Identification of Pharmacogenomic Biomarker Classifiers in Cancer Drug Development

Richard Simon

To Appear in *Pharmacogenomics, Anticancer Drug Discovery and Response*, (F Innocenti editor), Humana Press

Physicians need improved tools for selecting treatments for individual patients. Many syndromes traditionally viewed as individual diseases are heterogeneous in molecular pathogenesis and treatment responsiveness. This results in treatment of many patients with ineffective drugs and leads to the conduct of large clinical trials to identify small average treatment benefits for heterogeneous groups of patients. New genomic and proteomic technologies provide powerful tools for the selection of patients likely to benefit from a therapeutic without unacceptable adverse events. In this chapter we attempt to clarify how pharmacogenomic biomarker classifiers of the patients most likely to benefit from a drug can be identified and utilized during clinical development.

Keywords: Pharmacogenomics, biomarker, genomics, DNA microarray, clinical trial design, validation

Richard Simon, D.Sc.
Biometric Research Branch
National Cancer Institute
9000 Rockville Pike
Bethesda MD 20892-7434
U.S.A.
301.496-0975 (tel)
301.402-0560 (fax)

rsimon@mail.nih.gov

1. Introduction

Physicians need improved tools for selecting treatments for individual patients. For example, many cancer treatments benefit only a minority of the patients to whom they are administered. Being able to predict which patients are most likely to benefit would not only save patients from unnecessary toxicity and inconvenience, but might facilitate their receiving drugs that are more likely to help them. In addition, the current over-treatment of patients results in major expense for individuals and society, an expense which may not be indefinitely sustainable. In this paper we will address some key issues in the validation of pharmacogenomic classifiers.

2. Pharmacogenomic Biomarker Classifiers

Much of the discussion about disease biomarkers is in the context of markers which measure some aspect of disease status, extent, or activity. Such biomarkers are often proposed for use in early detection of disease or as a surrogate endpoint for evaluating prevention or therapeutic interventions. The validation of such biomarkers is difficult for a variety of reasons, but particularly because the molecular pathogenesis of many diseases is incompletely understood and hence it is not possible to establish the biological relevance of a measure of disease status.

A pharmacogenomic biomarker is any pre-treatment measurable quantity that can be used to select treatment; for example, the result of an immunohistochemical assay for a single protein, the abundance of a protein in serum, the abundance of mRNA transcripts for a gene in a sample of disease tissue or the presence/absence status of a specified germ-line polymorphism or tumor mutation. A pharmacogenomic biomarker classifier is a mathematical function that translates the biomarker values to a set of prognostic categories. These categories generally correspond to levels of predicted clinical outcome. With the advent of gene expression profiling, it is increasingly common to define composite pharmacogenomic biomarker classifiers based on the levels of expression of dozens of genes. For a fully specified classifier, however, all of the parameters and cut-points are specified for determining how to weight the different components and how to map the multivariate data into a defined set of categories. A completely defined classifier can be used to select patients and stratify patients for therapy in clinical trials that enable the clinical value of the classifier to be evaluated. Specifying only the genes involved does not enable one to structure prospective clinical validation experiments in which patients are assigned or stratified in prospectively well defined ways.

3 Types of Pharmacogenomic Biomarker Classifiers

Pharmacogenomic biomarkers can either be defined based on the known molecular target of the drug or empirically developed by comparing responders versus non-responders with regard to whole genome tumor characterizations such as transcript expression profiling. The former approach is preferable for a variety of reasons. First, a biomarker with a strong biological rationale linked to the mechanism of action of the drug is more satisfying than a “black box” classifier. Second, development time is likely to be shorter, and finally a reproducible assay for a single gene/protein biomarker may be implementable on a platform that does not require fresh/frozen tumor.

Mutation-targeted drugs are drugs which are specific to a mutated gene which is driving the growth of a tumor. For example, there is currently considerable interest in developing drugs specific for mutated B-raf which is present in approximately 60 percent of human melanomas. Such drugs have automatic pharmacogenomic biomarkers based on assaying for the presence of the mutation. In general, the mutation could either be a point mutation as in B-raf, gene amplification, or deletion.

Many current cancer drugs are selected to inhibit oncogenes that are mutated in some tumors. Although the drugs are not designed to be specific for the mutated forms of the associated proteins, presence or absence of a mutation can be used as a natural pharmacogenomic biomarker of the patients most likely to benefit from the drugs. Papadopoulous et al review the experience with molecularly targeted drugs of this type {Papadopoulos, 2006 #287}.

Although all chemotherapeutic drugs are “molecularly targeted” in the sense that they interact with specific intracellular components {Papadopoulos, 2006 #287}, in some cases the true target is not known when the drug is developed, in some cases there are multiple targets, and in many cases there is no obvious assay for determining the extent to which the target is driving tumor growth. In these cases one must generally use empirical methods to develop pharmacogenomic biomarker classifiers of the patients most likely to benefit from the drug. With this approach a training set of tumor specimens from patients who have responded to the drug is assayed and compared to a training set of specimens from patients who have not responded. The specimens are assayed, using either whole genome technology such as expression profiling or using assays based on candidate genes, and a predictive classifier is developed for identifying the tumors most likely to respond. In the next section, we will describe some aspects of the development of such classifiers using whole genome transcript expression profiling.

4. Developing Empirical Pharmacogenomic Classifiers Using Gene Expression

There are three components to the empirical approach of developing a predictive classifier. The first is determining which genes to include in the predictor. This is generally called “feature selection”. Including too many “noise variables” in the predictor usually reduces the accuracy of prediction. The second is specification of the mathematical function that will provide a predicted class label for any given expression vector. The third is parameter estimation. Most kinds of predictors have parameters that must be assigned values before the predictor is fully specified. For many kinds of predictors there is also a cut-point that must be specified for translating a quantitative predictive index into a predicted class label (eg 0 or 1) for binary class prediction problems.

Feature selection is usually based on identifing the genes that are differentially expressed among the classes when considered individually. For example, if there are two classes, one can compute a t-test or a modified t-test in which a hierarchical variance model is used for increasing the degrees of freedom for estimation of the gene-specific within-class variances {Wright, 2003 #284}. The logarithm of the expression measurements are used as the basis of the statistical significance tests. The genes that are significantly differentially expressed at a specified significance level are selected for inclusion in the class predictor. The stringency of the significance level used controls the number of genes that are included in the model. Although many computationally complex methods have been published to identify optimal sets of genes which together provide good discrimination, little compelling evidence currently exists that the computational effort of these methods is warranted.

Many algorithms have been used effectively with DNA microarray data for predicting of a binary outcome, e.g. response versus non-response. Dudoit et al. {Dudoit, 2002 #42} compared several algorithms using several publicly available data sets. A linear discriminant is a function

$$l(\underline{x}) = \sum_{i \in F} w_i x_i \qquad (1)$$

where x_i denotes the logarithm of the expression measurement for the i'th gene, w_i is the weight given to that gene, and the summation is over the set F of features (genes) selected for inclusion in the class predictor. For a two-class problem, there is a threshold value d, and a sample with expression profile defined by a vector $\underline{x}$ of values is predicted to be in class 1 or class 2 depending on whether $l(\underline{x})$ as computed from equation (1) is less than the threshold d or greater than d respectively.

Many types of classifiers are based on linear discriminants of the form shown in (1). They differ with regard to how the weights are determined. The oldest form of linear discriminant is Fisher's linear discriminant. To compute the weights for the Fisher linear discriminant, one must estimate the correlation between all pairs of genes that were selected in the feature selection step. The study by Dudoit et al. indicated that Fisher's linear discriminant did not perform well unless the number of selected genes was small relative to the number of samples. The reason is that in other cases there are too many correlations to estimate and the method tends to be un-stable and over-fit the data.

Diagonal linear discriminant analysis is a special case of Fisher linear discriminant analysis in which the correlation among genes is ignored. By ignoring such correlations, one avoids having to estimate many parameters, and obtains a method which performs better when the number of samples is small. Golub's weighted voting method {Golub, 1999 #12} and the Compound Covariate Predictor of Radmacher et al. {Radmacher, 2002 #86} are similar to diagonal linear discriminant analysis and tend to perform very well when the number of samples is small. They compute the weights based on the univariate prediction strength of individual genes and ignore correlations among the genes.

Support vector machines are very popular in the machine learning literature. Although they sound very exotic, linear kernel support vector machines do class prediction using a predictor of the form of equation (1). The weights are determined by optimizing a mis-classification rate criterion, however, instead of a least-squares criterion as in linear discriminant analysis (Ramaswamy et al. {Ramaswamy, 2001 #92}). Although there are more complex forms of support vector machines, they appear to be inferior to linear kernel SVM's for class prediction with large numbers of genes {Ben-Dor, 2000 #91}.

In the study of Dudoit et al. {Dudoit, 2002 #42}, the simplest methods, diagonal linear discriminant analysis, and nearest neighbor classification, performed as well or better than the more complex methods. Nearest neighbor classification is defined as follows. It depends on a feature set F of genes selected to be useful for discriminating the classes. It also depends upon a distance function $d(\underline{x},\underline{y})$ which measures the distance between the expression profiles $\underline{x}$ and $\underline{y}$ of two samples. The distance function utilizes only the genes in the selected set of features F. To classify a sample with expression profile $\underline{y}$, compute $d(\underline{x},\underline{y})$ for each sample $\underline{x}$ in the training set. The predicted class of $\underline{y}$ is the class of the sample in the training set which is closest to $\underline{y}$ with regard to the distance function d. A variant of nearest neighbor classification is k-nearest neighbor classification. For example with 3-nearest neighbor classification, you find the three samples in the training set which are closest to the sample $\underline{y}$. The class which is most represented among these three samples is the predicted class for $\underline{y}$. Tibshirani et al. ( ) developed a variant called shrunken centroid classification that combines the gene selection and nearest centroid classification components.

Dudoit et al. also studied some more complex methods such a Classification Trees and aggregated classification trees. These methods did not appear to perform any better than diagonal linear discriminant analysis or nearest neighbor classification. Ben-Dor et al. {Ben-Dor, 2000 #91} also compared several methods on several public datasets and found that nearest neighbor classification generally performed as well or better than more complex methods.

5. Developmental and Validation Studies

It is important to distinguish the studies which develop parmacogenomic classifiers from those which utilize such classifiers for targeting treatment selection or for evaluating the clinical utility of such classifiers. The vast majority of published prognostic marker studies are developmental. Developmental studies are often based on a convenience sample of patients for whom tissue is available but who are heterogeneous with regard to treatment and stage. Although there is a large literature on prognostic markers, few such factors are used in clinical practice. Prognostic markers are unlikely to be used unless they are therapeutically relevant and most developmental studies are not based on a cohort medically coherent enough to establish therapeutic relevance.

The patients included in a developmental study of a pharmacogenomic biomarker to be used in drug development should be appropriate to enable identification of patients who are most likely to benefit from the new drug in a pivotal study. For example, suppose that the pivotal study involves advanced disease patients who have failed first line treatment and involves comparing survivals for patients receiving the new drug to survivals for patients receiving palliative care.   Patients from single arm phase II trials of the new drug could be used to develop a pharmacogenomic biomarker classifier of those patients likely to respond to the new drug. Dobbin and Simon ( ) have studied sample size considerations for developmental studies of predictive binary classifiers and have indicated that generally at least 20 cases in each class are required. Consequently, a phase II database containing at least 20 responders and 20 non-responders would be needed for the development of a pharmacogenomic classifier to be used in the subsequent pivotal trials. This may require a larger phase II developmental program than is conventional.

If the pivotal study involves comparison of outcome for patients receiving a standard regimen C versus those receiving C plus the new drug, then development of a gene expression based classifier is more complex. The classifier could be developed based on phase II studies of patients receiving C plus the new drug, but unless one also studied patients receiving C without the new drug one would not know whether prediction was drug specific or just reflected general prognostic features of the tumors.

It is possible to develop pharmacogenomic predictors of risk of tumor progression rather than tumor response. Even if the patients are receiving the investigational drug as a single agent, however, it may not be clear to what extent the predictor reflects drug effect rather than non-specific disease pace.

As indicated in the previous paragraphs, there are limitations to the adequacy of a conventional phase II database for empirically developing a pharmacogenomic classifier for use in a pivotal study. In many ways the best resource for developing a pharmacogenomic biomarker classifier for use in a pivotal trial is a collection of pre-treatment tumor specimens from patients enrolled in such a pivotal trial. For example, archived material from a “failed” pivotal trial of the drug can be used to develop a biomarker classifier of patients most likely to benefit from the drug compared to the control. The classifier can be based on the actual endpoint used in the clinical trial or upon an intermediate endpoint such as progression-free survival for which there may be more events available. By “failed” pivotal trial, I mean a trial for the same target population of patients which did not establish a statistically significant benefit for the drug for the randomized patients as a whole. The classifier developed based on archived material in a failed pivotal trial should be considered to have the same status as a classifier based on a phase II database. That is, the classifier should be used to design a new pivotal trial that establishes the clinical benefit of the drug in a prospectively specified subset of patients. Using the same pivotal trial to develop a pharmacogenomic classifier and to test treatment effects in subsets determined by the classifier is generally not valid. Freidlin and Simon ( ) have shown how one pivotal trial can be potentially used for both purposes, however, if the set of patients used to develop the classifier is kept distinct from the set of patients used to evaluate treatment benefit. Generally, however, the studies should be kept separate. Developmental studies are exploratory, though they should result in completely specified binary classifiers. Studies on which claims of drug benefit are based should be non-exploratory, but should instead test prospectively defined hypotheses about treatment effect in a pre-defined patient population.

6. Estimates of Predictive Accuracy in Developmental Studies

Developmental studies are analogous to phase 2 clinical trials. They should include an indication of whether the pharmacogenomic classifier is promising and worthy of phase 3 evaluation. There are special problems in evaluating whether classifiers based on high dimensional genomic or proteomic assays are promising however. The difficulty derives from the fact that the number of candidate features available for use in the classifier is much larger than the number of cases available for analysis. In such situations, it is always possible to find classifiers that accurately classify the data on which they were developed even if there is no relationship between expression of any of the genes and outcome {Radmacher, 2002 #15}. Consequently, even in developmental studies, some kind of validation on data not used for developing the model is necessary. This “internal validation” is usually accomplished either by splitting the data into two portions, one used for training the model and the other for testing the model, or some form of cross-validation based on repeated model development and testing on random data partitions. This internal validation should not, however, be confused with external truly independent validation of the classifier.

The most straightforward method of estimating the prediction accuracy is the *split-sample* method of partitioning the set of samples into a training set and a test set. Rosenwald et al. {Rosenwald, 2002 #14} used this approach successfully in their international study of prognostic prediction for large B cell lymphoma. They used two thirds of their samples as a training set. Multiple kinds of predictors were studied on the training set. When the collaborators of that study agreed on a single fully specified prediction model, they accessed the test set for the first time. On the test set there was no adjustment of the model or fitting of parameters. They merely used the samples in the test set to evaluate the predictions of the model that was completely specified using only the training data. In addition to estimating the overall error rate on the test set, one can also estimate other important operating characteristics of the test such as sensitivity, specificity, positive and negative predictive values.

The split-sample method is often used with so few samples in the test set, however, that the validation is almost meaningless. One can evaluate the adequacy of the size of the test set by computing the statistical significance of the classification error rate on the test set or by computing a confidence interval for the test set error rate. Since the test set is separate from the training set, the number of errors on the test set has a binomial distribution.

Michiels et al. {Michiels #181} suggested that multiple training-test partitions be used, rather than just one. The split sample approach is mostly useful, however, when one does not have a well defined algorithm for developing the classifier. When there is a single training set-test set partition, one can perform numerous unplanned analyses on the training set to develop a classifier and then test that classifier on the test set. With multiple training-test partitions however, that type of flexible approach to model development cannot be used. If one has an algorithm for classifier development, it is generally better to use one of the cross-validation or bootstrap resampling approaches to estimating error rate (see below) because the split sample approach does not provide as efficient a use of the available data {Molinaro, 2005 #182}.

Cross-validation is an alternative to the split sample method of estimating prediction accuracy{Radmacher, 2002 #15}. Molinaro et al. describe and evaluate many variants of cross-validation and bootstrap re-sampling for classification problems where the number of candidate predictors vastly exceeds the number of cases.{Molinaro, 2005 #182} The cross-validated prediction error is an estimate of the prediction error associated with application of the algorithm for model building to the entire dataset.

A commonly used invalid estimate is called the *re-substitution* estimate. You use all the samples to develop a model. Then you predict the class of each sample using that model. The predicted class labels are compared to the true class labels and the errors are totaled. It is well-known that the re-substitution estimate of error is highly biased for small data sets and the simulation of Simon et al.{Simon, 2003 #12} confirmed that, with a 98.2 % of the simulated data sets resulting in zero misclassifications even when no true underlying difference existed between the two groups.

Simon et al.{Simon, 2003 #12} also showed that cross-validating the prediction rule after selection of differentially expressed genes from the full data set does little to correct the bias of the re-substitution estimator: 90.2 % of simulated data sets with no true relationship between expression data and class still result in zero misclassifications. When feature selection was also re-done in each cross-validated training set, however, appropriate estimates of mis-classification error were obtained; the median estimated misclassion rate was approximately 50%.

7. Use of Pharmacogenomic Classifiers in New Drug Development

With a pharmacogenomic classifier for predicting which patients are likely to benefit from an *available* treatment regimen, the emphasis should be on validation of the clinical utility of using the classifier. With an experimental therapy, however, the emphasis should be on demonstrating effectiveness of the drug in a population identified by the classifier. Simon and Maitournam {Simon, 2004 #160; Simon, 2006 #278} demonstrated that use of a genomic classifier for focusing a clinical trial in this manner can result in a dramatic reduction in required sample size, depending on the sensitivity and specificity of the classifier for identifying such patients. Not only can such targeting provide a huge improvement in efficiency in phase III development, it also provides an increased therapeutic ratio of benefit to toxicity and results in a greater proportion of treated patients who benefit.

Simon and Maitournam consider use of the Targeted Design shown in Figure 1. During pre-clinical and phase I/II clinical development one identifies a fully specified classifier of which patients have a high probability of responding to the experimental drug. That classifier is then used to select patients for phase III trial. This is a form of enrichment design. Table 1 shows the number of events required in order to have 80% statistical power for comparing exponential survival times using the design of Figure 1 if the treatment results in a halving of the hazard in the patients selected for study using the classifier. The number of events shown in Table 1 is compared to the number of events required in a standard clinical trial if the classifier is not used to select patients for randomization (Table 2). The table assumes that the treatment is not effective for the classifier negative patients. More extensive results on relative efficiency of the targeted and untargeted designs are described by Simon and Maitournam {Maitournam, 2005 #207; Simon, 2004 #160}.

For many molecularly targeted drugs, however, the appropriate assay for selecting patients is not known and development of a classifier based on comparing expression profiles for phase II responders versus phase II non-responders may be the best approach. In such instances, one may not have sufficient confidence in the genomic classifier developed in phase II to use it for excluding patients in phase III trials as in Figure 1. It may be better in this case to accept all conventionally eligible patients, and use the classifier in the pre-defined analysis plan.

Figure 2 shows the Marker by Treatment Interaction Design discussed by Sargent et al.{Sargent, 2005 #202} and by Pusztai and Hess{Pusztai, 2004 #203}. Both marker positive and marker negative patients are randomized to the experimental treatment or control. The analysis plan either calls for separate evaluation of the treatment difference in the two marker strata or for testing the hypothesis that the treatment effect is the same in both marker strata. When this design is used for development of an experimental drug, an appropriate analysis plan might be to utilize a preliminary test of interaction; if the interaction is not significant at a pre-specified level, then the experimental treatment is compared to the control overall. If the interaction is significant, then the treatment is compared to the control within the two strata determined by the marker. The sample size planning for such a trial and determination of the appropriate significance level for the preliminary interaction test require further study.

Freidlin and Simon{Freidlin, 2005 #208} proposed an alternative analysis plan for the design of Figure 2. They suggested that the overall null hypothesis for all randomized patients is tested at the 0.04 significance level. A portion, e.g. 0.01, of the usual 5 percent false positive rate is reserved for testing the new treatment in the subset predicted by the classifier to be responsive. The analysis starts with a test of the overall null hypothesis, without a preliminary test of interaction. If the overall null hypothesis is rejected, then one concludes that the treatment is effective for the randomized population as a whole and that the classifier is not needed. If the overall null hypothesis is not rejected at the 0.04 level, then a single subset analysis is conducted; comparing the experimental treatment to the control in the subset of patients predicted by the classifier as being most likely to be responsive to the new treatment. If the null hypothesis is rejected, then the treatment is considered effective for the classifier determined subset. This analysis strategy provides sponsors an incentive for developing genomic classifiers for targeting therapy in a manner that does not unduly deprive them of the possibility of broad labeling indications when justified by the data.

8. Conclusions

Physicians need improved tools for selecting treatments for individual patients. The genomic technologies available today are sufficient to develop such tools. There is not broad understanding of the steps needed to translate research findings of correlations between gene expression and prognosis into robust diagnostics validated to be of clinical utility. This paper has attempted to identify some of the major steps needed for such translation.

Acknowledgements

Thanks to Dr. Wenyu Jiang for the computing of Tables 1 and 2.

Figure Legends

Figure 1. Targeted clinical trial design for evaluating a new experimental therapy. A biomarker classifier is developed for identifying those patients most likely to respond to the new treatment (E). Only those patients are randomized to E versus the control treatment. The patients predicted less likely to respond (marker negative) are off study. The targeted design is most useful in cases where the biomarker classifier has a strong biological rationale for identifying responsive patients and where it may not be ethically advisable to expose marker negative patients to the new treatment.

Figure 2. Stratified analysis design for evaluating a new experimental treatment (E) relative to a control (C). The status of a biomarker based classifier of the likelihood of responding to E is utilized in a prospectively specified analysis plan. The biomarker classifier is not just used for stratifying the randomization. Alternative analysis plans are described in the text.

Hazard Ratio for Marker + Patients	Number of Events Required
0.5	74
0.67	200

Table 1: Approximate number of events required for 80% power with 5% two-sided log-rank test for comparing arms of design shown in Figure 3. Only marker + patients are randomized. Treatment hazard ratio for marker + patients is shown in first column. Time-to-event distributions are exponential and all patients are followed to failure.

Proportion of Patients Marker +	Approximate Number of Events Required
20%	5200
33%	1878
50%	820

Table 2: Approximate number of events required for 80% power with 5% two-sided log-rank test for comparing arms of design shown in Figure 1. Randomized arms are mixtures of marker – and marker + patients. Hazard ratio for marker – patients is 1 for the two treatment groups and 0.67 for marker + patients. All patients are followed to failure.

Figure 1

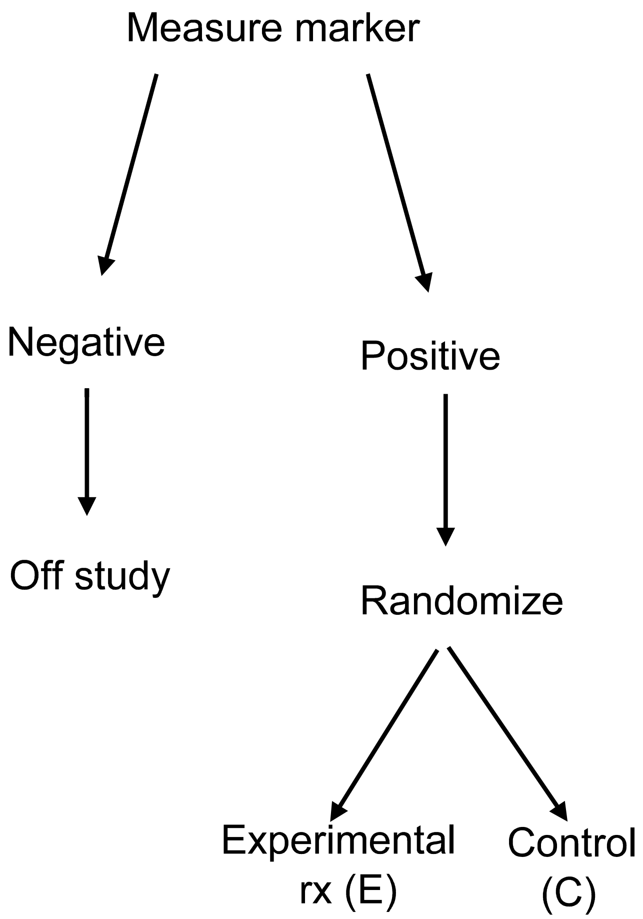
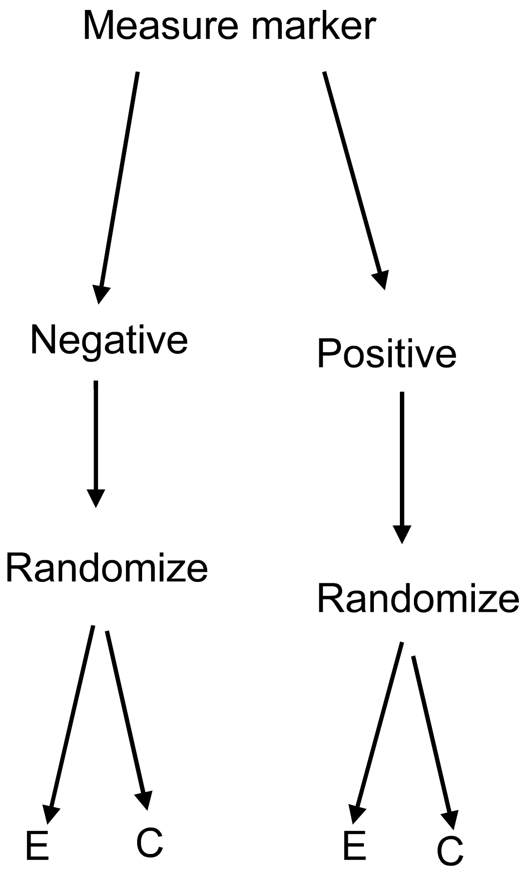

Measure marker
Negative
Positive
Off study
Randomize
Experimental rx (E)
Control (C)

Figure 2


Measure marker
Negative
Positive
Randomize
Randomize
E
C
E
C