Valuation of the Use of Biomarkers Predictive of Drug Efficacy to Enrich Responders in Oncology Drug Clinical Development

by

David Wine

B.A., Psychology
B.S. Engr., Computer Science and Engineering
University of Pennsylvania, 1985

SUBMITTED TO
The Sloan School of Management and
The Harvard-MIT Division of Health Sciences and Technology

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREES OF
Master of Science in Management and
Master of Science in Health Sciences and Technology
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY, June 2006

Signature of Author: _____
May 26, 2006

Certified by: _____
A. Gregory Sorensen, MD
Associate Professor of Radiology, HMS, MGH, HST
Thesis Advisor

Certified by: _____
Teodoro Forcht Dagi, MD, MBA
Senior Lecturer
Harvard-MIT Division of Health Sciences and Technology
Thesis Advisor

Certified by: _____
Dalia Cohen, PhD
Global Head of Functional Genomics
Novartis
Thesis Reader

Accepted by: _____
Sloan School of Management

Accepted by: _____
the Harvard-MIT Division of Health Sciences and Technology

Valuation of the Use of Biomarkers Predictive of Drug Efficacy to Enrich Responders in Oncology Drug Clinical Development

by

David Wine

Submitted to the Sloan School of Management and to
the Harvard-MIT Division of Health Sciences and Technology

on February 21, 2006

in Partial Fulfillment of the Requirements for the Degrees of

Master of Science in Management and
Master of Science in Health Sciences and Technology

Abstract

I study several aspects of the value in performing oncology clinical trials using screening biomarkers to preferentially select and enroll responders. From trial reports and investigational reports on potential biomarkers, I construct a series of six cases comparing the trial as conducted to a hypothetical trial using different screening and eligibility criteria. These cases illustrate, within limits of the model, what difference the use of a plausible biomarker test may have on trial size, cost, number of patients screened, and number of patients exposed to experimental treatment without benefit.

Thesis supervisor: A. Gregory Sorensen, MD
Associate Professor of Radiology

Thesis supervisor: Teodoro Forcht Dagi, MD, MBA
Senior Lecturer

Acknowledgements & Dedication

I am indebted to my thesis advisors Greg Sorensen at Harvard Medical School, and Teo Dagi with MIT, who helped especially in the focus of this thesis. And I owe thanks to Dalia Cohen, Head of Global Functional Genomics at Novartis, who graciously agreed to be an outside reader. I have enjoyed and valued our conversations on cancer drugs and development programs. Our discussions formed the basis for my treatment of ethics concerns in this thesis.

I owe special acknowledgements to Isaac Manke, Ben Eckert, Tim Gasperoni, Reuben Cummings, and Fabio Thiers. Together we first addressed this subject in a term paper entitled "Biomarkers in Clinical Development of Cancer Drugs: Challenges and Opportunities" for the HST class Principles and Practice of Drug Development.

Thank you to Professor Anthony Sinskey, MIT Professor of Microbiology and Health Sciences and Technology, the Faculty Director for the drug development course mentioned above. Professor Sinskey was a great help as my academic advisor in the Biomedical Enterprise Program. And thank you to Professor Ernie Berndt, who represents such a strength of the Sloan School, HST, and the Biomedical Enterprise Program at Harvard and MIT.

I thank donors to the Biomedical Enterprise Program, and I thank the Hugh Hampton Young Memorial Fellowship committee, who supported my work financially.

Valuation of the use of biomarkers predictive of drug efficacy to enrich responders in oncology drug clinical development.

This thesis aims to quantify part of the value derived from utilizing a biomarker in the clinical testing of an oncology drug. In particular value obtained from use of a biomarker predictive of drug-efficacy or -resistance to enrich for better clinical effect in the population studied. The value, in human and financial terms, can come from several sources. Those benefits with which I am concerned here are

- to enroll fewer patients, and so reduce the size of the trial,
- to expose fewer patients to experimental treatment without benefit,
- to reduce the duration of the trial.

Reference trials compared with hypothetical trials using biomarkers

To quantify these benefits I analyze a series of reference cases. Each of these references is a clinical trial of an oncology drug for which there is or may be a biomarker predictive of efficacy or resistance. For each reference I posit a certain change to the trial—either adding a biomarker when none was used, or using a different biomarker, or taking one away when one was used. Typically I create the hypothetical case from trial results and a report of an investigation into the predictive value of a potential biomarker. An example of such an investigation is an evaluation of PTEN deficiency as a marker of resistance to trastuzumab [1].

The point is that whichever of the reference or hypothetical case selects patients to get ***greater effect***, that case will be able to show its effect at the same confidence levels while enrolling a ***smaller number of patients***.

Cases will be detailed below; briefly the drugs and clinical applications are

- trastuzumab, metastatic breast cancer

- tamoxifen, metastatic breast cancer
- erlotinib, non-small cell lung cancer

Biomarkers in these cases

By biomarkers I mean any test at all which can provide new information about a patient's likelihood to respond to treatment. Practically proteomic tests and expression tests are avenues of development. The tests used in the cases reported here are tests for expression, immunohistochemistry (IHC), fluorescence in situ hybridization (FISH), and a measure of cell proliferation. Other tests such as tests of methylation, reported race, and gender all fall within the scope of biomarker for this discussion.

Oncology

A word about the interest of predictive biomarker tests in cancer and cancer drug development:

Cancers carry genetic changes, often many, some of which have proven to be characteristic of prognosis or sometimes instructive in treatment. An example is BCR/Abl, a fusion gene, and tyrosine kinase protein characteristic, even defining of chronic myeloid leukemia (CML).

As for imatinib, lately developed to inhibit BCR/Abl in CML, tamoxifen for blockade of the estrogen receptor (ER) has been used in the US since 1977 for breast cancer [2].

Understanding that the etiology of cancer is genetic, we may expect that some cancers are different from others in ways that can guide treatment, as a result of the particular genetic changes that have been dealt to the cells of the malignancy.

patient selection for established treatments

Patient selection for established treatments is clear and well defined when the use of a drug inhibiting a single target proves to be successful, and when

that target is overexpressed or constitutively active. This is so with Her2 as target of trastuzumab and biomarker for predicted efficacy of treatment in breast cancer. And is so with BCR/Abl as target of imatinib and biomarker in CML. And with the estrogen receptor as target of tamoxifen and biomarker in breast cancer.

patient selection for investigational drugs

But patient selection is not always clear for new prospective drugs. Gefitinib had very low response rates, and in odd populations. It was granted accelerated approved for non-small cell lung cancer after phase II results but has had its labeling amended and its use cut back after failing to show survival benefit. Gefitinib inhibits by intention the epidermal growth factor receptor (EGFR), also called variously erbB1 or Her1, which is part of a homologous family of signal transduction proteins involved in growth, and found disregulated in cancers.

A correlation of EGFR expression and efficacy of gefitinib was investigated, but did not prove to be, although particular mutations related with sensitivity to gefitinib have been reported.

EGFR as a target of cancer treatment

As an emblematic example of a cancer therapy target, EGFR has monoclonal antibodies targeting its cell surface epitopes and small molecule tyrosine kinase inhibitors targeting its intracellular region. Like Her2 EGFR is a cell surface protein. As of 2005 as reported in [3] these antibodies (biologics) and drugs were in development or in marketing for Her-1 (EGFR) and Her-2 (Table 1, Table 2):

Agent	Characteristic	Target	Tumor Type	Stage
Cetuximab	Chimeric	HER-1	Colon H&N, NSCLC, pancreas	Marketed Phase III
ABX-EGF	Human	HER-1	Colon, renal	Phase III
EMD-7200	Humanized	HER-1	H&N, ovarian, colon, cervix	Phase II
h-R3	Humanized	HER-1	H&N	Phase II
Pertuzumab	Humanized	HER-2	Breast, ovarian, prostate, NSCLC	Phase II
Trastuzumab	Humanized	HER-2	Breast	Marketed

Table 1. Her2 family monoclonal antibodies in development or marketing. Table from [3]. *H&N, head and neck; NSCLC, non-small-cell lung cancer.*

Tyrosine Kinase Inhibitors Designed to Target the HER Family				
Agent	Irreversible	Target	Tumor Type	Stage
Gefitinib	No	HER-1	NSCLC	Marketed
Erlotinib	No	HER-1	NSCLC, pancreas	Marketed
Lapatinib	No	HER-1/2	Breast	Phase III
CI-1033	Yes	Pan HER	SCC, skin	Phase II
EKB-569	Yes	HER-1	Colon	Phase II
BMS-599626	No	HER-1/2	—	Phase I
AEE788	No	HER-1/2 Anti-VEGFR	—	Phase I

Table 2. Her2 family small molecule tyrosine kinase inhibitors. Table from [3]. *SCC, squamous cell carcinoma; NSCLC, non-small-cell lung cancer.*

value to select patients for greater effect

In treatment, it is of course critical to the lives of patients to know what treatments are most likely to bring benefit. But before a drug has become approved for marketing, it becomes important to know which patients are most likely to benefit from its use. This is so for the same human terms as in treatment with marketed drugs, and for the principle of not wishing to expose patients to experimental treatment without promise of benefit. But in clinical development—with its major impact on drug development expense—effective patient selection promises savings in cost and time.

These are the reasons for this investigation into the value of patient selection to improve response in trials, and why in oncology.

Approach

The rough method of analysis is to collect data on the effect of the treatment in a reference case, with or without biomarker as the case may be. I propose different trial eligibility so the treated subgroups in the reference case and the hypothetical case will have different effects from treatment (effect usually as response rate). To keep the same p value and power of between the studies, the sample size will change. It is the ratio of these two sample sizes which is a primary result in the case.

The choice to utilize a biomarker to select patients implies costs and other considerations; these are many, variable, and bear discussion. A few are costs to develop the biomarker test, for its validation, deployment and use, and considerations toward marketing and adoption of the drug. And some other potential benefits may also attend the choice, such as earlier marketing, competitive advantage of higher efficacy rates, useful patent life, revenue from the biomarker test itself, and potentially improved chances of regulatory approval.*

My main intent in this analysis is to quantify just several of the putative benefits—how the decision to use biomarkers may influence trial size and by extension its cost, trial duration, and patients treated without benefit.

Point of view

It's important to note that I am only making this investigation on drugs which did ultimately gain regulatory approval. Before making choices of clinical trial design, one would not know whether treatment for the indication in question would be ultimately approvable. The value of using a biomarker to select patients for enrollment will depend on which of the two types of therapy it turns out to be, approvable, or not. For example, the value obtained by the extension of useful patent term would not apply to both approvable and non-approvable therapies. Savings in the cash costs of the

* By construction each of the reference and hypothetical case has the same chances of showing effect. This is not the same as achieving regulatory approval, which one can argue is more likely in case the effect of treatment is larger.

trial would apply to both. I place this aspect of valuation—prospectively considering drugs which will ultimately fail—outside my scope, primarily because data are sparse on drug development failures.

I draw on Vernon and Hughen's working paper [4] on the economics of pharmacogenomics, on Simon and Maitournam's work on the efficiency of targeted trial designs [5] and on Manke *et al.* [6] in which this general analysis method was earlier used.

The cases

Trastuzumab trials without Her2 as a biomarker

Two main cases follow in the use of trastuzumab, in treatment of metastatic breast cancer. The reference case is from a phase III study [7] enrolling 938 metastatic breast cancer patients, with eligibility open to those testing for Her2 with an immunohistochemical (IHC) score of 2 or more (IHC2+).

I study several hypothetical cases in comparison to this reference. The first and second suppose that testing for Her2 was not to have been required for enrollment.

Trastuzumab trials with Her2 and PTEN

The third case on the same reference trial supposes now the use of a biomarker predictive of resistance, to be used in addition to Her2 testing. This biomarker is modeled with data from a report on the use of PTEN deficiency as a marker of resistance to trastuzumab [1]. A fourth variation is studied of an incremental improvement to the PTEN test, using PTEN plus a Her2 test with better specificity.

Tamoxifen, in the treatment of metastatic breast cancer.

This case applies a report of a potential biomarker to the prospect of selecting invasive breast cancer patients least likely to relapse under treatment with tamoxifen. The potential biomarker is progesterone receptor (PgR) status measured with ___ and tumor proliferative status measured with ^{3}H-thymidine labeling index (TLI)

Erlotinib, in the treatment of non-small cell lung cancer

Erlotinib is a small molecule EGFR inhibitor, has been approved for the treatment of non-small cell lung cancer. The erlotinib trial showed EGFR expression to have predictive value. These data are used to construct a hypothetical trial in which patients are enrolled who have overexpression of EGFR.

Drug	Brand Name	Drug Type	Target	Cancer Type
erlotinib	Tarceva	tyrosine kinase inhibitor (TKI)	EGFR (Her1) (erbB1)	non-small cell lung cancer
trastuzumab	Herceptin	mAb	Her2 (neu) (erbB2)	invasive and metastatic breast cancer
tamoxifen	Nolvadex	competitive anti-estrogen	estrogen receptor (ER)	invasive and metastatic breast cancer

Table 3. Summary table of drugs, their drug targets, and cancers used in the cases studied.

Ethical Stance

In considering the decision to use predictive biomarkers, each choice entails ramifications: commercial, clinical, ethical. I discuss the ethical nature of the decision, exploring several factors which may become significant when one considers a particular drug development case.

I'll define the decision under ethical consideration as this:

> To decide whether to use a predictive biomarker test in an oncology clinical trial, with a rationale to select subjects for an expectation of greater effect.

In cases where a feasible choice exists to use a predictive biomarker to enrich responders, is it ethical to do so? Is it ethical not to do so? One's answer will depend on the particular situation. Here I bring out several considerations which might bear on the decision, with the intention that these considerations would serve as a framework with which to examine a particular situation.

The decision bears not only on clinical trials, but clinical practice post-marketing should the drug be approved.

patients exposed to trial

One kind of ethical value holds in exposing fewer patients to trials.

One measure of the risk patients engage might simply be the *number of patients in the experimental treatment group*, the fewer the better if effectiveness can be shown.

This number of people are exposed to experimental treatment, and so take the risk the drug will not be effective for them. Any selection criteria that increase effect will be preferable on this score.

Another measure, which one can make once the trial has been conducted, or which one can guess at before, is the *number of individuals exposed to experimental treatment without benefit*. This is the value quantified in the

cases presented here, which mitigate toward the choice of using a biomarker for selection.

Society and individuals only accept the risk of lack of efficacy and adverse effects for the chance the treatment will come into practice. And when the treatment does not come into practice, if we could have known, we would rather not have done the trial.

risk of drug failure

It is preferable to avoid both

- false rejection of an effective drug, and
- false acceptance of an ineffective drug.

In the case of false rejection, there are two ills—patients are exposed to trials uselessly, and a good drug will not later be available to treat patients. In the case of false acceptance, patients are exposed to trials of an ineffective drug, and patients later on will be treated with an ineffective drug.

How can the use of predictive biomarkers affect the likelihood of false rejection and false acceptance? In the analysis presented here, the probability of type I and type II errors are fixed so the biomarker and non-biomarker cases were comparable. This does not mean that in practice the likelihood of approval will be the same when a trial shows efficacy.

There can be several influences which may make a trial with fewer patients and greater effect more likely to achieve approval. For one, trial sponsors may in practice power the biomarker trial higher, in a trade-off between cost and risk of false rejection. Also, given that statistical significance has been reached, one may suspect that approval is more likely if effect is greater, because of the greater potential benefit to patients of the drug. These factors mitigate toward using a biomarker test for patient selection.

A trial less likely to be falsely rejected is to be preferred in this from an ethical viewpoint.

creation of orphan populations

Selection of drug candidates must take into account the size of the ultimate market for those drugs. When considering a collection of targeted treatments, in aggregate this effect could create a set of less served cancer patients. As the groups of unserved patients become small niches, it follows that they will generate less interest by industry.

Whereas regulatory agencies promote the development of drugs for orphan indications, it seems unlikely that the orphan niches which might come into being on the fringes will be seen as orphan indications. Incentives provided by regulatory authorities to promote development of orphan drugs would not apply to these populations.

It is not necessarily so that such an outcome, creating unserved niches, is unethical or inappropriate. It may be considered that these unserved niches exist either way, whether they are recognized or not.

For a drug that has an association between effect and biomarker status, that association will exist whether or not the fact of it is known, and whether or not drug prescribing is based on such biomarker status. So consider the outcome in each of the clinical trial kinds—with patients selected on the basis of biomarker and not. Suppose a biomarker selected trial yields a drug approval, and creates an "orphan population" of patients for whom the drug is not demonstrated to work. In the case of a drug approved after a non-selected trial one may argue that the "orphan population" is still the same population, only now we don't recognize them as less likely responders.

The full force of this argument is weakened when we consider that we will know more about the likely non-responders if we have tested the drug on the full population, and this may provide better guidance to the treatment of biomarker positive individuals. Even given some patients' negative biomarker status they may still be best off taking the drug, only we may not know this to be so if a biomarker were used as a selection criterion in trials.

lower response may not mean the drug is not clinically significant

While trials may be made more efficient by selecting patients more likely to respond, it might happen that even those biomarker negative individuals who are less likely to respond would still find that their best choice is treatment with the drug. When this is the case and when targeted trials are conducted there is risk of harm to the biomarker negative population.

When trials exclude biomarker negative patients one can expect the drug to be approved only for biomarker positive patients. This will in turn affect health provider policies and physicians' prescribing practice.

One may hope that following initial approval for a narrow indication, further trials might be done, to widen approval for other indications. But they may not, and I view further trials as if they may not be made, from our vantage point for ethical analysis at the point of decision on how to design a single clinical trial. Or physicians might prescribe the drug off-label, if they have evidence it is the best choice for some biomarker negative patients.

The key difference, perhaps, if a targeted trial is conducted, is that physicians and regulators would have no data from this trial on the response of biomarker negative individuals. If the biomarker negative population has lesser response, yet still clinically significant response, one might never know based on the targeted clinical trial.

Here is a risk of harm to the biomarker negative population, in case they have clinically significant response to the drug. If a pooled trial were to have been conducted, the biomarker negative population would be approved and prescribed for treatment with the new drug, and would benefit. If a targeted trial were to have been conducted, these same individuals would not have an approved treatment, and their physicians would not have evidence of the drug's benefit.

While there is commercial incentive to make trials smaller, there is also significant incentive to keep indications wide. The best thing that can be said

about this risk of harm to excluded patients is that drug companies have a strong interest to reach for wider indications, at the time of initial clinical trials and in follow-on applications. This interest counterbalances the incentive to make trials smaller, and blunts the tendency toward exclusion.

cases where exclusion criteria vary together with race or ethnic group

One ethical concern may arise in the circumstance that one race or ethnic group is highly represented among those excluded from trials, and then later treatment, according to biomarker status. On the one hand it may be unethical to test the drug on biomarker negative individuals who appear less likely to respond. On the other hand we would rather not leave a racial or ethnic group underserved.

use of a response biomarker for to screen for continued treatment

Use of a biomarker may avoid some of the objections against , in the case that the biomarker for selection is in fact a biomarker for response to treatment. It could be possible, for example, to try an experimental treatment on enrollees in the trial, and keep for study only those for whom an imaging study shortly after initiation of treatment shows that the treatment is acting on the tumor. It can be hoped that this sort of screening may be more specific for response, but statistical likelihood aside, such a screen may be seen as more fair.

where is the biomarker negative patient better off?

One lens on the decision is to consider the individuals who are biomarker negative, indicating they have a lower likelihood of response, and the prospect of including them in a trial which *doesn't* use a biomarker for selection.

If what one believes about the biomarker and its relation to response implies biomarker negative patients are better off with the control group treatment, one may argue that a trial using a biomarker should be considered.

commercial feasibility of biomarker test

Suppose a biomarker test can clearly distinguish patients likely to respond from those unlikely to respond, yet can not be feasibly implemented in clinical practice. In such a case, it would be hard to argue for approval of the drug to be used without the biomarker. Then trials would have to be conducted without using the biomarker for patient selection. However, if trial sponsors or physicians conducting the trial can know who will not respond, can a trial ignoring this information be ethically conducted? The answer may depend on what alternatives are available for the biomarker negative group. The main concern I see when this situation occurs is that drug companies will have incentive to avoid gathering more data showing the relationship of biomarker status to response.

how to handle uncertainty

How should we treat uncertainty in the quantities that form the context for the decision to use a predictive biomarker? How narrow should our estimates be before considering a trial using the biomarker? These estimates are uncertain—biomarker prevalence, drug response conditioned on biomarker status, even knowledge that the drug is at all effective. At the time of clinical drug development, we expect more often to be unsure of these values than to know them well.

One (somewhat dissatisfying) approach to this question is to make the evaluations underlying the decision not only given one's best estimates, but also while varying estimates within reasonable ranges. If the extremes of such evaluations all point toward the same decision, one may feel more comfortable. However, in case the extremes point toward opposite decisions (which I consider more likely), there is little additional guidance to be gained from the exercise.

further trials when biomarkers are used for control and experimental drug

Tricky issues of trial design are raised if standard of care treatment has different inclusion/exclusion criteria than experimental. In essence the

populations in the control group and experimental group are different when different biomarkers are used to establish eligibility for each treatment .

If standard of care were to demand the use of the established biomarker/drug, and the novel drug uses a different biomarker then certain problems arise in the design of the trial. How are subjects to be selected for randomization? Select without the biomarker and include/exclude after randomization? Then the experimental and control populations would be composed differently, and measurement of treatment effect over control would be spurious. Select only those who are positive for the biomarkers needed for both of the treatments? Then the control and experimental groups are composed the same but the population studied would not be representative of the population meant to be treated.

summary

This discussion of ethical concerns mostly lend support to the use of predictive biomarkers for patient selection, yet exposes certain hazards to the approach. It is hoped that analysis of a particular drug development program in the light of the issues discussed here will provide some guidance in this decision.

Modeling sample size

I follow some of Vernon and Hughen's terminology (Table 4) [4] and use Simon and Maitournam's sample size equations to determine the first result of this model, ρ_{BR}, (Table 5) the sample size ratio—the ratio between the size of a trial in which only biomarker positive patients are enrolled and the size of one in which the biomarker test is not performed.

λ	Proportion of the population for whom biomarker test is positive
ε	Efficacy† for those without biomarker, above control.
δ	Improvement in efficacy among positive biomarker population.
$E_B = (\varepsilon+\delta)$	Efficacy rate in the biomarker positive population, above control.
$E_R = (\varepsilon+\lambda\delta)$	Efficacy rate in pooled population of positive and negative biomarker test, above control.
α	Probability of rejecting the null hypothesis when it is true. That is, the chance of concluding that the treatment is effective if it has no effect or if it is deleterious. This is required to be .05 typically.
β	Probability of failing to reject the null hypothesis given it is false. That is, the chance of failing to show efficacy for a treatment that does have the expected effect. Often (1- β) is described as the power of a trial. In practice I take β to be typically 0.1

Table 4. Symbols from [4] for the calculation of sample sizes in a reference trial versus a related hypothetical case.

A statement of the relation between required trial size and effect (and α and β), in the terms of Vernon and Hughen [4], is

$$N = \left(\frac{1}{E} - 1\right)(z_\alpha + z_\beta)^2$$

Equation 1. Relation between sample size, α, β, and effect. [4]

† Efficacy in these definitions refers to the treatment groups, usually as response rate, above control.

Simon and Maitournam use a more sophisticated sample size equation, and perhaps more closely following practice. I adopt this equation, shown fully in Appendix A. [‡]

In my analysis I pick N and E from a trial report. Using the characteristics of some contemplated biomarker, I estimate the effect which would follow in the use of such a biomarker to select patients. The relative sample sizes of the targeted to the untargeted trials is a main result.

$$\rho_{BR} = \frac{N_B}{N_R}$$

N_R is the size and E_R is effect from a reference trial; I propose a value of the effect E_B in a hypothetical trial; and from these obtain the corresponding sample size of a trial using the biomarker, N_B.[§]

I provide by construction that α and β are each the same between actual and hypothetical trials.

ρ_{BR}	The *sample size ratio*, between the sample size required to demonstrate effect in the biomarker positive population and the sample size to demonstrate effect in the pooled biomarker positive and negative population.
N_R	The number of patients required to detect the effect $\varepsilon+\lambda\delta$, in the pooled population of biomarker positive and negative, with the parameters α and β, without using biomarker for eligibility.
N_B	The number of patients required to detect the effect $\varepsilon+\delta$, with the parameters α and β, while using the biomarker for eligibility.

Table 5. Symbols used to model changes in trial size.

[‡] Both sets of sample size equations yield results that are mostly similar but vary by as much as 17%.

[§] The actual trial data form the *reference* case. Characteristics of a contemplated biomarker form the *hypothetical* case. I abuse this language in the first example in which it is reversed—a hypothetical trastuzumab trial considered to have been done without Her2 as an eligibility criterion.

Scope and validation of the model

Such a model as I present, to estimate value that would be gained in a hypothetical trial in comparison to a reference trial, is meant to illustrate. The reference cases and biomarker characteristics are chosen to create a plausible comparison. In any actual particular case of a new trial the equations underlying this model would be applied using values relevant to that case. In practice other considerations outside this model would certainly be part of the analysis.

Validation

In validating a model, one would ideally wish to

a) ensure the model is calculating what it is meant to calculate without error, that it matches a distinct statement of what the model does.

b) compare with any independent implementations of the algorithm.

c) note where the model breaks down, what it does and doesn't take into account; consider assumptions which may be violated and so introduce error.

d) compare with any independent estimates of a similar type.

e) compare the results of the model's predictions with actual outcomes.

Of course there is the final bulwark against a renegade error, the critical reader.

detecting potential errors in model's behavior

The model takes a few inputs from the two cases to be compared, the reference trial and the hypothetical one in which a different selection criterion is used to select patients more likely to respond to treatment. The main intermediate result is the ratio of the two sample sizes. Outputs depend on this and others of the inputs.

The model is described in the preceding section and in following sections. To assure the model is acting reasonably, I reproduced a few synthetic cases developed by Simon and Maitournam [5].

I also ran a series of synthetic data on the author's website, http://linus.nci.nih.gov/~simonr/boep.html. These results matched mine down to the tolerance of integer rounding.

Samples of the synthetic data used in this validation are found in Appendix B.

testing versus an independently developed implementation

I also ran a series of synthetic data on the author's website, http://linus.nci.nih.gov/~simonr/boep.html. These results matched mine down to the tolerance of integer rounding.

Samples of the synthetic data used in this validation are found in Appendix B.

limitations and scope of the model

There are assumptions made in the model which are not always (often?) correct, and these cause systematic error, or bias. Assumptions, their violation, and its effect on the results are discussed along the definition of the model, and in the sections of relevant cases.

Even within the scope of this model, there are some assumptions of the model at variance with full reality; these bear some consideration.

One assumption is that the control group will have the same outcome independent of whether the trial is done with the biomarker. This is untrue generally (but with an effect unknown to us). I note later in the case of Her2 in breast cancer in what way the resulting estimates should be in error because of this assumption.

Another assumption made in [4] is that control group effect is 0. The equations used here, from [5], do not make this assumption.

compare with any independent estimates

The only closely related estimate I know of is on a powerpoint slide [8] by Art Levinson as CEO of Genentech. In this a claim is made on the value obtained by using Her-2 to select patients for trastuzumab—Appendix C. In this appendix I compare this statement with my results.

compare predictions with actual outcomes

One of the very attractive approaches to validating a model is to compare it's predictions with later outcomes. However, it is not desirable, feasible, or ethical to test the same drug for the same indication with targeted and non-targeted trial designs.

Modeling reduction in cost

I use the cost of a single patient in a cancer clinical trial to be $10676 in 2005 dollars, based on an estimate of $9000 in 1998 [9], and inflated by the consumer price index [10].

Vernon and Hughen [4] argue that trial costs form a preponderance of the cost of drug development, so that $(N_R-N_B) / N_R = 1-\rho_{BR}$ (Table 6) approximates the reduction in trial costs and development costs overall, when using the biomarker test to select patients.

$1-\rho_{BR} = (N_R-N_B)/N_R$	Variable reduction in trial costs, of biomarker case versus reference.
$C_V = (N_R-N_B)\times$ <*single patient cost*>	Variable cost savings from smaller trial size.

Table 6. Variable reduction in trial costs.

Modeling the number of patients screened

Any benefits from being able to make earlier development decisions or from earlier marketing depend on the duration of clinical trials. The duration of trials will depend in part on the number of patients treated, but also on the number of patients who are screened, and their availability.

Fewer patients enrolled, assuming the number of clinical trial centers remains unchanged, implies a shorter enrollment time and so a shorter trial time.** However, depending on the proportion of the population for whom the biomarker test is positive (λ) more patients may have to be screened. Whether having to screen more patients leads to a longer trial time depends critically on the rate that new patients become available, and whether more trial centers are opened. This question will be idiosyncratic, depending on the incidence rate of the cancer, the catchment size of the centers, and the particulars of the eligibility requirements. Rather than making weakly based assumptions I model this question as far as identifying the number of

** I neglect the time to make the biomarker screening, but this should be considered in practice.

patients screened, as N_B/λ (Table 7) [4, 5]. The number of patients screened in the reference case is trivially N_R.

N_B/λ	The number of patients screened in the biomarker case.

Table 7. Screening size of trial.

Modeling the difference in patients treated without benefit

When comparing a reference trial with a hypothetical case, I report my estimate of the difference in the number of patients exposed to experimental treatment without benefit. I admit this may be a fuzzy concept; even so this is how I define it here:

- I assume that the control group is getting standard of care (half of those enrolled by assumption), and for this reason I count none of them in this estimate. The numbers treated in each case are $N_R \times 0.5$ and $N_B \times 0.5$.
- I treat effect for a single patient to be binary, so that a treatment effect of 60% in a group is taken to mean 40% of the patients do not benefit.
- Those patients given the experimental treatment without effect in the reference case are $N_R \times 0.5 \times (1-E_R)$. Those in the biomarker case $N_B \times 0.5 \times (1-E_B)$.

$N_{\Delta FailTx} = 0.5 \times ((1-E_R) \times N_R - (1-E_B) \times N_B)$	The expected number of biomarker negative patients treated with the experimental treatment who do not benefit.

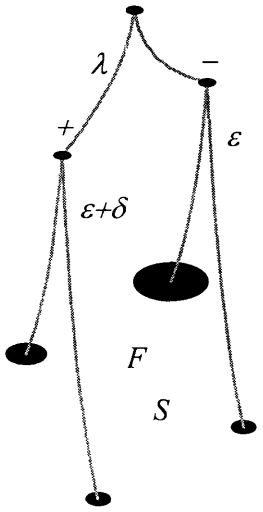
Table 8. The difference in number of patients treated without benefit.

$N_{\Delta FailTx}$ is what I define as the difference in number of patients exposed to experimental treatment without benefit (Table 8). A large number means that the biomarker+ selection spares many individuals from treatment without benefit.

Case modeling format

This diagram and table summarize what the biomarker test means in the context of population of eligible patients and their response to treatment. As a whole it represents the reference trial in which both biomarker + and – groups are treated in their proportion in the population. Effect is net of control group effect.

trastuzumab Her2 IHC+ v. none



λ		0.28	
ε		.10	
ε+δ		.13	
–	F	65%	
+	F	22%	
–	S	7%	
+	S	6%	

biomarker +/–
F fail treatment, S succeed

The inputs and outputs of the model are shown in the case table.

drug name, biomarker condition 1 v. condition 2		
λ		*proportion biomarker+*
ε		*efficacy for biomarker-, net of control*
δ		*Δ efficacy for biomarker+*
$E_B\ (\varepsilon+\delta)$		*efficacy for biomarker+, net of control*
$E_R\ (\varepsilon+\lambda\delta)$		*efficacy for pooled biomarker +/-, net of control*
		treatment effect in control group
		treatment effect in biomarker-
		treatment effect in biomarker+
ρ_{BR}		*sample ratio*
N_B		*sample size biomarker case*
N_R		*sample size no-biomarker case*
N_B/λ		*patients screened to find biomarker+*
$(N_R-N_B)/N_R$		*% reduction in variable cost*
C_V		*variable cost savings*
$N\Delta_{FailTx}$		*patients saved failure under treatment*
$P(+\|succeed)$		*sensitivity of the biomarker test*
$P(-\|fail)$		*specificity of the biomarker test*

Some figures are filled in directly or after some calculation, from clinical trials and biomarker studies. These input values are entered in the left of the column; and values that descend from these, to satisfy the relations of the model, are entered in the right of the column.

N_R or N_B may be taken from the reference case, or N_R is arbitrarily set to 1000 to illustrate the ratio.

A note on what is meant here by "efficacy" or "effect." Effect is represented as a proportion, measured in different ways, in keeping with the clinical trial reports. Data from the cases studied here use variously complete response, partial response, objective response rate, five year relapse-free survival, and overall survival. Response is used synonymously. In this model of clinical trials, the null hypothesis is that the experimental treatment group and

control group effect are not different (two sided). There is usually no placebo, but rather standard of care treatment for the control group. Computations of sample size use effect net of control group effect. ε, N_B, *and* N_R all refer to effect above control. Equation 1 uses effect in the sense of net effect.

Trastuzumab trials without Her2 as a biomarker

I now consider the hypothetical case in which trastuzumab is imagined to have been tested without using the Her2 biomarker.

A phase III trial studied the use of trastuzumab plus chemotherapy versus chemotherapy alone in Her2-overexpressing metastatic breast cancer [11].

treatment condition	N	rate of response
trastuzumab plus anthracycline	143	56%
anthracycline	138	42%
trastuzumab plus paclitaxel	92	41%
paclitaxel	96	17%
trastuzumab plus chemotherapy	235	50%
chemotherapy	234	32%

Table 9. Phase III results for trastuzumab plus chemotherapy versus chemotherapy alone in Her2-overexpressing metastatic breast cancer. Her2-overexpression was determined as IHC 2+. [7, 11]

Pooling together each of the trastuzumab treated groups, compared to no trastuzumab:

treatment condition	N	rate of response
trastuzumab	470	50%
no trastuzumab	468	27%

Table 10. Response rate pooled across treatment subgroups in pivotal phase III trial.

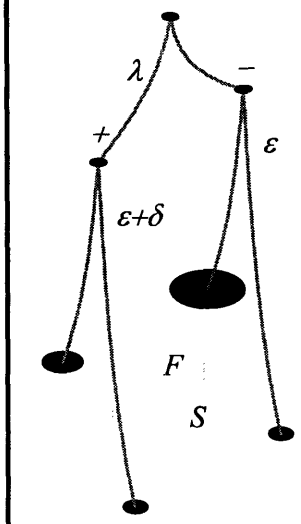
All patients enrolled in these trials had Her2 over-expression of IHC 2+. Since Her2 is reported to be overexpressed in 25-30% of metastatic breast cancer[12], I take λ for 27.5%. We know the effect in the biomarker positive population, from Table 10; this is $(\varepsilon+\delta)$, 50%. The population size N_B, 938 is also from Table 10.

We do not, from this phase III study, know ε, the efficacy of trastuzumab plus standard of care for breast cancers which *don't* express Her2. Paclitaxel monotherapy in other studies gave response rates of 21% to 68% in metastatic breast cancer. [12] This implies ε should be something above

21%, but this is probably an underestimate, if only because of false negative Her2 assays, and because Her2 is prognostic for negative outcome with chemotherapy.

One phase II study examined clinical outcomes associated with the various Her2 assays. This study enrolled a group who were positive and a group of patients who were negative for Her2 expression [12]. The overall response for IHC+ patients was 69%, and for IHC- patients response was 46%. Thus from that study $\varepsilon^* = 46\%$, $(\varepsilon^*+\delta^*) = 69\%$, and so $\delta^* = 13\%$. If we adopt $\delta = \delta^* = 13\%$, directly we get these values:

sample table



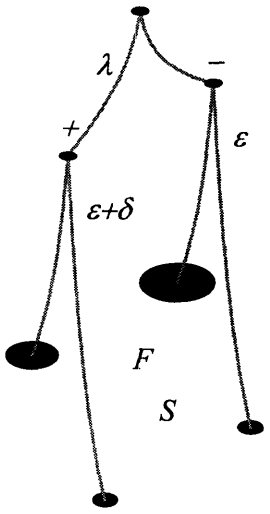
λ		*0.28*	
ε		*.10*	
ε+δ		*.23*	
−	F	*65%*	
+	F	*22%*	
−	S	*7%*	
+	S	*6%*	

biomarker +/−
F fail treatment, S succeed

trastuzumab Her2 IHC+ v. none		
λ	*0.28*	*proportion biomarker+*
ε	*10%*	*efficacy for biomarker-, net of control*
δ	*13%*	*Δ efficacy for biomarker+*
$E_B\ (\varepsilon+\delta)$	*23%*	*efficacy for biomarker+, net of control*
$E_R\ (\varepsilon+\lambda\delta)$	*14%*	*efficacy for pooled biomarker +/-, net of control*
	27%	*treatment effect in control group*
	37%	*treatment effect in biomarker-*
	50%	*treatment effect in biomarker+*
ρ_{BR}	*0.38*	*sample ratio*
N_B	*938*	*sample size biomarker case*
N_R	*2490*	*sample size no-biomarker case*
N_B/λ	*3350*	*patients screened to find biomarker+*
$(N_R-N_B)/N_R$	*62%*	*% reduction in variable cost*
C_V	*$16,566,726*	*variable cost savings*
$N\Delta_{FailTx}$	*714*	*patients saved failure under treatment*
$P(+\|succeed)$	*0.47*	*sensitivity of the biomarker test*
$P(-\|fail)$	*0.75*	*specificity of the biomarker test*

Case 1. Trastuzumab trial as conducted with Her2 biomarker compared with open enrollment.

A similar analysis in [6] based on different sources forms the basis for the inputs to this case:

trastuzumab Her2 IHC+ v. none (a)				
				
		λ	*0.34*	
		ε	*7%*	
		$\varepsilon+\delta$	*9%*	
		− F	*61%*	
		+ F	*29%*	
		− S	*5%*	
		+ S	*5%*	
biomarker +/− F fail treatment, S succeed				

trastuzumab Her2 IHC+ v. none (a)		
λ	0.34	*proportion biomarker+*
ε	7%	*efficacy for biomarker-, net of control*
δ	9%	*Δ efficacy for biomarker+*
$E_B\ (\varepsilon+\delta)$	16%	*efficacy for biomarker+, net of control*
$E_R\ (\varepsilon+\lambda\delta)$	10%	*efficacy for pooled biomarker +/-, net of control*
	29%	*treatment effect in control group*
	36%	*treatment effect in biomarker-*
	45%	*treatment effect in biomarker+*
ρ_{BR}	0.42	*sample ratio*
N_B	469	*sample size biomarker case*
N_R	1122	*sample size no-biomarker case*
N_B/λ	1379	*patients screened to find biomarker+*
$(N_R-N_B)/N_R$	58%	*% reduction in variable cost*
C_V	$6,971,662	*variable cost savings*
$N\Delta_{FailTx}$	308	*patients saved failure under treatment*
$P(+\|succeed)$	0.54	*sensitivity of the biomarker test*
$P(-\|fail)$	0.68	*specificity of the biomarker test*

Case 2. Trastuzumab trial as conducted with Her2 biomarker compared with open enrollment, from Manke *et al.* [6]

In both these cases the trial using the biomarker was less than half the size. What is really remarkable, though, is the number of individuals screened relative to the size of the trial. In each case the enrollment would ultimately be around one third of the screened population. Generally speaking a large number to screen may mean a considerably longer enrollment period, or more centers with their attendant cost.

Screening size, remember is strongly dependent on λ, as simply N_B/λ.

PTEN deficiency as a biomarker of trastuzumab resistance

It has been reported that PTEN deficiency contributes to trastuzumab resistance, in vitro, in animal xenograft, and in a small group of cancer patients [1]. This case compares the reference case mentioned above, trastuzumab for metastatic breast cancer, conducted using Her2 IHC2+ as a biomarker, with a hypothetical trial using PTEN inactivation as additional marker of resistance, enrolling only Her2 IHC2+ PTEN+ patients.

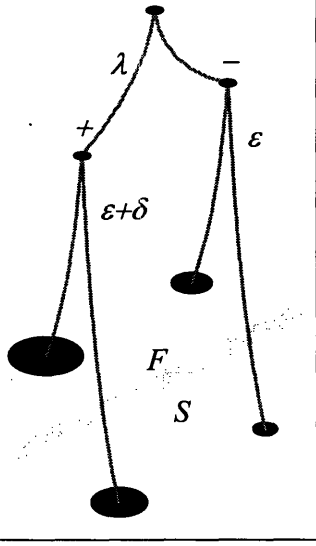
Treatment effect is 65.8% (Table 11). Frequency of PTEN IRS5+among the Her2+ study participants was 38/47 so I take the probability of finding a patient to test biomarker+ to be 0.81. The control group response is taken to be the same as in the reference, 27%. Biomarker negative now means Her2 IHC2+; response in this population is 50% from the reference trial. (Case 3).

PTEN status	Response Rate
deficient (IRS 0-4)	11.1%
positive (IRS 5+)	65.8%

Table 11. Response rate for PTEN deficient and positive patients. Cutoff for positive PTEN assessment at immunoreactive score (IRS) 7+. Response as complete response or partial response.

Biomarker	Proportion of IHC2+ who are also PTEN IRS5+
PTEN- IRS 0-4|Her2 IHC2+	11/47
PTEN IRS 5+|Her2 IHC2+	38/47

Table 12. Proportion of Her2 IHC2+ study participants who are PTEN+ and PTEN-.

trastuzumab Her2 IHC+ PTEN+ v. Her2 IHC+				
				
		λ	*0.81*	
		ε	*23%*	
		$\varepsilon+\delta$	*16%*	
	–	F	*15%*	
	+	F	*49%*	
	–	S	*4%*	
	+	S	*31%*	
biomarker +/– *F fail treatment, S succeed*				

trastuzumab Her2 IHC+ PTEN+ v. Her2+		
λ	*0.81*	*proportion biomarker+*
ε	*23%*	*efficacy for biomarker-, net of control*
δ	*16%*	*Δ efficacy for biomarker+*
$E_B(\varepsilon+\delta)$	*39%*	*efficacy for biomarker+, net of control*
$E_R(\varepsilon+\lambda\delta)$	*36%*	*efficacy for pooled biomarker +/-, net of control*
	27%	*treatment effect in control group*
	50%	*treatment effect in biomarker-*
	66%	*treatment effect in biomarker+*
ρ_{BR}	*0.85*	*sample ratio*
N_B	*802*	*sample size biomarker case*
N_R	*938*	*sample size no-biomarker case*
N_B/λ	*991*	*patients screened to find biomarker+*
$(N_R-N_B)/N_R$	*15%*	*% reduction in variable cost*
C_V	*$1,456,693*	*variable cost savings*
$N\Delta_{FailTx}$	*56*	*patients saved failure under treatment*
$P(+\|succeed)$	*0.88*	*sensitivity of the biomarker test*
$P(-\|fail)$	*0.23*	*specificity of the biomarker test*

Case 3. Hypothetical trastuzumab trials using biomarker of HER2 IHC+ plus PTEN activity.

Selection based on PTEN in this example yields an incremental improvement over Her2 IHC2+. Although response is higher among the biomarker+ group in this case (16% higher), the overall reduction in variable cost is modest

(15%). This is attributable in part to the small proportion of the patients less likely to respond who are excluded (19%).

PTEN deficiency with a more stringent Her2 test

When Her2 expression is measured with fluorescence in situ hybridization (FISH), the specificity of the PTEN+ Her2+ biomarker test is even greater.

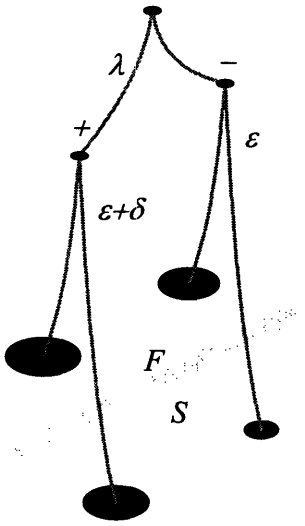
This hypothetical case supposes enrollment criteria of PTEN IRS5+ Her2 FISH+, compared with the reference case using Her2 IHC2+ as above. Response rate is 71% for patients with this biomarker. (Table 13). The probability λ of finding biomarker+ is taken to be the compound probability of FISH+ | Her2 IHC2+ (39/47), and PTEN IRS5+ | FISH+ (31/39) (Table 12 and Table 14).

Biomarker	response rate
PTEN- (IRS 0-3) Her2 IHC2+ FISH+	12.5%
PTEN IRS4+ Her2 IHC2+ FISH+	71%

Table 13. Response rate is strikingly lower among Her2 FISH+ PTEN-, compared with Her2 FISH+ PTEN+[1], or Her2 IHC2+ .

Biomarker	Proportion of IHC2+ who are also Her2 FISH+
FISH+|Her2 IHC2+	39/47
FISH-|Her2 IHC2+	8/47

Table 14. Proportion of Her2 IHC2+ study participants who are FISH+ and FISH-.

trastuzumab Her2 FISH+ PTEN+ v. Her2 IHC+				
				
		λ	*0.66*	
		ε	*23%*	
		ε+δ	*21%*	
	–	F	*26%*	
	+	F	*37%*	
	–	S	*8%*	
	+	S	*29%*	
biomarker +/– *F fail treatment, S succeed*				

trastuzumab Her2 FISH+ PTEN+ v. Her2+ IHC2+		
λ	*0.66*	*proportion biomarker+*
ε	*23%*	*efficacy for biomarker-, net of control*
δ	*21%*	*Δ efficacy for biomarker+*
$E_B\ (\varepsilon+\delta)$	*44%*	*efficacy for biomarker+, net of control*
$E_R\ (\varepsilon+\lambda\delta)$	*37%*	*efficacy for pooled biomarker +/-, net of control*
	27%	*treatment effect in control group*
	50%	*treatment effect in biomarker-*
	71%	*treatment effect in biomarker+*
ρ_{BR}	*0.70*	*sample ratio*
N_B	*661*	*sample size biomarker case*
N_R	*938*	*sample size no-biomarker case*
N_B/λ	*1002*	*patients screened to find biomarker+*
$(N_R-N_B)/N_R$	*30%*	*% reduction in variable cost*
C_V	*$2,955,807*	*variable cost savings*
$N\Delta_{FailTx}$	*111*	*patients saved failure under treatment*
$P(+\|succeed)$	*0.79*	*sensitivity of the biomarker test*
$P(-\|fail)$	*0.42*	*specificity of the biomarker test*

Case 4. Hypothetical trastuzumab trials using biomarker of Her2 FISH+ PTEN+ compared with trial as conducted using Her2 IHC2+.

Replacing Her2 IHC2+ with Her2 FISH+ yielded an improvement of 30% in savings versus 15% (Case 3). This difference is attributable in large part to

the exclusion of more patients less likely to respond, and in a smaller degree to the marginally higher response rate among biomarker+ patients.

Note that these are both cases in which the number of patients screened is close to the reference population for the pooled biomarker +/- trial.

A practical note on the validity of this biomarker: The evidence in [1] for PTEN inactivity as a marker of clinical resistance is thin; this limits our confidence that the estimation parameters used as inputs to the model are even close to true. The magnitude of the predictive value of PTEN activity, though, is large, and source of a considerable benefit as modeled.

Tamoxifen for invasive breast cancer

This case applies a report of a potential biomarker to the prospect of selecting invasive breast cancer patients least likely to relapse under treatment with tamoxifen.

The anti-estrogen drug tamoxifen was first approved for the treatment of metastatic breast cancer in postmenopausal patients in the United Kingdom in 1973 and subsequently in the United States in 1977 [2]. Prior to this, removal of the ovaries (or ablation) was performed to treat breast cancer, the first report of which was in 1896. [13, 14] Tamoxifen is used today in a number of breast cancer indications.

The study by Scarpi, *et al.* [2] aimed to investigate the relationship between outcome in node positive estrogen receptor-positive (ER+) invasive breast cancer patients treated with tamoxifen and the status of several potentially relevant pre-treatment biomarkers. In particular, in this study, progesterone receptor (PgR) and tumor proliferative status measured with ^{3}H-thymidine labeling index (TLI) using were found in multivariate analysis to have independent value to predict relapse (Table 15).

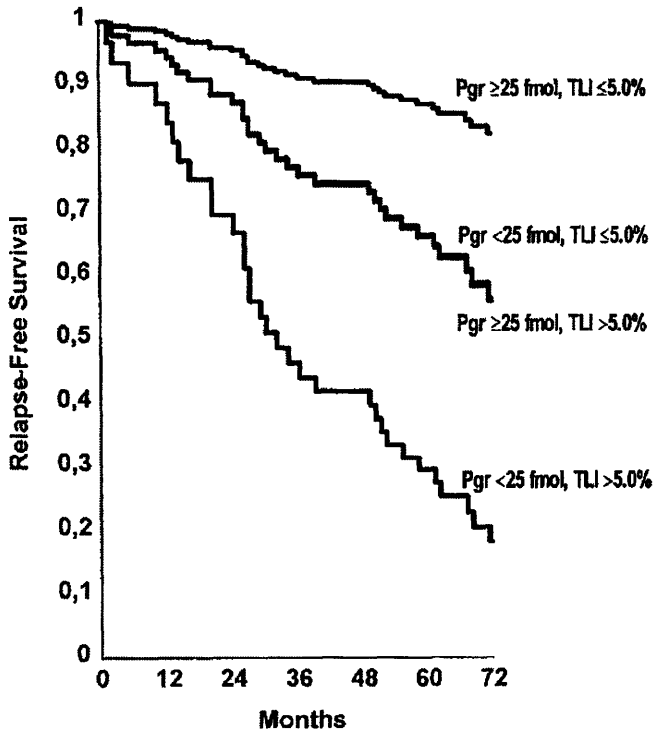


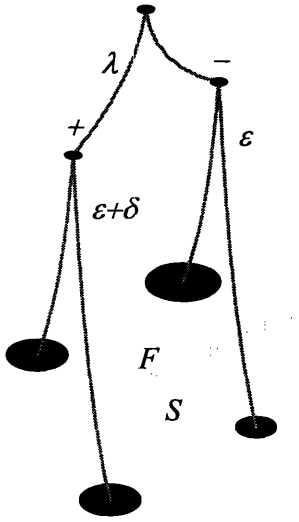
Table 15. Relapse free survival for subgroups determined by presence of progesterone receptor and by proliferative index [2].

	Biomarker Status			
	overall	PgR+/TLI-	PgR-/TLI- or PgR+/TLI+	PgR-/TLI+
Five year relapse-free survival (RFS)	75%	87%	66%	29%
			62%	
group size	119/119	62/119	57/119	

Table 16. Relapse-free survival at five years for the study group and subgroups divided by biomarker status.

This case presents something of a problem because the trials leading to the development of tamoxifen would have been conducted in a different regulatory and overall health care context than trials we could propose today. There is no suitable reference trial to select from. For reference purposes, I propose that we are comparing the development of tamoxifen, as if it were new, with a standard-of-care that has one half of tamoxifen's effect as demonstrated in these clinical results.

As in the case with PTEN, assuming these data do indicate a true predictive association of the biomarker, those values are probably somewhat different than the values I extract here. In using these data (Table 16) for this case analysis I do not imply that our confidence in the predictive value of these biomarkers is strong enough for one to make clinical or drug development decisions on this foundation. Nor did the authors make such a claim. In practice validation will come with successive trials measuring the biomarker.

tamoxifen ER+ PGR+ TLI- v. ER+			
			
	λ	*0.52*	
	ε	*24%*	
	ε+δ	*25%*	
	− F	*36%*	
	+ F	*26%*	
	− S	*12%*	
	+ S	*26%*	
biomarker +/− *F fail treatment, S succeed*			

tamoxifen ER+ PGR+ TLI- v. ER+		
λ	*0.52*	*proportion biomarker+*
ε	*24%*	*efficacy for biomarker-, net of control*
δ	*25%*	*Δ efficacy for biomarker+*
$E_B\ (\varepsilon+\delta)$	*50%*	*efficacy for biomarker+, net of control*
$E_R\ (\varepsilon+\lambda\delta)$	*38%*	*efficacy for pooled biomarker +/-, net of control*
	38%	*treatment effect in control group*
	62%	*treatment effect in biomarker-*
	87%	*treatment effect in biomarker+*
ρ_{BR}	*0.54*	*sample ratio*
N_B	*543*	*sample size biomarker case*
N_R	*1000*	*sample size no-biomarker case*
N_B/λ	*1043*	*patients screened to find biomarker+*
$(N_R-N_B)/N_R$	*46%*	*% reduction in variable cost*
C_V	*$4,875,411*	*variable cost savings*
$N\Delta_{FailTx}$	*175*	*patients saved failure under treatment*
$P(+\|succeed)$	*0.69*	*sensitivity of the biomarker test*
$P(-\|fail)$	*0.58*	*specificity of the biomarker test*

Case 5. Hypothetical tamoxifen trial comparing all patients with best responding biomarker subgroup, PgR+/TLI-.

A dramatic reduction in seen in sample size when adding the hypothetical biomarker test, the variable cost savings is 39%.

The control treatment effect was contrived for this case to be half the tamoxifen reference response. If one substitutes more effective treatment for the controls, the gains in efficiency are improved (Table 17).

Assumed control effect	% savings in variable cost
25%	42%
30%	44%
35%	45%
37.5%	46%
40%	46%
45%	48%
50%	50%

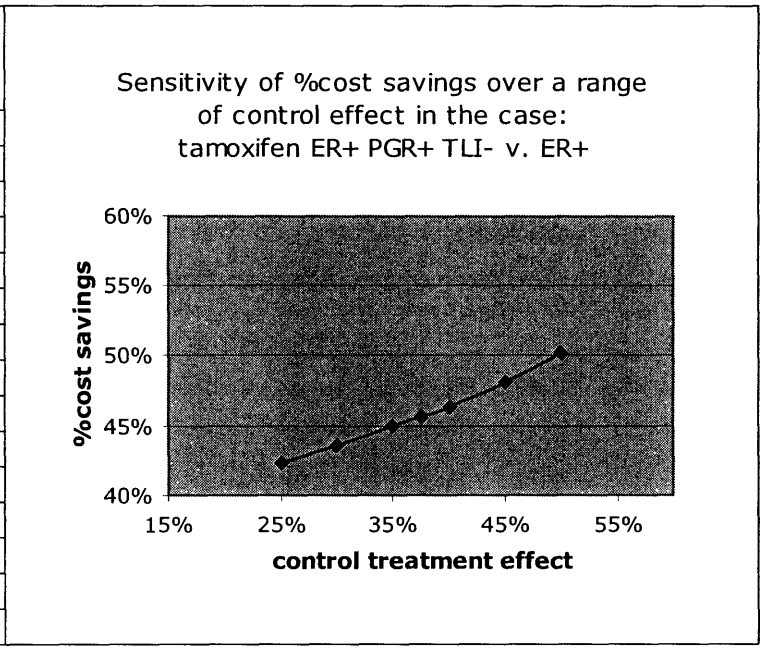


Table 17. Cost savings shows modest sensitivity in this case to control group effect.

Erlotinib for advanced non-small cell lung cancer

Erlotinib, an EGFR inhibitor, has been approved for the treatment of advanced or metastatic non-small cell lung cancer. These data are taken from the approval summary [15]. Although EGFR expression was not shown to be a good predictor of response to treatment with another EGFR inhibitor gefitinib, the erlotinib trial did show EGFR expression to have predictive value.

	N	Effect
erlotinib	424	8.96%
placebo	210	0.95%

Table 18. Tumor response (complete response or partial response) in treated and placebo groups [15].

	N	Effect
EGFR-	61	3.28%
EGFR+	69	11.6%
Unknown	294	9.52%

Table 19. Tumor response (complete response or partial response) in treated group according to EGFR status [15].

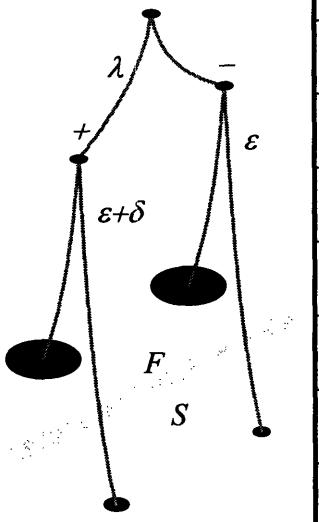
This set of data presents two difficulties for inclusion. One is that EGFR status is known in a subset of patients. This "over-determines" the values for effect in the biomarker positive, negative, and pooled population. I make an estimate for these (Table 20) taking care to give priority to the effect in the (largest) unknown biomarker group.

Group	N	Effect
estimated EGFR- in unknown	294*61/130 = 138	=(3.28/9.52)*8.96% =3.09%
estimated EGFR+ in unknown	294*69/130 = 156	=(11.6/9.52)*8.96% =10.92%
known+estimated EGFR-	138+61=199	=(3.09%*199+61*3.28%)/199 =**4.10%**
known+estimated EGFR+	156+69=225	=(10.92%*156+69*11.6%)/225 =**11.1%**
Pooled	424	8.96%

Table 20. Reconciling the known and unknown EFGR status of the treatment group. I assume that EGFR status in the unknown group is in the same ratio as in the known, and apportion the effect in the pooled group according to the ratio of effect in the known EGFR+ and EGFR- subgroups.

The second difficulty is that the placebo group is not the same size as the treatment group, in contradiction to an assumption of the model. For the purpose of the model I force the size of the placebo group to match the treatment group.

A contradiction to an assumption in the model which would affect model results is that EGFR+ status in the control group is a marker of poor prognosis [15]. (The model assumes control group effect is the same in both the reference and hypothetical case.) This implies that in a trial enrolling only EGFR+ patients the treatment effect will be easier to show, making any differences shown in the model more conservative than they should be. However, since control effect is so close to zero, I don't believe the distortion is significant.

erlotinib EGFR+ v. none				
				
		λ	0.53	
		ε	3.2%	
		ε+δ	7.0%	
	−	F	45%	
	+	F	48%	
	−	S	1%	
	+	S	5%	
biomarker +/− *F fail treatment, S succeed*				

erlotinib EGFR+ v. none		
λ	*0.53*	*proportion biomarker+*
ε	*3.2%*	*efficacy for biomarker-, net of control*
δ	*7.0%*	*Δ efficacy for biomarker+*
$E_B\ (\varepsilon+\delta)$	*10%*	*efficacy for biomarker+, net of control*
$E_R\ (\varepsilon+\lambda\delta)$	*6.9%*	*efficacy for pooled biomarker +/-, net of control*
	1.0%	*treatment effect in control group*
	4.1%	*treatment effect in biomarker-*
	11.1%	*treatment effect in biomarker+*
ρ_{BR}	*0.62*	*sample ratio*
N_B	*527*	*sample size biomarker case*
N_R	*848*	*sample size no-biomarker case*
N_B/λ	*994*	*patients screened to find biomarker+*
$(N_R-N_B)/N_R$	*38%*	*% reduction in variable cost*
C_V	*$3,422,584*	*variable cost savings*
$N\Delta_{FailTx}$	*158*	*patients saved failure under treatment*
$P(+\|succeed)$	*0.78*	*sensitivity of the biomarker test*
$P(-\|fail)$	*0.49*	*specificity of the biomarker test*

Case 6. Hypothetical erlotinib trial comparing the reference trial to one in which only EGFR+ patients are enrolled.

Discussion

The quantitative results are summarized here:

case	*variable cost reduction*	*patients screened to find biomarker + patients*	*patients screened in trial without biomarker*	*patients saved failure under treatment*
trastuzumab Her2 IHC+ v. none	62.3%	3350	2490	714
trastuzumab Her2 IHC+ v. none (a)	58.2%	1379	1122	308
trastuzumab Her2 IHC+ PTEN+ v. Her2 IHC+	14.5%	991	938	56
trastuzumab Her2 FISH+ PTEN+ v. Her2 IHC+	29.5%	1002	938	111
tamoxifen ER+ PGR+ TLI- v. ER+	45.7%	1043	1000	175
erlotinib EGFR+ v. none	37.8%	994	848	158

Table 21. Case summary. *Variable cost reduction*. Example: costs of $80M with biomarker versus $100M without is said to measure a 20% variable cost reduction. Another component of trial size is the number of *patients screened* in the biomarker trial to find the requisite number of biomarker positive patients to randomize. This is meaningful in relation to the number of patients screened in an untargeted trial. *Patients saved failure under treatment* is the *expected* difference, in the targeted and untargeted design, of the patients given experimental treatment who will not benefit from it.

A wide range is not a surprise, as biomarker prevalence, and treatment response conditioned on biomarker status vary widely among the cases, along with other factors.

worth of one biomarker over another

Comparisons between choices of biomarker are interesting, such as with this case

- trastuzumab Her2 IHC+ PTEN+ v. Her2 IHC+

and this one

- trastuzumab Her2 FISH+ PTEN+ v. Her2 IHC+,

which compare designs using progressively more refined biomarkers.

It is remarkable to me how the substitution of FISH for IHC in those cases can wring an extra 15% enrollment size reduction, 29.5% variable cost savings with FISH versus 14.5% with IHC. However the inputs to the model, the biomarker values were chosen from literature as plausible, not well supported; so it would be prudent to hold some suspicion that the values are way off base.

screening size

Targeted trials which are hungry for many patients to screen are those in which

a) biomarker positive patients are relatively rare.

b) the drug has a small improvement of effect over control treatment.

A trastuzumab trial held as a reference case would be more than twice larger if it were to have been done without using a biomarker; but it would have screened 25% less patients.

I have mentioned that in some cases the assumption that the control group in the biomarker and the reference case have the same effect is incorrect. In the case of Her2 for breast cancer, it is clearly incorrect, because Her2 is a poor prognostic indicator [16]. In that case which compares the trial as conducted versus the hypothetical case in which no biomarker was used (Case 1), the effect of this poor assumption is to attribute worse prognosis to the control group in the hypothetical no biomarker case than we truly expect. This leads to an overestimate of effect in the untargeted case. In the end, the error causes the model's value for savings to be an underestimate.

In the case where we compare the Her2 IHC2+ biomarker with Her2 FISH+, the assumption probably does not cause as large an inaccuracy, since one may expect the two control populations to be more alike in their response.

Appendix A.

The sample size equations used in this work, excerpted from [5].

The Ury and Fleiss expression for the sample size of the untargeted design is:

$$n = (n_a/4)[1 + \sqrt{2w + 1}]^2$$

where

$$n_a = [z_\alpha \sqrt{2\bar{p}\bar{q}} + z_\beta \sqrt{p_e(1 - p_e) + p_c(1 - p_c)}]^2/\bar{\delta}^2,$$

$$\bar{\delta} = \gamma\delta_0 + (1 - \gamma)\delta_1,$$

$$p_e = p_c + \bar{\delta},$$

$$\bar{p} = (p_c + p_e)/2,$$

$$\bar{q} = 1 - \bar{p}, \text{ and}$$

$$w = \bar{\delta}/[(z_\alpha + z_\beta)^2\bar{p}\bar{q}].$$

The constants z_α and z_β denote the 100 (1-α) and 100 (1-β) percentiles of the standard normal distribution.

For the targeted design we add the symbol T. The response probability in the experimental group is $p_e^T = p_c + \delta_1$ and the expression for the sample size becomes:

$$n^T = (n_a^T/4)[1 + \sqrt{1 + 2w_T}]^2$$

where,

$$n_a^T = [z_\alpha \sqrt{2\bar{p}_T\bar{q}_T} + z_\beta \sqrt{p_e^T(1 - p_e^T) + p_c(1 - p_c)}]^2/\delta_1^2,$$

$$\bar{p}_T = (p_e^T + p_c)/2,$$

$$\bar{q}_T = 1 - \bar{p}_T \text{ and}$$

$$w_T = \delta_1/[(z_\alpha + z_\beta)^2\bar{p}_T\bar{q}_T]$$

The relative efficiency of the two designs with regard to number of randomized patients is therefore given by equation (A) with f defined by

$$f = \left[\frac{z_\alpha \sqrt{2\bar{p}\bar{q}} + z_\beta \sqrt{p_e(1 - p_e) + p_c(1 - p_c)}}{z_\alpha \sqrt{2\bar{p}_T\bar{q}_T} + z_\beta \sqrt{p_e^T(1 - p_e^T) + p_c(1 - p_c)}} \frac{1 + \sqrt{1 + 2w}}{1 + \sqrt{1 + 2w_T}}\right]^2.$$

A Rosetta stone relating the terms used in this model to the formulas I use from Simon & Maitournam[5].

term used here	*example*	*Definition*	*correspondence to Simon & Maitournam*
λ	0.28	*proportion biomarker+*	$1-\gamma$
ε	10%	*efficacy for biomarker-, above control*	δ_0
δ	13%	*Δ efficacy for biomarker+*	$\delta_1 - \delta_0$
$E_B\ (\varepsilon+\delta)$	23%	*efficacy for biomarker+, above control*	δ_1
	14%	*treatment effect in control group*	p_c
	27%	*treatment effect in biomarker- group*	$p_c+\delta_0$
	37%	*treatment effect in biomarker+ group*	$p_c+\delta_1$
$E_R\ (\varepsilon+\lambda\delta)$	50%	*efficacy for pooled biomarker +/-, above control*	$\delta_0\ (\gamma) + \delta_1\ (1-\gamma)$
ρ_{BR}	0.57	*sample ratio*	$1\ /\ (n\ /\ n^T)$
N_B	938	*N biomarker case*	
N_R	1658	*N no-biomarker case*	
N_B/λ	3350	*patients screened to find biomarker+*	
$(N_R-N_B)/N_R$	43%	*% reduction in variable cost*	
C_V	$7,687,791	*variable cost savings*	
$N\Delta_{FailTx}$	355	*patients saved failure under treatment*	
$P(+\|succeed)$	0.47	*sensitivity of the biomarker test*	
$P(-\|fail)$	0.75	*specificity of the biomarker test*	

Appendix B.

Validation samples

This table shows an example of some of the synthetic data that was run through the website associated with [5]. [††]

p_c	*0.2*	*0.3*	*0.3*	*0.2*	*0.2*	*0.2*	*0.2*
γ	*0.5*	*0.3*	*0.3*	*0.5*	*0.5*	*0.5*	*0.5*
δ_1	*0.3*	*0.3*	*0.3*	*0.4*	*0.4*	*0.4*	*0.4*
δ_0	*0*	*0.2*	*0.2*	*0.2*	*0.2*	*0.2*	*0.2*
α	*0.05*	*0.05*	*0.05*	*0.05*	*0.05*	*0.005*	*0.0005*
β	*0.1*	*0.1*	*0.01*	*0.2*	*0.1*	*0.1*	*0.1*

See appendix A for symbol definitions. The calculation of the model presented here and from the website agreed down to the tolerance of integer rounding errors.

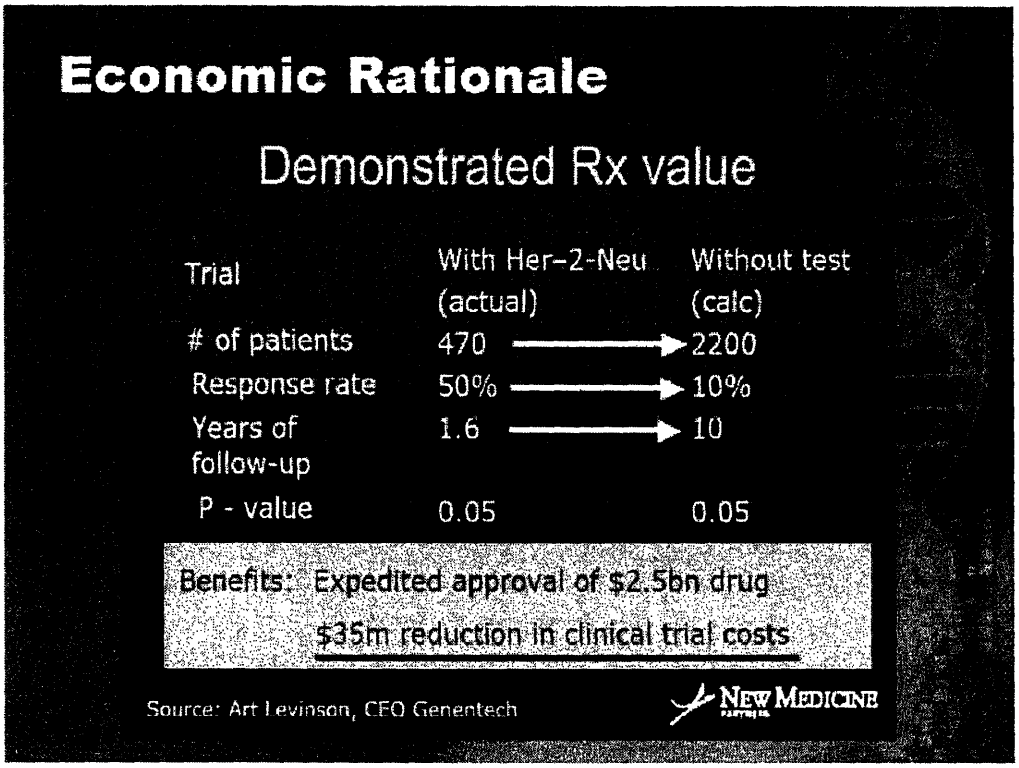
The following examples are synthetic data used in [5]. The results from my model match the website calculator results, and to the tolerance of visual inspection, match the results published in [5].

p_c	*0.1*	*0.1*	*0.5*	*0.5*	*0.1*	*0.1*	*0.5*	*0.5*
γ	*0.5*	*0.5*	*0.5*	*0.5*	*0.4*	*0.4*	*0.4*	*0.4*
δ_1	*0.2*	*0.4*	*0.2*	*0.4*	*0.2*	*0.4*	*0.2*	*0.4*
δ_0	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
α	*0.05*	*0.05*	*0.05*	*0.05*	*0.05*	*0.05*	*0.05*	*0.05*
β	0.10	0.10	0.10	0.10	*0.10*	*0.10*	*0.10*	*0.10*

[††] http://linus.nci.nih.gov/~simonr/boep.html

Appendix C.

This slide by Art Levinson, as CEO of Genentech, states that the value of performing the trastuzumab clinical trials using Her2 as a predictive biomarker was faster approval of a $2.5 billion drug, and a $35 million reduction in trial costs.



source:[8]

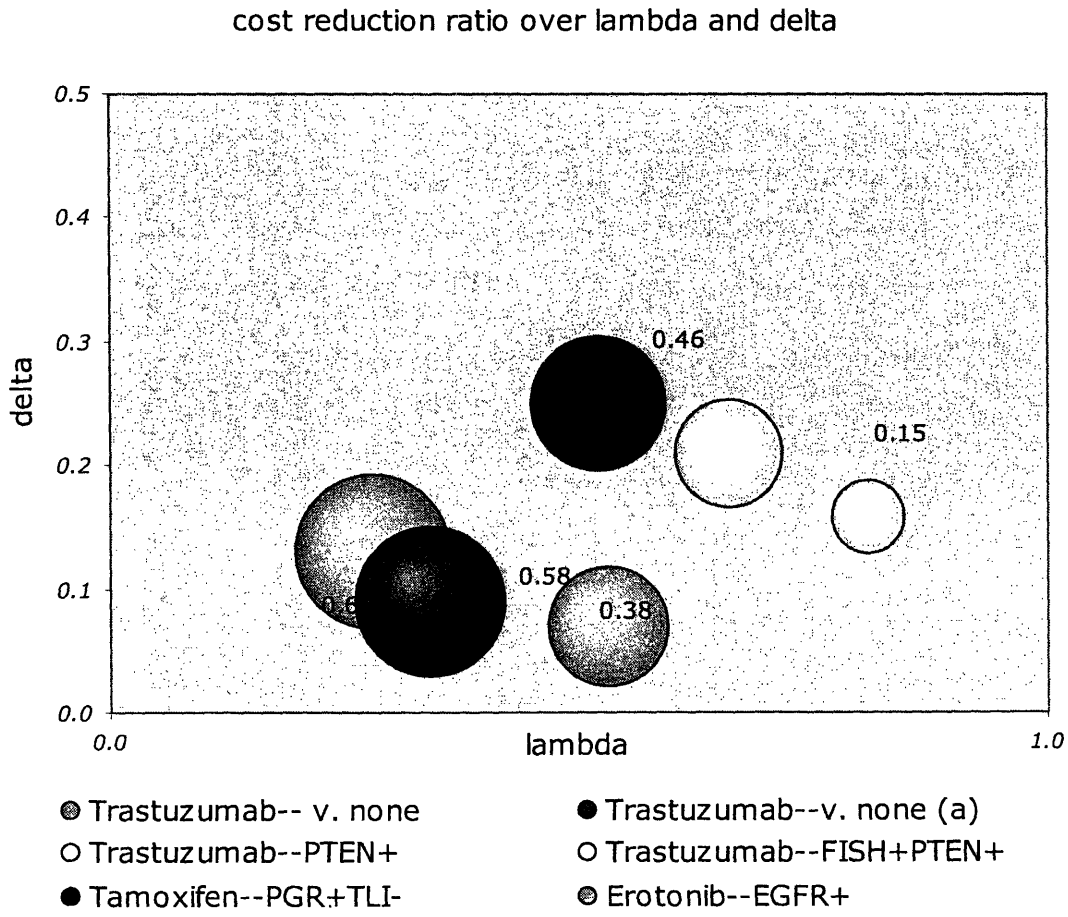
	with Her2 (actual trial)	*without Her2 (hypothetical trial)*
number of patients	*470*	*2200*
response rate	*50%*	*10%*
years of follow-up	*1.6*	*10*

Although based on a different trial data, it is interesting to compare the estimates of savings reported here with this estimate from Genentech.

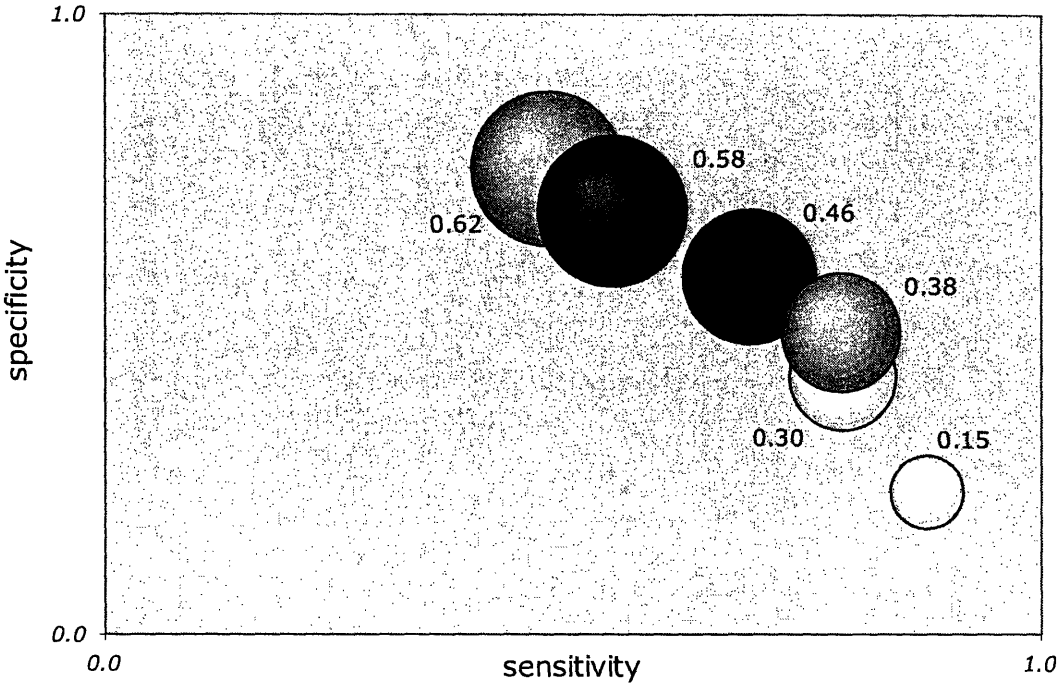
The sample size ratio of 470/2200 = 0.21 reported by Levinson compares with 0.38 in the case examined here. The variable cost reduction estimated in my model is $16.6M, based on a sample size of 938. This is a great deal less than the $35M stated by Genentech.

Note that the stated reduction in follow-up time, given the market value of the drug, far outweighs the reduction in clinical trial costs. Although few drugs have even close to this commercial value, this result implies that this source of value should be carefully modeled in practice.

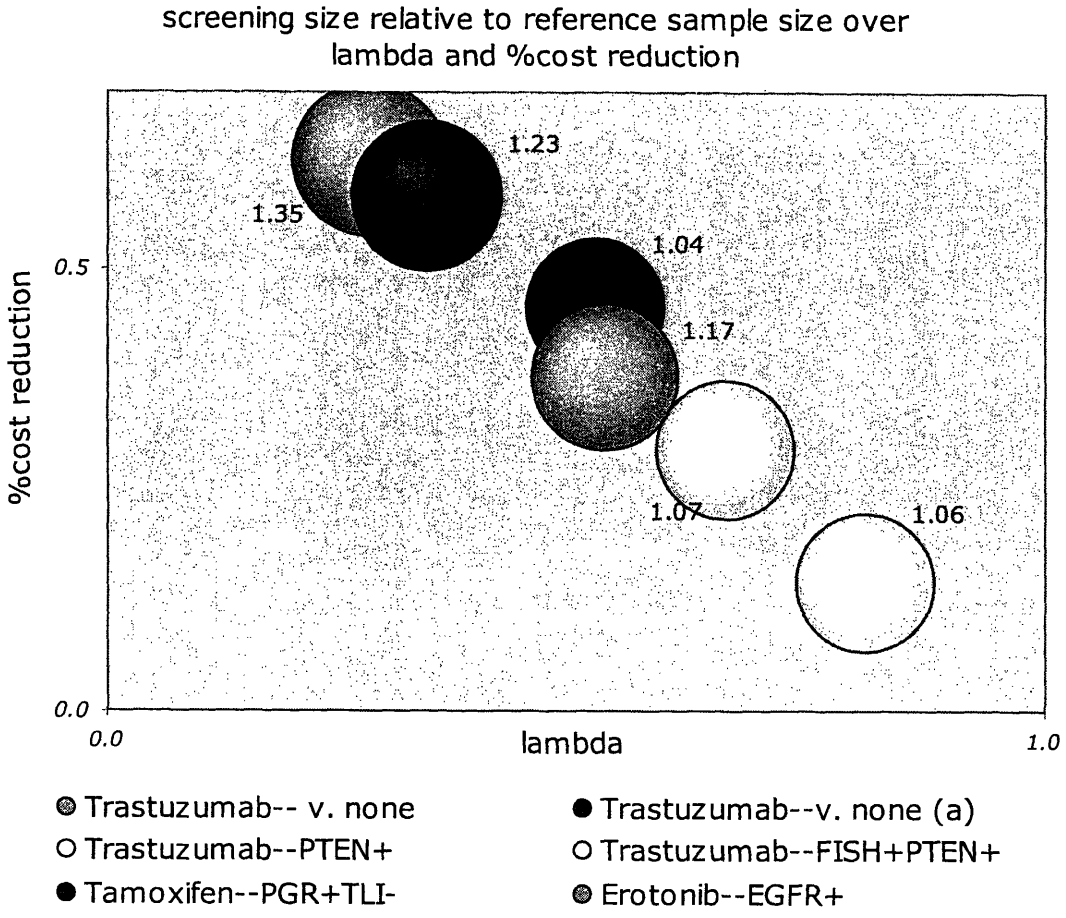
Exploratory charts

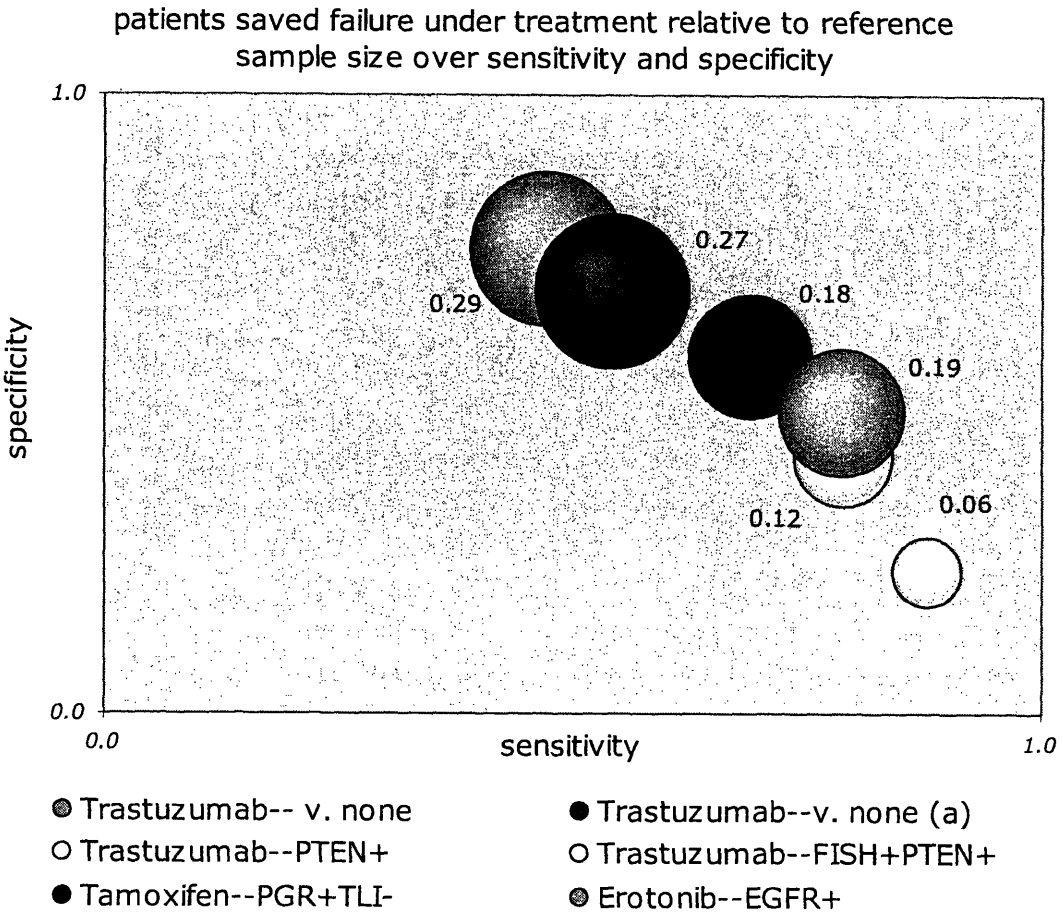


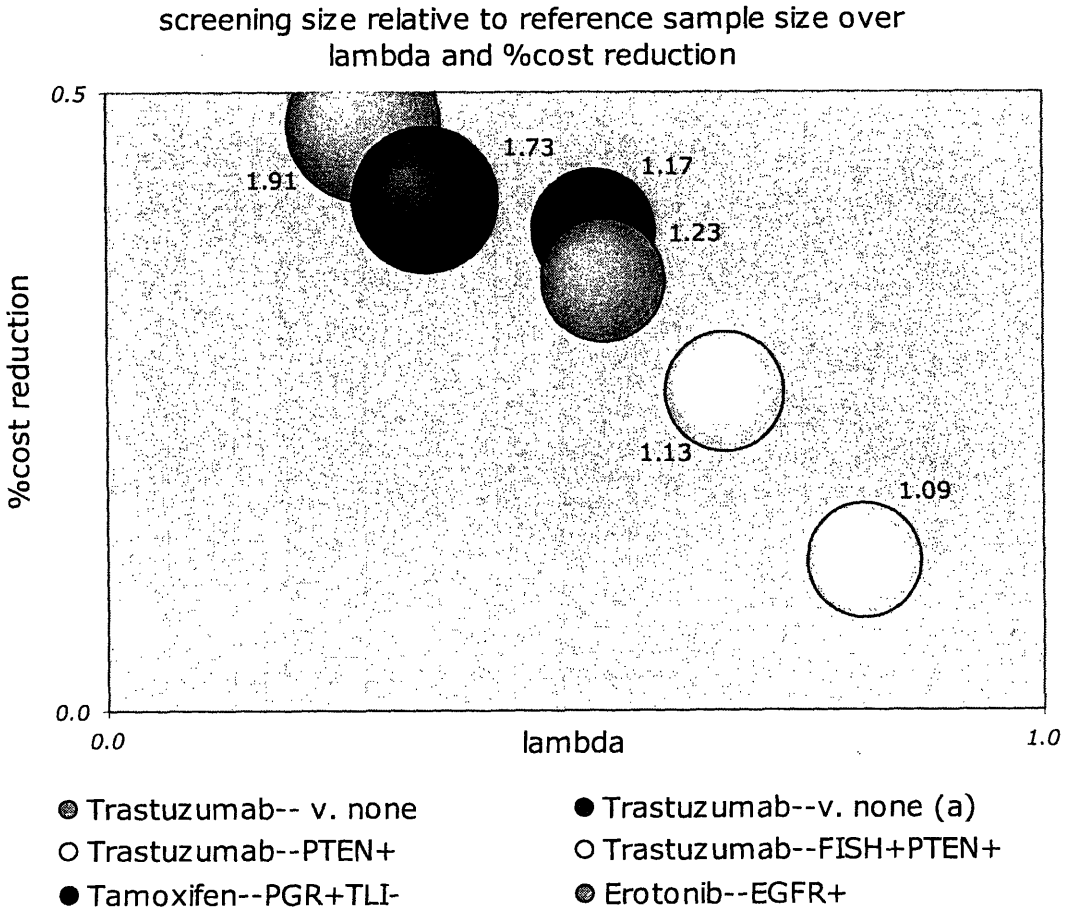
cost reduction ratio over sensitivity and specificity



- Trastuzumab-- v. none
- Trastuzumab--v. none (a)
- Trastuzumab--PTEN+
- Trastuzumab--FISH+PTEN+
- Tamoxifen--PGR+TLI-
- Erotonib--EGFR+


screening size relative to reference sample size over lambda and %cost reduction
1.23
1.35
1.04
1.17
1.07
1.06
0.5
0.0
%cost reduction
0.0
lambda
1.0
Trastuzumab-- v. none
Trastuzumab--v. none (a)
Trastuzumab--PTEN+
Trastuzumab--FISH+PTEN+
Tamoxifen--PGR+TLI-
Erotonib--EGFR+


patients saved failure under treatment relative to reference sample size over sensitivity and specificity
1.0
0.0
specificity
0.0
sensitivity
1.0
0.29
0.27
0.18
0.19
0.12
0.06
Trastuzumab-- v. none
Trastuzumab--PTEN+
Tamoxifen--PGR+TLI-
Trastuzumab--v. none (a)
Trastuzumab--FISH+PTEN+
Erotonib--EGFR+


screening size relative to reference sample size over lambda and %cost reduction
0.5
0.0
%cost reduction
0.0
lambda
1.0
1.91
1.73
1.17
1.23
1.13
1.09
Trastuzumab-- v. none
Trastuzumab--v. none (a)
Trastuzumab--PTEN+
Trastuzumab--FISH+PTEN+
Tamoxifen--PGR+TLI-
Erotonib--EGFR+

References

1. Nagata, Y., et al., *PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients.* Cancer Cell, 2004. **6**(2): p. 117-27.
2. Scarpi, E., et al., *Biomarker prediction of clinical outcome in operable breast cancer patients treated with tamoxifen.* Breast Cancer Res Treat, 2001. **68**(2): p. 101-10.
3. Baselga, J. and C.L. Arteaga, *Critical update and emerging trends in epidermal growth factor receptor targeting in cancer.* J Clin Oncol, 2005. **23**(11): p. 2445-59.
4. Vernon, J.A. and W.K. Hughen, *The Future of Drug Development: The Economics of Pharmacogenomics*, in *Working Paper 11875*. 2005, National Bureau of Economic Research: Cambridge, MA.
5. Simon, R. and A. Maitournam, *Evaluating the efficiency of targeted designs for randomized clinical trials.* Clin Cancer Res, 2004. **10**(20): p. 6759-63.
6. Manke, I., et al., *Biomarkers in Clinical Development of Cancer Drugs: Challenges and Opportunities*. 2004, Massachusetts Institute of Technology.
7. 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**: p. 14-21.
8. Press, M. and S.A. Seelig, *Lessons Learned from the Development of a Diagnostic to Predict Response to Herceptin*, in *Targeted Medicine: From Concept to Clinic symposium*. 2004, Scientific American: New York, NY.
9. Parexel, *Parexel's Pharmaceutical R& D Statistical Sourcebook 1998*. 1998, Waltham, MA.
10. Sahr, R.C., *Inflation Conversion Factors for Dollars 1665 to Estimated 2015*. 2005.
11. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* N Engl J Med, 2001. **344**(11): p. 783-92.
12. Seidman, A.D., et al., *Weekly trastuzumab and paclitaxel therapy for metastatic breast cancer with analysis of efficacy by HER2 immunophenotype and gene amplification.* J Clin Oncol, 2001. **19**(10): p. 2587-95.
13. **Beatson, G.T., *On the treatment of inoperable cases of carcinoma of the mamma: suggestions for a new method of treatment with illustrative cases.* Lancet, 1896. ii: p. 104-107.**
14. Clarke, M.J., *Ovarian ablation in breast cancer, 1896 to 1998: milestones along hierarchy of evidence from case report to Cochrane review.* Bmj, 1998. **317**(7167): p. 1246-8.
15. Johnson, J.R., et al., *Approval summary for erlotinib for treatment of patients with locally advanced or metastatic non-small cell lung cancer after failure of at least one prior chemotherapy regimen.* Clin Cancer Res, 2005. **11**(18): p. 6414-21.

16. Ross, J.S., et al., *The Her-2/neu gene and protein in breast cancer 2003: biomarker and target of therapy.* Oncologist, 2003. **8**(4): p. 307-25.