the markers), more complex adoptive randomization designs could be applied to preferentially randomize patients into better performing marker arms. It is expected that only a small fraction of the total patient population will be positive for any one of the markers because the sensitive population is small (assuming that the marker is reasonably sensitive and specific). Therefore, this design implies that a large number of patients will be screened during the second step of the study to find marker-positive individuals who are eligible for therapy. The exact number needed to screen will depend on marker prevalence. To maximize eligibility for treatment among the screened patients, it is desirable to test multiple different predictors simultaneously on each case. Patients who are positive for a particular marker will receive treatment in parallel treatment arms. Several different drug and marker pairs could also be evaluated simultaneously in a single study.

Discussion

Gene expression profiling during phase II clinical trials is increasingly done in an attempt to discover predictive markers. These studies might not yield reliable drug-specific markers in many instances due to the confounding effect of coordinated expression of thousands of genes that are associated with clinical phenotype (which might also predict response to therapy) and because of inherent technical noise in the microarray data. In small studies with few responses, these factors could have a profound effect on the results. However, the transcriptional profile results obtained during phase II studies can provide an opportunity to prospectively test a priori defined gene expression–based predictors. Predictive markers can be tested in the clinic in a similar manner as drugs are tested and similar trial design principles can be applied. When the activity of a new drug is unknown in unselected patients, but there is also at least one potential predictor of response, we suggest that a tandem two-step phase 2 trial design could answer both questions: (*a*) does the drug have sufficient activity in unselected patients, and if it does not, (*b*) can the predictor enrich the responder population.

Technically, the two steps of the study could be separated and run as independent studies, one for unselected patients and another for marker-positive patients. However, there are several reasons why keeping them together might yield a more seamless trial. Running two separate studies takes more time and is more expensive. Because any particular predictor will define only a relatively small subset of patients who are eligible for therapy, it is appealing to simultaneously test several predictors so that fewer patients are turned away as marker negative and therefore ineligible for treatment. Multiple distinct predictors for the same drug as well as several different drug and predictor pairs could be tested simultaneously in a single, parallel, multiarm trial.

The study design that we propose evaluates the candidate response marker based on its positive predictive value (e.g., how often response is seen among marker-positive cases). A high enough positive predictive value is necessary for clinical utility, but alone, it is not sufficient to make a marker clinically useful. Sensitivity also has to be considered and that it is influenced by false negative cases (i.e., patients who respond but are marker-negative). A crude estimation of sensitivity may be made by using information from the first step of the tandem design when unselected patients are included, but to define the sensitivity and specificity of the marker more precisely, a separate study is needed. However, response markers with low positive predictive values are clinically not useful and need not be investigated further.

Any clinical study that prospectively evaluates patient selection methods will require a tissue biopsy. Early single-agent phase II studies are often conducted in the metastatic patient population, and therefore, deep tissue and organ biopsies might be required for marker evaluation. Invasive procedures to obtain biopsies for correlative science studies have traditionally been avoided for considerations of patient discomfort, fear of adverse events, and cost. However, serious complication rates from fine-needle aspirations of abdominal organs or body cavity lymph nodes are substantially lower than serious, grade 3 to 4 treatment-related adverse events during most investigational chemotherapy trials. For example, a study described complication rates encountered during 10,766 ultrasonographically guided abdominal fine-needle aspirations and reported 0.18% (n = 22) major complications including peritoneal bleeding and 0.018% death rate (n = 2; ref. 20). Well-informed patients may elect to take these risks to participate in studies that evaluate the value of personalizing treatment.

There are other clinical trial strategies to prospectively discover and validate molecular predictors of response. An adoptive clinical trial design was recently described that incorporates pharmacogenomic predictor discovery and validation into a

Table 3. Estimated β distribution for minimum CB = 25% with 90% credible intervals

N	Lower bound of CB rate	Upper bound of CB rate
30	0.15	0.40
40	0.16	0.38
50	0.17	0.36
120	0.19	0.32

traditional randomized phase III study (21). This design includes (*a*) identification of the sensitive patients during the initial phase of the phase III trial (i.e., predictor discovery), (*b*) a statistical test for treatment effect for the marker-positive patients that are accrued in the remainder of the trial (i.e., validation set), and (*c*) a properly powered statistical test for overall treatment effect using results from all randomized patients (i.e., traditional phase III end point). This is an appealing strategy because it incorporates marker discovery and validation into a single randomized study without compromising the ability to detect an overall treatment effect by traditional criteria. Unfortunately, it requires a large randomized trial. The current article describes an alternative strategy that can be incorporated into phase II testing of novel drugs.

References

1. Ayers M, Symmans WF, Stec J, et al. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel/FAC chemotherapy in breast cancer. J Clin Oncol 2004;22:2284–93.
2. Chang JC, Wooten EC, Tsimelzon A, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet 2003;362:362–9.
3. Iwao-Koizumi K, Matoba R, Ueno N, et al. Prediction of docetaxel response in human breast cancer by gene expression profiling. J Clin Oncol 2005;23:422–31.
4. Hess KR, Anderson K, Symmans W, et al. Pharmacogenomic predictor of sensitivity to preoperative paclitaxel and 5-fluorouracil, doxorubicin, cyclophosphamide chemotherapy in breast cancer. J Clin Oncol 2006;24:4236–44.
5. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005;23:7332–41.
6. Pounds S, Morris SW. Estimating the occurrence of false positive and false negatives in microarray studies by approximating and partitioning the empirical distribution of *P*-values. Bioinformatics 2003;19:1236–42.
7. Pusztai L, Ayers M, Stec J, et al. Gene expression profiles obtained from single passage fine needle aspirations (FNA) of breast cancer reliably identify prognostic/predictive markers such as estrogen (ER) and HER-2 receptor status and reveal large scale molecular differences between ER-negative and ER-positive tumors. Clinical Cancer Res 2003;9:2406–15.
8. Gruvberger S, Ringner M, Chen Y, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res 2001;61:5979–84.
9. Sotiriou C, Wirapati P, Loi S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 2006;98:262–72.
10. Anderson K, Hess KR, Kapoor M, et al. Reproducibility of gene expression signature based predictions in replicate experiments. Clin Cancer Res 2006;12:1721–7.
11. 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;17:2639–48.
12. Vogel CL, Cobleigh MA, Tripathy D, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 2002;20:719–26.
13. Gong Y, Yan K, Lin F, et al. Determination of oestrogen-receptor status and ERBb2 status of breast carcinoma: a gene-expression profiling study. Lancet Oncol 2007;8:203–11.
14. Desai KV, Xiao N, Wang W, et al. Initiating oncogenic event determines gene-expression patterns of human breast cancer models. Proc Natl Acad Sci U S A 2002;99:6967–72.
15. Sweet-Cordero A, Mukherjee S, Subramanian A, et al. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet 2005;37:48–55.
16. Bild A, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
17. Huang F, Reeves K, Han X, et al. Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: rationale for patient selection. Cancer Res 2007;67:2226–38.
18. Simon R. Clinical trials in cancer. In: DeVita VT, Hellman S, Rosenberg SA, editors. Cancer, principals and practice of oncology. Philadelphia: Lippincott Williams Wilkins; 2001. p. 521–38.
19. Thall PF, Simon R. A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. Control Clin Trials 1994;15:463–81.
20. Fornari F, Civardi G, Cavanna L, et al. The Cooperative Italian Study Group. Complications of ultrasonically guided fine-needle abdominal biopsy. Results of a multicenter Italian study and review of the literature. Scand J Gastroenterol 1989;24:949–55.
21. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 2005;11:7872–8.