Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 308–320
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Development and validation of biomarker classifiers for treatment selection

Richard Simon*

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA

Available online 12 June 2007

Abstract

Many syndromes traditionally viewed as individual diseases are heterogeneous in molecular pathogenesis and treatment responsiveness. This often leads to the conduct of large clinical trials to identify small average treatment benefits for heterogeneous groups of patients. Drugs that demonstrate effectiveness in such trials may subsequently be used broadly, resulting in ineffective treatment of many 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 spite of the large literature on developing predictive biomarkers, there is considerable confusion about the development and validation of biomarker-based diagnostic classifiers for treatment selection. In this paper we attempt to clarify some of these issues and to provide guidance on the design of clinical trials for evaluating the clinical utility and robustness of pharmacogenomic classifiers.
Published by Elsevier B.V

Keywords: Pharmacogenomics; Biomarker; Genomics; DNA microarray; Clinical trial design; Validation

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 (e.g. Bast and Hortobagyi, 2004; Johnson and Janne, 2005). 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.

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 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 messenger ribonucleic acid (mRNA) transcripts for a gene in a sample of disease tissue or the presence/absence status

* Tel.: +1 301 496 0975; fax: +1 301 402 0560.
E-mail address: rsimon@mail.nih.gov.

0378-3758/$ - see front matter Published by Elsevier B.V
doi:10.1016/j.jspi.2007.06.010

of a specified germline polymorphism or tumor mutation. A pharmacogenomic 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 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.

In this paper we will address some key issues in the development and validation of pharmacogenomic classifiers.

2. Developmental and validation studies

It is important to distinguish the studies which develop pharmacogenomic classifiers from those which evaluate the clinical utility of such classifiers. The vast majority of published prognostic marker studies are developmental and are not adequate for establishing the clinical utility and robustness of a classifier (Simon and Altman, 1994). 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. The studies are generally performed in an exploratory manner with no specified eligibility criteria, no primary endpoint or hypotheses and no defined analysis plan. The analysis often includes numerous analyses of different endpoints and patient subsets. Often there are multiple candidate biomarkers to evaluate, multiple ways of measuring and combining the candidate biomarkers. Such an informal approach is appropriate in a developmental study so long as one recognizes that the same study cannot be used to evaluate the clinical value of the resulting biomarkers or classifiers. The developmental study is exploratory and directed to hypothesis formation. The purpose of developmental studies should be to develop completely specified classifiers and completely specified hypotheses that can be tested in subsequent validation studies.

3. Development of multi-component classifiers

Four main components to developing a classifier are: (i) feature selection; (ii) selecting a prediction model; (iii) fitting the prediction model to training data; and (iv) estimating the prediction error that can be expected in future use of the model with independent data.

3.1. Feature selection

Feature selection is often important in developing an accurate classifier. It is well known from the theory of linear regression that including too many "noise variables" in the predictor reduces the accuracy of prediction. A noise variable is a variable that is not related to the thing being predicted. For microarray studies the number of noise variables may be orders of magnitude greater than the number of informative variables.

The most commonly used approach to feature selection is to identify 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 Mann–Whitney test for each gene. The log-ratios or log-intensity measurements are generally used as the basis of the statistical significance tests. The genes that are differentially expressed at a specified significance level are selected for inclusion in the class predictor. The stringency of the significance level controls the number of genes that are included in the model. If one wants a class predictor based on a small number of genes, the threshold significance level is made very small. Some statisticians fail to distinguish between "class comparison" problems, where the objective is to identify differentially expressed genes, and "class prediction" problems, where the objective is to do accurate prediction. Class comparison analyses are often appropriate when the objective is understanding biological mechanisms; e.g. what genes get expressed or repressed during wound healing of the kidney. Class prediction analyses are often appropriate for medical problems when the objective is predicting response to a specific treatment. Criteria such as false discovery rate are relevant for class comparison problems because it is useful to know what proportion of the genes reported as differentially expressed among the conditions represent false positives. For class prediction problems, however, the