Genomic biomarkers for a binary clinical outcome in early drug development microarray experiments

Peer-reviewed author version

VAN SANDEN, Suzy; SHKEDY, Ziv; BURZYKOWSKI, Tomasz; Gohlmann, Hinrich W. H.; TALLOEN, Willem & BIJNENS, Luc (2012) Genomic biomarkers for a binary clinical outcome in early drug development microarray experiments. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 22 (1), p. 72-92.

DOI: 10.1080/10543406.2010.504906
Handle: http://hdl.handle.net/1942/13629

Genomic Biomarkers for a Binary Clinical Outcome in Early Drug Development Microarray Experiments

Suzy Van Sanden[1]*, Ziv Shkedy[1], Tomasz Burzykowski[1],

Hinrich W.H. Göhlmann[2], Willem Talloen[2], Luc Bijnens[2]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics,

Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium

[2]Johnson & Johnson, PRD, Turnhoutseweg 30, B-2340 Beerse, Belgium

Abstract

In this paper, we discuss methods to select three different types of genes (treatment related, response related, or both) and investigate if they can serve as biomarkers for a binary outcome variable. We consider an extension of the joint model introduced by Lin *et al.* (2010) and Tilahun *et al.* (2010) for a continuous response. As the model has certain drawbacks in a binary setting, we also present a way to use classical selection methods to identify subgroups of genes, which are treatment and/or response related. We evaluate their potential to serve as biomarkers by applying DLDA to predict the response level.

Keyword: Biomarkers; BW-ratio; Categorical Data; Joint modeling; Microarrays.

Running Head: Genomic Biomarkers in Microarray Experiments

*Corresponding author. Tel: +32-11-26 82 81; Fax: +32-11-26 82 99; Email address: suzy.vansanden@uhasselt.be

1 Introduction

A biomarker (biological marker) can be defined as a physical sign or laboratory measurement that serves as an indicator for biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Lesko and Atkinson, 2001; Biomarkers Definitions Working Group, 2005). When the measurement in question is the expression of a gene, we refer to the gene as a genomic biomarker. In the remaining part of the paper, the term biomarker refers to a genomic biomarker.

In recent years, microarray experiments have been used as a tool to select genomic biomarkers. Microarrays provide information on the expression levels of a large number of genes at the same time. Therefore, a substantial pool of potential biomarkers can be evaluated simultaneously. The purpose of these biomarkers is not limited to prediction of the outcome. They can also give researchers insight into the biological processes associated with the response of interest if the biomarker is mechanistically relevant.

The use of gene-expression values to predict a specific response is well documented in literature. For example, Nguyen *et al.* (2002) included the expression levels of a number of genes as covariates in a Cox proportional hazard model to predict survival of patients with diffuse large B-cell lymphoma and breast cancer patients. The genes were selected by applying the partial least squares method, which maximizes the covariance between the response and a linear combination of the gene-expression data. In 2006, Tan *et al.* presented a similar approach. They used the partial least squares method for data reduction in the context of cytotoxicity experiments. Bair *et al.* (2006) published a paper regarding a supervised principal components analysis to predict survival. In their motivating example, they use gene-expression measurements from DNA microarrays as predictor variables.

All biomarker experiments mentioned above were conducted to study the association between gene-expression and the response, as illustrated in panel *a* of Figure 1, for the purpose of finding genes that can predict the response. In this paper, we consider experiments, in which a treatment was administered to the subjects prior to taking response and gene-expression measurements. We focus on the association between the treatment, the gene-expression measurements, and the outcome of primary interest. The setting is illustrated in panel *b* of Figure 1. This type of study mainly takes place in early drug development.

FIGURE 1

The goal of such an experiment is multifold. To get a better understanding of the biological processes taking place as a result of treatment, researchers wish to identify genes that are affected by it, or genes that are related to the phenotypic response to the treatment, such as toxicity or efficacy. One of the objectives of this type of experiment is thus to obtain lists of genes, of which the expression is affected by the treatment, related to the response, or both. Another goal is to evaluate if some of these genes can actually serve as biomarkers for the primary outcome variable. In other words, is the relation strong enough so that the genes in question can be used to predict the phenotypic response?

Lin *et al.* (2010) proposed a method of obtaining the aforementioned lists of genes for the case of a continuous outcome variable. In this paper, we focus on a binary outcome variable. To our knowledge, no methods for selection of genomic biomarkers for a binary outcome have been proposed in the literature. Figure 2 presents, using simulated data, an example of the main types of genes, which we can encounter for this particular setting. In this example, we consider a balanced design with 12 subjects, two treatments (1 and 2), and two levels of the response variable (0 and 1). Four (resp. two) of the six subjects that received treatment 1 (resp. 2) have response value 0.

The first panel of Figure 2 displays a gene capable of separating the two treatment groups (Type I). However, this type of gene is not able to make a distinction between the response levels. The opposite is true for the gene in the second panel (Type II). The gene represented in the third panel (Type III), can be used to predict the response level, as indicated by the full line. It can however also distinguish reasonably well between the two treatments (dotted line), with a misclassification rate of only two out of twelve. The gene in the last panel (Type IV) shows a good separation between the two treatment groups and is predictive for the response level after adjustment for the treatment effect.

FIGURE 2

When we focus on the search for finding genes that separate between the treatments, as a result of the relationship between the treatment and the response, we can obtain Type I genes, but also Type III or Type IV genes. We will term these genes potential *therapeutic biomarkers* (treatment-related genes). When we search for genes, for which the expression level is related to the response level (with or without adjustment for treatment effects), this can lead to Type II, III, or IV genes. They are called potential *prognostic biomarkers* (response-related genes). Search methods that focus on finding genes with both properties will deliver a third type of potential biomarkers, which are both therapeutic and prognostic. The results of the different searches will provide an answer to the first research question presented above, but not yet to the second one. Potential biomarkers will be evaluated by their ability to classify the subjects with respect to the levels of the response variable. A gene of either of the three types of potential biomarkers (therapeutic, prognostic or both) will be used as a biomarker for the clinical outcome variable only if it is also predictive for that outcome.

There are several methods to search for the potential biomarkers. Lin *et al.* (2010) have proposed

to use a joint model for gene-expression data and a continuous response variable. Tilahun *et al.* (2010) have used a similar approach to jointly model gene and metabolite expression. We can extend this approach to the binary case. However, while the approach seems to work well in a continuous setting, we will show that it has some drawbacks in case of a binary outcome. We therefore also investigate the use of classical selection methods for handling categorical data for discovering therapeutic and/or prognostic biomarkers.

The paper is organized as follows. In Section 2, two case-studies are introduced. We propose the methodology for finding biomarkers for the primary response variable in Section 3. The methods are applied to the case-studies, and the results are presented in Section 4. A short discussion and conclusions are presented in Sections 5.

2 Data

The data analyzed in this paper consist of outcomes from two randomized pre-clinical experiments: a behavioral study and a toxicology study. For each subject in both experiments, information is available about a treatment group, a clinical endpoint, and gene-expression. The aim of the analysis is to identify treatment- and/or response-related genes, and to evaluate if they can be used as genomic biomarkers, i.e., can be used to predict the clinical outcome. In what follows, we describe the two experiments in more detail.

2.1 Behavioral Study

The behavioral study is an experiment concerning compulsive checking disorder (Szechtman *et al.*, 2001). The disorder is induced by treating the animals with a chemical compound.

Twenty-four rats were randomized equally into two groups. The first group received the active compound (T), while the second was given a solvent (P). After receiving treatment, the rats had to complete an open field test (Szechtman *et al.*, 2001). The data indicated how often a rat went back to its home base in the open field. The home base was defined as the area where the animal spent the longest cumulative time. Animals showing the signs of the disorder (meaning that the compound has successfully induced the symptoms, characteristic of the disorder) were characterized by displaying, for example, an increased frequency of visits to the home base.

The clinical outcome of the experiment is a binary variable. A rat is considered to be a responder (visit=1) when the total number of visits the rats made to the home base lies above the median number of visits to the home base, and a non-responder otherwise (visit=0). Table 1 displays the number of responders and non-responders for both treatment groups. The data indicate that the level of compulsive checking substantially differs between rats that were treated with the solvent as compared to those treated with the compound.

TABLE 1

After completion of the experiment, a sample was taken from the thalamus part of the brain of the rats and used to obtain microarray measurements for 5644 genes. The data, from the Affymetrix Rat Genome U34A arrays, were summarized using the Affymetrix microarray suite software (MAS) Version 5.0, and normalized using quantile normalization.

2.2 Toxicology Study

Kidney vacuolation was observed in Sprague-Dawley rats treated with different compounds. The purpose of the toxicology study was to examine changes in gene-expression in the kidney, following treatment for 28 days, to gain information relevant to the underlying mechanism of the kidney

vacuolation and to the assessment of the toxicological consequences of the vacuolation.

In the study, 100 rats were randomized equally to three treatment and one control groups. The response variable, toxicity, consisted of four different levels, indicative of none to high toxicity. For 38 animals, about 10 per treatment group, there were also data available from a microarray containing the expression values of approximately 31,000 genes. An overview of the treatment and toxicity values for the animals with microarray measurements is displayed in Table 2.

The gene-expression measurements were obtained by using Affymetrix GeneChip Rat Genome 230 2.0 arrays. The collected data were summarized using the MAS 5.0 software. Data from all arrays were made comparable by using quantile normalization.

TABLE 2

To obtain a dichotomized version of the toxicity variable, the two first levels are combined, as well as the two last levels. The newly created response variable contains only two levels, corresponding to a low and a high toxicity. A problem with the dataset is the sparseness, which is most problematic in the control and in the second treatment groups. Therefore, parts of the analysis presented in the paper will be limited to the comparison of the first and third treatment group only.

3 Methodology

In this section, we describe two approaches to find the potential biomarkers, of the types described in Section 1, for a binary response. The first involves joint modeling of the continuous gene-expression data and a binary response variable. The model is a variant of the one proposed by Renard *et al.* (2002) for the evaluation of surrogate endpoints in clinical trials, and corresponds to the model used by Lin *et al.* (2010) for a continuous response. The second approach is based

on biomarker selection using the ratio of the between and within-group sums of squares. In what follows, the methods, together with a discussion of their benefits and disadvantages, are presented. Details about the implementation of the methods are given in the Appendix.

3.1 The Joint Modeling Approach

Consider an experiment, in which n subjects received a particular treatment. In the experiment, the value for a binary response variable (Y) of interest is measured for each subject together with the gene-expression (X) for m genes. Let Z_j indicate the treatment group subject j belongs to, X_{ij} denote the gene-expression for gene i of subject j, and Y_j^* be the level of a latent continuous variable underlying the response Y_j. It is assumed that Y_j^* is normally distributed. Note that the observed outcome Y_j is considered to be a binary variable, defined as follows:

$$Y_j = \begin{cases} 1 & Y_j^* > 0, \\ 0 & Y_j^* \leq 0. \end{cases} \tag{1}$$

We assume the following gene-specific joint model for the latent outcome variable Y_j^* and the gene-expression X_{ij}:

$$\begin{cases} X_{ij} = \mu_i + \alpha_i \; Z_j + \varepsilon_{X_{ij}}, \\ Y_j^* = \mu_Y + \beta \; Z_j + \varepsilon_{Y_j}, \end{cases} \tag{2}$$

where

$$\begin{pmatrix} \varepsilon_{X_{ij}} \\ \varepsilon_{Y_j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{X_i}^2 & \sigma_{X_iY} \\ \sigma_{X_iY} & 1 \end{pmatrix} \right]. \tag{3}$$

The resulting probit model formulation for the observed binary outcome Y_j and the gene-expression X_{ij} is

$$\begin{cases} X_{ij} \sim N(\mu_i + \alpha_i \; Z_j, \sigma_{X_i}^2), \\ Y_j \sim B(p_j), \\ \Phi^{-1}(p_j) = \mu_Y + \beta \; Z_j, \end{cases} \tag{4}$$

where $B(p_j)$ denotes the Bernoulli distribution with the success probability $p_j = P(Y_j = 1)$, Φ is the standard normal cumulative distribution function.

Based upon this model, we can first test the null hypothesis $H_0 : \beta = 0$ to see if the response is influenced by the treatment. For a gene to be considered a potential therapeutic biomarker, the gene-specific null hypothesis $H_0 : \alpha_i = 0$, to test if the gene is differentially expressed between the treatments, has to be rejected versus the alternative hypothesis $H_1 : \alpha_i \neq 0$. A gene can serve as a prognostic biomarker if the null hypothesis $H_0 : \rho_i = 0$ is rejected versus the alternative hypothesis $H_1 : \rho_i \neq 0$, where $\rho_i = \sigma_{X_i Y}/\sigma_{X_i}$. Note that ρ_i is the so-called adjusted association proposed by Buyse *et al.* (1998) as a measure for surrogacy in the context of randomized clinical trials. Genes, for which the two above mentioned null hypotheses are rejected, are candidates for biomarkers that are both therapeutic and prognostic.

The interpretation of ρ_i is clear in a continuous setting. When it is different from zero, a linear association exists between the response and gene-expression values after adjustment for possible treatment effects. The probit model formulation for a binary and continuous response in (4), described by Renard *et al.* (2002), implies that the correlation coefficient can be interpreted as the correlation between the underlying latent variable Y_j^* and the gene-expression data X_{ij}, after correction for treatment.

The differentially expressed genes, discovered by hypothesis testing, do not always form the best subset for classification. Hypothesis testing reduces the risk of chance findings, which is important when we are interested in the individual genes. However, in a biomarker experiment with a binary response variable, we are looking for a group of genes, biomarkers, which are used together to build a good classifier. If we use hypothesis testing, too many genes may be filtered out, leading to loss of classification information, while the group of retained genes may be too small to reduce noise

(Amaratunga *et al.*, 2004). Instead, we can also rank the genes according to the value of the test statistic and select the top r genes to serve as biomarkers. The genes can be ranked based on the statistic used to test $H_0 : \alpha_i = 0$ (for potential therapeutic biomarkers) or $H_0 : \rho_i = 0$ (for potential prognostic biomarkers). A third possibility is to select genes ranked high on both lists (potential therapeutic/prognostic). For each gene, a new rank is calculated by taking the sum of the ranks from the therapeutic and prognostic gene lists.

3.2 The BW-criterion

Though the joint modeling framework can be used in a binary setting, it is not necessarily the best approach for the latter situation. We therefore consider another approach, using a well-known gene selection method that is often applied prior to classification, namely, the BW-ratio (Dudoit *et al.*, 2002). This ranking-based approach focuses on finding genes with the largest BW-ratio. The latter stands for the ratio of the between- and within-group sum of squares of gene-expression levels (Dudoit *et al.*, 2002). In a two-group setting, the BW-ratio reduces to the same statistic as the t-test. Genes are ranked according to the BW-ratio and the top r genes are retained as the genes of interest and candidate biomarkers.

If the BW-ratio is applied to the gene-expression levels and we consider the treatment variable Z_j as the group indicator, genes are selected for their association with treatment. We thus obtain potential therapeutic biomarkers. Potential prognostic biomarkers are obtained by considering the binary response variable Y_j as group indicator. The biomarkers are thus selected for their ability to separate between the levels of the clinical outcome variable. Note that with this approach, we cannot detect genes of type IV (see Figure 2). With the joint modeling approach, it is possible to detect also these genes, as the latter methods can correct for the treatment effect. When biomarkers are ranked high on both lists, they are associated with both treatment and response. Hence, they

are candidate therapeutic/prognostic biomarkers. For each gene, a new rank is calculated by taking the sum of the ranks from the therapeutic and prognostic gene lists. Note that the thus obtained lists of ranked genes will be denoted through the rest of the article as BW_{Treat}, BW_{Resp}, and $\text{BW}_{Treat/Resp}$ respectively.

3.3 Evaluation of Potential Biomarkers

For the evaluation of the biomarkers and the selection of the number of biomarkers to use (r), one possible approach is to look at the misclassification rate (MCR), when classifying subjects to the levels of the response variable based on the selected genes. For this purpose, we will use DLDA, diagonal linear discriminant analysis. Both Dudoit *et al.* (2002) and Van Sanden *et al.* (2007) discuss the performance of DLDA in the microarray setting and show that it outperforms many other classification methods (such as classification trees, SVM, etc.). To obtain a realistic estimation of the misclassification rate, we use double k-fold cross-validation (CV). The method works as follows. We split the arrays randomly in k subsets and leave them out one by one. Using the remaining data, we select the biomarkers and build the classifier. The latter is then used to classify the observations that were left out. After doing this for all subsets, we take the average misclassification rate. If k is equal to the number of arrays, the method is called leave-one-out cross-validation (LOOCV). If k is smaller than the number of arrays, the k-fold cross-validation procedure is repeated 100 times. Each time the data are split differently. The mean and standard error for the misclassification rate are then calculated.

While cross-validation is a well-known method for unbiased estimation of the error rate for classification procedures, there has been some discussion with regard to its performance when dealing with small sample sizes (Fu *et al.*, 2005). Under the latter condition, which is common for microarray data, the variability tends to be large. Fu *et al.* (2005) have proposed the bootstrap

cross-validation method (BCV) for estimating the error rate in small sample size settings. The procedure consists of drawing bootstrap samples and performing CV for each of them. Fu *et al.* (2005) have shown that this method outperforms many others, including cross-validation and the .632 and 0.632+ bootstrap methods (Efron and Tibshirani, 1993, 1997) in the aforementioned setting. According to the authors, the reason the methods works so well is that the prediction by CV performs better in bootstrap samples than in the original sample when its size is small.

The bootstrap samples in the BCV method are chosen such that they contain at least three distinct observations in every class. The average misclassification rate and standard error are then determined. We perform the bootstrap cross-validation method with 100 bootstrap samples and we apply LOOCV to each of them. Note that k-fold ($k <$ number of samples) and bootstrap cross-validation are too computationally intensive when combined with the joint modeling approach. They are therefore not performed.

4 Results

Both biomarker selection methods are applied to the case studies described in Section 2. We first present a straightforward analysis of the behavioral study. Secondly, we demonstrate the application of the methods to the more complicated toxicology study.

4.1 Behavioral Study

Before including the gene-expression data in the analysis, the effect of treatment on the binary clinical outcome variable is investigated using a Fisher's exact test. A significant treatment effect is found (p=0.0033).

4.1.1 The Joint Modeling Approach

After fitting the joint model, we test the hypotheses to find the potential therapeutic and/or prognostic markers at the 5% significance level. As the tests are performed for a large number of genes, correcting for multiple testing is necessary. The null hypothesis of no treatment effect on gene-expression, $H_0 : \alpha_i = 0$, is rejected for nine genes when using a t-test with the Bonferroni multiple testing procedure. The null hypothesis for the adjusted association, $H_0 : \rho_i = 0$, can not be rejected for any of the genes (likelihood ratio test). When using the FDR adjustment by applying the Benjamini and Hochberg procedure (BH, Benjamini and Hochberg (1995)), we find 10 therapeutic and no prognostic biomarkers. Based on either the Bonferroni or BH-procedure-selected therapeutic biomarkers, we obtain a misclassification rate for DLDA of 12.5%.

Instead of using hypothesis tests to select the biomarkers, we can also rank the genes according to the value of the test statistic and select the top r=2, 5, 10, 20, 40, 200, 500 or 1000 genes to serve as biomarkers. Note that the nine highest ranked therapeutic genes are those found by hypothesis testing using the Bonferroni approach.

Figure 3 and the first part of Table 3 show the results of classification using the top r selected therapeutic and/or prognostic biomarkers. Even without cross-validation (Figure 3, panel a), the best prediction is not always obtained with the prognostic biomarkers. When $r = 5$ genes are selected from the list of therapeutic or therapeutic/prognostic biomarkers, a lower misclassification rate is obtained as compared to selecting the top five prognostic biomarkers. This indicates that the association between the latent continuous response variable and gene-expression might not be useful to predict the level of the binary response variable. For leave-one-out cross-validation (Figure 3, panel b), an advantage of using therapeutic instead of prognostic biomarkers is clearly visible. It appears that genes, selected for their association with the treatment effect, can lead to a better

prediction of the response level as compared to the genes selected based on their association with the response. Note that the misclassification rate increases with the increasing number of selected biomarkers. This may indicate the presence of "noisy" genes that, when included, interfere with the classification.

FIGURE 3

TABLE 3

4.1.2 The BW-criterion

When we use the BW-criterion with the response variable *visit* as group indicator, we obtain a ranked list of biomarkers indicated by BW_{Resp} (prognostic). If treatment is used as group indicator, we denote the list of biomarkers as BW_{Treat} (therapeutic). The list of biomarkers ranked highly on both of them, is referred to as $BW_{Treat/Resp}$ (therapeutic/prognostic).

A strong correspondence exists between therapeutic markers selected by the joint model and the BW-approach. All the candidate therapeutic biomarkers selected by using the Bonferroni approach, and nine of the 10 candidate biomarkers selected by the BH procedure, are ranked highest on the BW_{Treat} list. A similar correspondence does not exist for the prognostic biomarkers.

The results of classification of the binary outcome variable obtained by using the complete dataset to build the classifier, and the results obtained with leave-one-out, 3-fold, and bootstrap cross-validation, are presented in the lower part of Table 3 and in Figure 4. Overall, the misclassification rate seems to be the lowest when using the therapeutic markers, with only a few exceptions, e.g., for $r > 10$ and no CV (Figure 4, panel *a*). For the 3-fold cross-validation procedure, some of the differences between the misclassification rates, obtained by using biomarkers selected from the

BW_{Resp}, BW_{Treat}, or $BW_{Treat/Resp}$ lists, are large. When comparing these differences with the standard error of the misclassification rates (see Table 3, section for 3-fold CV (BW-approach)), they even appear to be significant. On the other hand, the misclassification rate obtained with the bootstrap cross-validation procedure is only slightly different for the classifications based on the therapeutic, prognostic, or therapeutic/prognostic biomarkers (Figure 4, panel *d*).

FIGURE 4

4.2 Toxicology Study

For the toxicology study, the researcher's interest lies in finding biomarkers for toxicity, i.e., genes that are influenced by treatment and/or are correlated with toxicity. The therapeutic and/or prognostic biomarkers will give information about the effect of the treatment on gene-expression, about the relationship between the expression levels and toxicity measures, or about the way the treatment leads to toxicity. Furthermore, we can examine the ability of all three types of biomarkers to predict the toxicity levels.

An analysis similar to the one performed for the behavioral study is conducted for the toxicology study. However, the use of the joint modeling approach is restricted to the dichotomized version of the response variable, as the current software can only handle categorical data with two classes. Furthermore, as sparseness in the data leads to convergence problems, we only use the first and the third treatment group for the joint modeling approach. We can overcome this problem by using the BW-approach.

A Fisher's exact test indicates that there is no overall treatment effect on the dichotomized response variable ($p = 0.3034$, $n = 20$), unless we also include the animals, for which microarray data are not available ($p = 0.0003$, $n = 50$). There is thus not enough power to detect the difference

between the two treatments when using the reduced dataset. The small sample size of the dataset could also be a problem for other hypothesis tests performed during the analysis.

4.2.1 The Joint Modeling Approach

After fitting the joint model to the reduced dataset with the binary toxicity variable (low and high toxicity) and 10 rats from both the first and the third treatment group, we test the hypotheses to find the potential therapeutic and/or prognostic biomarkers at the 5% significance level. The null hypothesis $H_0 : \alpha_i = 0$ was rejected for 33 genes when using a t-test with the Bonferroni multiple testing procedure, while the null hypothesis for the adjusted association, $H_0 : \rho_i = 0$, could not be rejected for any of the genes (likelihood ratio test). Using the BH procedure, we find 705 therapeutic and no prognostic genes. Using the therapeutic biomarkers selected by either the Bonferroni or BH procedure, we obtain a misclassification rate for DLDA of 35%.

As explained in Section 3.1, hypothesis testing is not always the best option when selecting genes for performing classification. The genes are, therefore, also ranked according to the value of the test statistic. Figure 5 displays toxicity versus the gene-expression level for the highest-ranked gene on each of the lists. One therapeutic gene appears to be enough to separate the treatments based on gene-expression values (Figure 5, panel *a*). However, this gene is not appropriate for predicting toxicity. The prognostic biomarker is more suited for that purpose, but does not separate treatments (Figure 5, panel *b*). The therapeutic/prognostic biomarker separates the treatments to a certain degree and has some predictive power for toxicity (Figure 5, panel *c*). However, a combination of therapeutic and/or prognostic biomarkers will be required to predict toxicity with a suitable level of accuracy.

FIGURE 5

Figure 6 and the upper part of Table 4 display the results of classification of the samples to the toxicity levels, using the therapeutic and/or prognostic biomarkers. Without using cross-validation (Figure 6, panel *a*), the use of the prognostic biomarkers seems to yield the lowest misclassification rate, while the opposite is true for the therapeutic markers. This is to be expected, as the observations that are predicted were also used for selecting the therapeutic and prognostic biomarkers, and for building the classifier. From the cross-validation results, it appears that the prognostic biomarkers do not always lead to the lowest classification error (Figure 6, panel *b*). For example, when selecting less then 10 genes, the best prediction is obtained with the therapeutic or therapeutic/prognostic biomarkers. This is in contrast with the results obtained from the analysis without cross-validation (Figure 6, panel *a*). Note that in the cross-validation procedure, one particular sample, the only one belonging the "low toxicity - T3" cell of Table 2, could not be left out. Otherwise, there would be an empty cell in the dataset used for the selection of the biomarkers. However, when an empty cell occurs, the model building procedure does not converge. Therefore, we cannot use the joint modeling approach to select biomarkers based on the dataset without that particular sample.

FIGURE 6

TABLE 4

4.2.2 The BW-criterion

First, we discuss the analysis, in which the treatments of primary interest (T1 and T3) are included. Note that the response variable is binary, i.e., low and high levels of toxicity are considered. Ranking the genes according to the BW-ratio for the toxicity and/or treatment variable leads to the list of therapeutic and/or prognostic biomarkers. The lists are again indicated by BW_{Resp}, BW_{Treat}, and

$BW_{Treat/Resp}$ respectively.

In Figure 7, toxicity is plotted versus the gene-expression level for the highest ranked gene on each of the lists. Note that the therapeutic biomarker (Figure 7, panel *a*) is the same gene as obtained with the joint model (see Figure 5). Also for this study, a strong correspondence exists between this type of markers selected by the model and the BW-ratio. Thirty of the 33 candidate therapeutic biomarkers selected by using the Bonferroni procedure, and 677 of the 705 candidate biomarkers obtained by using the BH procedure, are ranked the highest on the BW_{Treat} list. A similar correspondence does not apply to the prognostic biomarkers. Furthermore, the prognostic gene displayed in Figure 7 (panel *b*) shows a slightly better separation between toxicity levels compared to that observed for the gene selected based on the joint model (see Figure 5, panel *b*).

FIGURE 7

Table 4 and Figure 8 display the misclassification rates for the different cross-validation methods. A similar conclusion can be drawn as for the modeling approach, irrespectively of whether LOOCV is used or not (Figure 8, panels *a* and *b*). The misclassification rate for the prognostic and therapeutic/prognostic biomarkers is, however, most of the times smaller, as compared to the modeling approach. Using the 3-fold cross-validation (Figure 8, panel *c*), there is only a slight difference visible between the three types of biomarkers. Taking into account the estimates of the standard error, we can conclude that the small observed deviations are not statistically significant (Table 4, section for 3-fold CV (BW-approach)). However, as we are only using 2/3 of the already small collection of samples for selection of biomarkers and building a classifier, this validation method may not be the most reliable. Bootstrap cross-validation shows that the best results are obtained with the prognostic markers. But again, the observed differences are not statistically significant.

FIGURE 8

As mentioned before, the BW-approach can be applied to the categorical toxicity variable with four levels and/or can include all treatment groups. Figure 9 displays the results for the analysis using the dichotomized toxicity response (panel a) and the categorical toxicity response (panel b). In both analysis, all four treatment groups were used. Note that the misclassification rate for the categorical response is higher than that for the binary response. This is most likely caused by the large number of empty cells in Table 2. For this reason, the bootstrap cross-validation method is not applied here. When the data are sparse, drawing bootstrap samples with enough distinct observations for every category is not possible.

A second pattern, which can be observed from panel b of Figure 9, is that none of the biomarkers (prognostic, therapeutic, therapeutic/prognostic) can be identified as the best biomarker for classification.

FIGURE 9

4.3 Stability of the Gene Selection Process

To get an understanding of why the treatment related genes can lead to a better prediction of the response level, we investigate the consistency of the lists of prognostic, therapeutic, and therapeutic/prognostic biomarkers. We focus on the BW-approach, as the latter generally leads to better results compared to the joint modeling approach. For each of the two studies, we create two new datasets by each time leaving out one observation at random. We then select the different types of biomarkers using the BW-approach and compare the thus obtained lists of genes between the two datasets. The ratio, the number of genes appearing in the top r of the gene lists for both datasets divided by r (r=2, 5, 10, 20, 40, 200, 500 or 1000), is calculated for the three types of biomarkers for both the behavioral and toxicology study. The whole process is repeated 100 times and the

average of the ratio is displayed in Figure 10. For the behavioral study, it is clear that, when the number of selected biomarkers is smaller than 20, more genes are found in common for the lists of therapeutic and therapeutic/prognostic biomarkers compared to the list of prognostic biomarkers. When r becomes larger, the difference disappears and the average ratio stabilizes around 0.8. The difference still exists for the toxicology study between the therapeutic and prognostic biomarkers, but is less outspoken for the prognostic versus the therapeutic/prognostic biomarkers.

FIGURE 10

5 Discussion

The modeling approach for the binary case, presented in this paper, is a direct extension of the joint model for gene-expression levels and a continuous response variable considered by Lin *et al.* (2010). Adapting the existing method seemed to be the most logical first step. However, we have also shown that it is not necessarily the best method for categorical (binary) data, as the technique has limitations. The joint model can only be applied to binary response variables. Furthermore, sparseness in the data can lead to problems with the convergence of the model. The issue has been observed in the toxicology study.

While hypothesis testing provides a clear decision rule to decide which genes can serve as biomarkers, it is often too restrictive. For instance, for the optimal choice of r (i.e. the number of biomarkers, leading to the lowest misclassification rate for a certain method/type of biomarker), a smaller misclassification rate can be obtained based on the prognostic and/or therapeutic biomarkers selected by the ranking-based approach compared to those found by hypothesis testing. On the other hand, the question remains how to choose the optimal number of biomarkers to use for classification.

The BW-criterion, which can be applied to categorical data with more than two categories, is introduced as an alternative to the modeling technique. There is a strong correspondence between the joint modeling and the BW-approach, with regard to the selection of therapeutic biomarkers. They lead to similar results in both the behavioral and toxicology study. For both methods, the ranking of the genes is obtained by using a form of a t-statistic (BW-ratio for two groups reduces to a t-statistic). However, for joint model, this test statistic is subject to the correlation between the gene-expression and the response. It is therefore expected that the aforementioned correspondence between the two methods depends on the size of this correlation. For the two studies considered in this article, we could not reject the null hypothesis that the correlation is zero for any of the genes. This indicates that the correlation, if it exists, is very small, and it thus explains the equivalence of the two methods regarding the therapeutic biomarkers.

For the joint modeling approach in a binary setting, the definition of a prognostic biomarkers is not straightforward. It is focused on a linear relationship between the gene-expression and a latent continuous outcome variable underlying the observed binary variable, after correction for treatment. We therefore consider an alternative definition of a prognostic biomarker for a binary response variable. We look at the ability of the gene to separate the samples between the levels of the response variable. For this, we use the BW-ratio as a criterion, with the response variable as the group indicator.

The sample sizes of the case studies made available to us by the pharmaceutical industry, are small. Due to the cost price of microarrays, this is a common problem in early drug development studies. As a results, we might lack the power to detect potential biomarkers through hypothesis testing. However, when biomarkers are selected using a ranking-based approach rather than determined by hypothesis testing, we can always select the top r genes that lead to the lowest misclassification rate. Furthermore, when computationally possible, we can apply bootstrap cross-

validation. According to Fu *et al.* (2005), the method performs well with samples of sizes as small as 16.

Another consequence of the small sample size of the toxicology study, is the high misclassification rate observed, especially when classifying the samples to four toxicity levels. While we have demonstrated that the method can, in principle, be applied to categorical response variables with more than two levels, we do not recommend it when the number of observations per cell is as small, as it is in the case of the toxicology study.

When the data are used for selecting biomarkers, building the classifier, and predicting the outcome, the lowest misclassification rate is often obtained for genes selected based on BW_{Resp}, while the highest misclassification rate usually occurs when using BW_{Treat}. This is to be expected, because BW_{Resp} selects genes that best classify the samples of that particular dataset to the levels of the response variable. It is therefore true in general. However, when using a more suitable method to estimate the misclassification rate, like a cross-validation approach, we have seen in the behavioral study that the best results are not always obtained when using BW_{Resp}. The genes with expression values associated with treatment, or both treatment and the response, often lead to a better prediction of the response level of a new sample. This occurrence can most likely be attributed to the following two observations. For the behavioral study, the response is strongly related to the treatment. Furthermore, we have found that the selection of treatment-related, or both treatment- and response-related genes is more stable compared to the selection of genes only related to the response. In studies that show these similar characteristics, we thus expect to see similar results.

For the toxicology study, the number of significant therapeutic biomarkers is much larger when using the BH multiple testing procedure, as compared to the Bonferroni method. This is not

unusual, since the latter controls the FWER, while the BH procedure only offers weak control over the FWER (Benjamini and Hochberg, 1995). However, the inclusion of the larger set of biomarkers in the classification procedure, selected using the BH method as compared to the Bonferroni method, has no effect on the estimated misclassification rate, which is in both cases 35%. This appears to be the lowest obtainable misclassification rate based on the potential therapeutic markers (see Figure 8, panel *a*). Thus, adding more biomarkers does not increase the information for classification.

Concluding Remarks

Typical biomarker experiments involve only gene-expression measurements and a response variable. We present an analysis for a more complicated situation, with also a treatment variable. The outcome in question is categorical (binary). Two distinct approaches are considered for biomarker selection. First, we extend the joint model used for a continuous response to the binary case. While the method works under certain conditions, it has several limitations in the binary setting and is computationally intensive. We therefore consider another approach, based on the BW-criterion. This method works for binary data, but can also be applied to a categorical outcome variable with more than two levels.

Both methods lead to the identification of subgroups of genes (potential biomarkers) related to treatment, response, or both. We are thus able to detect genes that can indicate treatment effect (therapeutic) and/or can discriminate between the response levels (prognostic). However, the potential biomarkers obtained with the BW-approach lead in general to a lower misclassification rate compared to those obtained with the joint modeling approach. Furthermore, using the information regarding the treatment that was administered, or combining the information available on both treatment and response, can sometimes increase the predictive power of the method and reduce the number of biomarkers necessary for a good prediction. In summary, the methods presented in the paper can form a tool for the pharmaceutical industry for biomarker detection and prediction

of response levels by using information from functional genomic experiments.

Acknowledgment

We gratefully acknowledge support of the IAP research network P6/03 of the Belgian Government (Belgian Science Policy).

References

Amaratunga, D., Cabrera, J. (2004) *Exploration and Analysis of DNA microarray and protein array data.* Hoboken, NJ: Wiley-Interscience-John Wiley and Sons, Inc.

Bair, E., Hastie, T., Paul, D., Tibshirani, R. (2006). Prediction by supervised principlal components. *The Journal of the American Statistical Association* 101, 119–137.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57(1), 289–300.

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* 69, 89-95.

Burzykowski, T., Molenberghs, G., Buyse, M. (2005). *The evaluation of surrogate endpoints.* New York: Springer.

Buyse, M., Molenberghs, G. (1998. The validation of surrogate endpoints in randomized experiments. *Biometrics* 54, 186–201.

Dudoit, S., Fridlyand, J., Speed, T.P. (2002). Comparison of discrimination methods for the clas-

sification of tumors using gene-expression data. *Journal of the American Statistical Association* 97(457), 77-87.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* 92, 548-560.

Fu, W.J., Carroll, R.J., Wang, S., (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* 21(9), 1979–1986.

Lesko, L.J., Atkinson, A.J. (2001). Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: Criteria, validation, strategies. *Annual Review of Pharmacology & Toxicology* 41, 347–366.

Lin, D., Shkedy, Z., Molenberghs, G., Goehlmann, H., Talloen, W., Bijnens, L. (2010) Selection and Evaluation of Gene-specific Biomarkers in Pre-clinical and Clinical Microarray Experiments. *Online Journal of Bioinformatics* (accepted for publication).

Molenberghs, G., Verbeke, G. (2005). *Discrete Longitudinal Data*. New York: Springer-Verlag.

Nguyen, D.V., Rocke, D.M. (2002). Parrial least square proportional hazard regression for aplication to DNA microarray survival data. *Bioinformatics* 18(12), 1625–1632.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M. (2002) Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical*, 44, 921–935.

SAS Institute, Inc. (2004) The GLIMMIX Procedure (Experimental). Cary, NC: SAS Institute, Inc.

Szechtman, H., Eckert, M.J., Tse, W.S. Boersma, J.T., Bonura, C.A., McClelland, J.Z., Culver,

K.E., Eilam, D. (2001). Compulsive checking behavior of quinpirole-sensitized rats as an animal model of Obsessive-Compulsive Disorder (OCD): form and control. *BMC Neuroscience* 2, Art.4.

Tan, Y., Shi, L., Hussen, S.M., Xu, J., Tong, W., Frazier, J.M., Wang, C. (2006). Integrating time course microarray gene-expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium. *Bioinformatics* 22(1), 77–87.

Tilahun, A., Lin, D., Shkedy, Z., Geys, H., Alonso, A., Peeters, P., Talloen, W., Drinkenburg, W., Ghlmann, H., Gorden, E., Bijnens, L., Molenberghs, G. (2010). Genomic Biomarkers for Depression: Feature-Specific and Joint Biomarkers. *Statistics in Biopharmaceutical Research*, ahead of print.

Van Sanden, S., Lin, D., Burzykowski, T. (2007). Performance of classification methods in a microarray setting: a simulation study. *Biocybernetics and Biomedical Engineering* 27(3), 15–28.

Appendix: Computational Issues

The Joint modeling Approach

Recent software developments have made it possible to jointly model two or more response variables, even if they do not have the same distributions. These features have been implemented, e.g., in the GLIMMIX procedure in SAS 9.1 (SAS Institute, 2004; Molenberghs *et al.*, 2005). The correlation between the measurements of the response variables is modeled directly by the specification of the covariance matrix Σ of the residuals, specified in equation (3). In GLIMMIX, this is done by the random statement discussed below.

The combination of normal and multinomial response variables has not been included in the

procedure. It is, however, possible to jointly model normal and binary data.

To use the GLIMMIX procedure for multiple responses, the data have to be transformed from the multivariate (responses in separate variables) to the univariate form (responses combined in one variable). An additional variable is needed to indicate the distribution of a particular response. The code to adjust the data is given below:

```
data dataset;
length dist $7;
set dataset;
response=resp;
dist = "Binary";
link = "PROBIT";
output;
response=geneexpres;
dist = "Normal";
link = "ID";
output;
run;
```

To fit the gene-specific joint model presented in equation (3), the following GLIMMIX statement is used:

```
*The full model;
proc glimmix data=dataset;
class dist treat animal;
```

```
model response(event=FIRST) = dist dist*treat/noint s
dist=byobs(dist);
random _residual_ / subject=animal type=chol;
run;
```

Note that the covariance matrix, given in equation (2), is specified using the random statement `random _residual_ / subject=animal type=chol;`

The result of the hypothesis test for a treatment effect on gene-expression can be found in the output. To perform the likelihood ratio test for the adjusted association, we also have to fit the reduced model, which assumes a covariance equal to zero between the response and the gene-expression. Hence, we adjust the random statement and specify a diagonal covariance matrix by using the option "type=un(1)". The likelihood ratio test, carried out for these two models, determines if the hypothesis, adjusted association equal to zero, can be rejected. The SAS code for the reduced model is given below:

```
*The reduced model;
proc glimmix data=dataset;
class dist treat animal;
model response(event=FIRST) = dist dist*treat/noint s
dist=byobs(dist);
random _residual_ / subject=animal type=un(1);
run;
```

The BW-criterion

The BW-approach and the classification procedure are implemented in R. Existing R functions can be used, namely *stat.bwss* from the `sma` package to obtain the BW-ratio and *stat.diag.da* from that same package to perform DLDA.

Tables legends

Table 1: Frequency table for the binary outcome in the behavioral study.

Table 2: Number of rats with different toxicity levels for the toxicology study.

Table 3: Mean misclassification rate (standard error) for classification to the dichotomized response variable of the behavioral study. The upper part of the table contains the results for therapeutic and/or prognostic biomarkers obtained by the joint modeling approach, while the lower part presents the results for BW_{Resp}, BW_{Treat} and $BW_{Treat/Resp}$. (CV: cross-validation, LOOCV: leave-one-out cross-validation, BCV: bootstrap cross-validation).

Table 4: Mean misclassification rate (standard error) for classification to the dichotomized response variable of the toxicology study. The upper part of the table contains the results for therapeutic (comparison between T1 and T3) and/or prognostic biomarkers obtained by the joint modeling approach, while the lower part presents the results for BW_{Resp}, BW_{treat} and $BW_{Treat/Resp}$. (CV: cross-validation, LOOCV: leave-one-out cross-validation, BCV: bootstrap cross-validation).

Figure legends

Figure 1: Diagrams of different types of biomarker experiments. X, Z, and Y represent, respectively, gene-expression data, treatment, and clinical endpoint. Panel *a*: biomarker experiment to study the association between gene-expression and the response; Panel *b*: biomarker experiment to study the effect of a treatment on the gene-expression and the response, and to study the association between the gene-expression and the response.

Figure 2: Scatterplot of the response versus gene-expression for different types of genes. The data are simulated. The vertical lines indicate a possible gene-expression value that can be used as a threshold to separate between treatment groups (dotted) or response levels (full).

Figure 3: Mean misclassification rate for the response, using therapeutic and/or prognostic biomarkers obtained by the joint model for gene-expression and the dichotomized response variable of the behavioral study.

Figure 4: Mean misclassification rate for the dichotomized response variable of the behavioral study, using therapeutic and/or prognostic biomarkers obtained with the BW-approach.

Figure 5: Toxicity versus gene-expression level for the highest ranked therapeutic, prognostic, and therapeutic/prognostic biomarkers, respectively.

Figure 6: Mean misclassification rate for toxicity, using therapeutic (comparison between T1 and T3) and/or prognostic biomarkers obtained by the joint model for gene-expression and the dichotomized response variable of the toxicology study.

Figure 7: Toxicity versus gene-expression level for highest ranked genes on the BW_{Treat}, BW_{Resp} and $BW_{Treat/Resp}$ lists, respectively.

Figure 8: Mean misclassification rate for the dichotomized response variable of the toxicology

study, using therapeutic (comparison between T1 and T3) and/or prognostic biomarkers obtained with the BW-approach.

Figure 9: Mean misclassification rate for the response variable of the toxicology study, using therapeutic and/or prognostic biomarkers obtained with the BW-approach. LOOCV is used and all four treatment groups are considered.

Figure 10: Average proportion of genes in common between the top r genes of the two lists obtained by applying the BW-approach separately to two subsets of the data. The subsets are obtained by randomly dropping one observation and the whole process is repeated 100 times.

Tables

Table 1: Frequency table for the binary outcome in the behavioral study.

	Treatment		
Visit	T	P	
0	2	10	12
1	10	2	12
	12	12	24

Table 2: Number of rats with different toxicity levels for the toxicology study.

	Treatment				
Toxicity	C	T1	T2	T3	
none (0)	10	1	0	0	11
low (1)	0	3	0	1	4
medium (2)	0	6	5	3	14
high (3)	0	0	3	6	9
	10	10	8	10	38

Table 3: Mean misclassification rate (standard error) for classification to the dichotomized response variable of the behavioral study. The upper part of the table contains the results for therapeutic and/or prognostic biomarkers obtained by the joint modeling approach, while the lower part presents the results for BW_{Resp}, BW_{Treat} and $BW_{Treat/Resp}$. (CV: cross-validation, LOOCV: leave-one-out cross-validation, BCV: bootstrap cross-validation).

Setting/type of biomarker or BW ratio			r = 2	5	10	20	40
Joint Model	No CV	Prognostic	0.21	0.17	0.08	0.00	0.04
		Therapeutic	0.13	0.13	0.13	0.17	0.17
		Ther/Prog	0.17	0.04	0.08	0.00	0.08
	LOOCV	Prognostic	0.63	0.71	0.67	0.33	0.33
		Therapeutic	0.13	0.13	0.13	0.17	0.17
		Ther/Prog	0.63	0.38	0.38	0.33	0.38
BW-approach	No CV	Response	0.13	0.13	0.13	0.13	0.04
		Treat	0.13	0.13	0.13	0.17	0.17
		Treat/Resp	0.13	0.13	0.13	0.17	0.13
	LOO CV	Response	0.50	0.21	0.21	0.21	0.29
		Treat	0.13	0.13	0.13	0.17	0.17
		Treat/Resp	0.13	0.13	0.17	0.17	0.17
	3-fold CV	Response	0.46	0.42	0.40	0.37	0.39
			(0.0879)	(0.0967)	(0.0883)	(0.0918)	(0.0980)
		Treat	0.14	0.13	0.13	0.17	0.20
			(0.0215)	(0.0091)	(0.0156)	(0.0388)	(0.0476)
		Treat/Resp	0.22	0.21	0.20	0.23	0.28
			(0.0760)	(0.0610)	(0.0613)	(0.0625)	(0.0764)
	BCV	Response	0.19	0.18	0.17	0.18	0.17
			(0.1287)	(0.1011)	(0.0976)	(0.0926)	(0.0900)
		Treat	0.15	0.14	0.15	0.16	0.19
			(0.0826)	(0.0800)	(0.0832)	(0.0739)	(0.0809)
		Treat/Resp	0.16	0.15	0.16	0.18	0.18
			(0.0980)	(0.0891)	(0.0956)	(0.0904)	(0.0885)

Table 4: Mean misclassification rate (standard error) for classification to the dichotomized response variable of the toxicology study. The upper part of the table contains the results for therapeutic (comparison between T1 and T3) and/or prognostic biomarkers obtained by the joint modeling approach, while the lower part presents the results for BW_{Resp}, BW_{treat} and $BW_{Treat/Resp}$. (CV: cross-validation, LOOCV: leave-one-out cross-validation, BCV: bootstrap cross-validation).

Setting/type of biomarker or BW ratio			r				
			2	5	10	20	40
Joint Model	No CV	Prognostic	0.20	0.15	0.10	0.10	0.10
		Therapeutic	0.35	0.35	0.35	0.35	0.35
		Ther/Prog	0.35	0.35	0.35	0.30	0.35
	LOOCV	Prognostic	0.53	0.42	0.37	0.37	0.32
		Therapeutic	0.37	0.37	0.37	0.37	0.37
		Ther/Prog	0.37	0.37	0.37	0.42	0.42
BW-approach	No CV	Response	0.10	0	0	0	0
		Treat	0.35	0.35	0.35	0.35	0.35
		Treat/Resp	0.20	0.20	0.30	0.30	0.25
	LOOCV	Response	0.40	0.30	0.30	0.40	0.40
		Treat	0.35	0.35	0.35	0.35	0.35
		Treat/Resp	0.40	0.35	0.35	0.35	0.35
	3-fold CV	Response	0.40	0.38	0.35	0.34	0.35
			(0.0984)	(0.0728)	(0.0722)	(0.0679)	(0.0660)
		Treat	0.34	0.35	0.35	0.34	0.34
			(0.0499)	(0.0482)	(0.0497)	(0.0491)	(0.0458)
		Treat/Resp	0.35	0.36	0.36	0.36	0.36
			(0.0920)	(0.0784)	(0.0726)	(0.0734)	(0.0729)
	BCV	Response	0.18	0.17	0.16	0.15	0.14
			(0.0903)	(0.0836)	(0.0781)	(0.0719)	(0.0657)
		Treat	0.34	0.33	0.33	0.32	0.32
			(0.1237)	(0.1104)	(0.0986)	(0.0989)	(0.0989)
		Treat/Resp	0.26	0.25	0.25	0.25	0.25
			(0.1214)	(0.1074)	(0.1063)	(0.0991)	(0.0987)

Figures

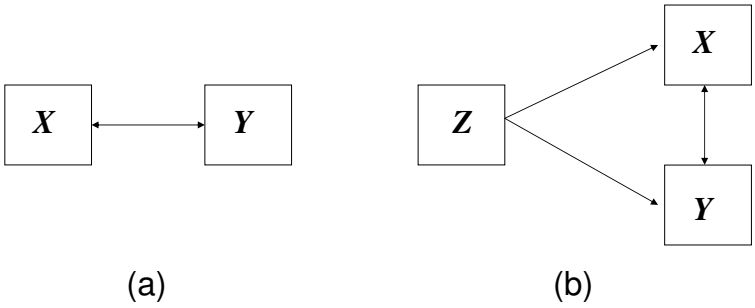
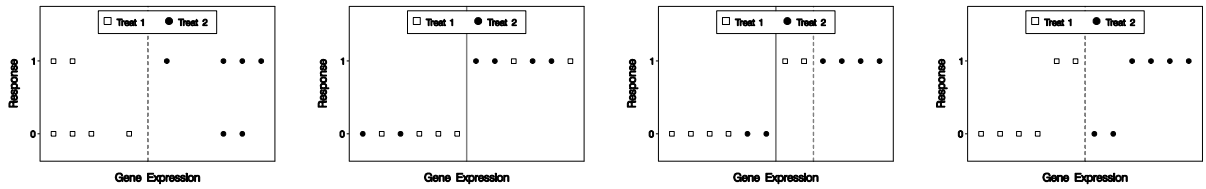
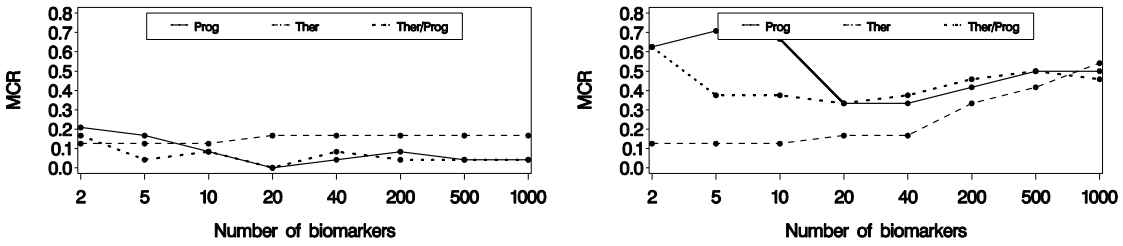


Figure 1: Diagrams of different types of biomarker experiments. X, Z, and Y represent, respectively, gene-expression data, treatment, and clinical endpoint. Panel *a*: biomarker experiment to study the association between gene-expression and the response; Panel *b*: biomarker experiment to study the effect of a treatment on the gene-expression and the response, and to study the association between the gene-expression and the response.



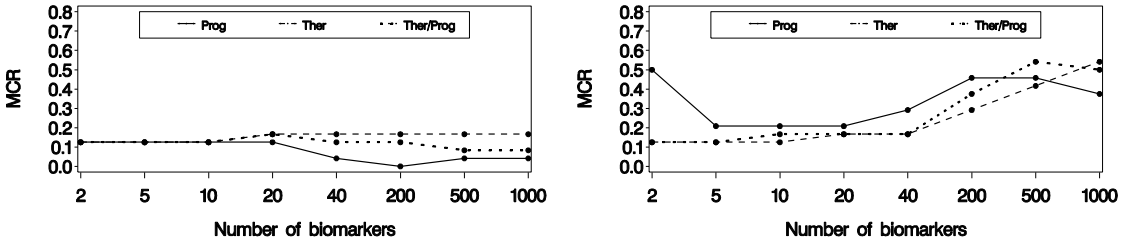
(a) Type I (b) Type II (c) Type III (d) Type IV

Figure 2: Scatterplot of the response versus gene-expression for different types of genes. The data are simulated. The vertical lines indicate a possible gene-expression value that can be used as a threshold to separate between treatment groups (dotted) or response levels (full).

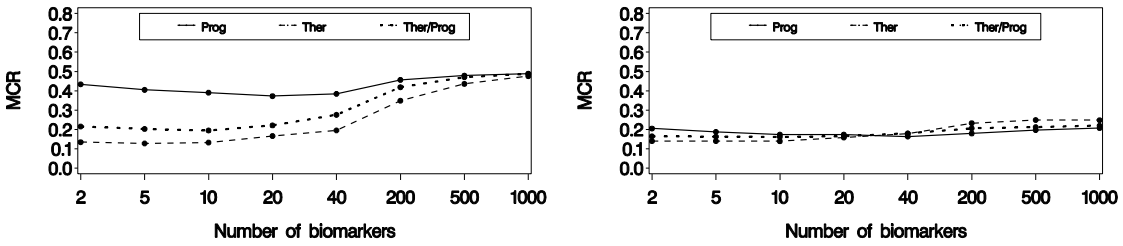


(a) Without CV (b) LOOCV

Figure 3: Mean misclassification rate for the response, using therapeutic and/or prognostic biomarkers obtained by the joint model for gene-expression and the dichotomized response variable of the behavioral study.

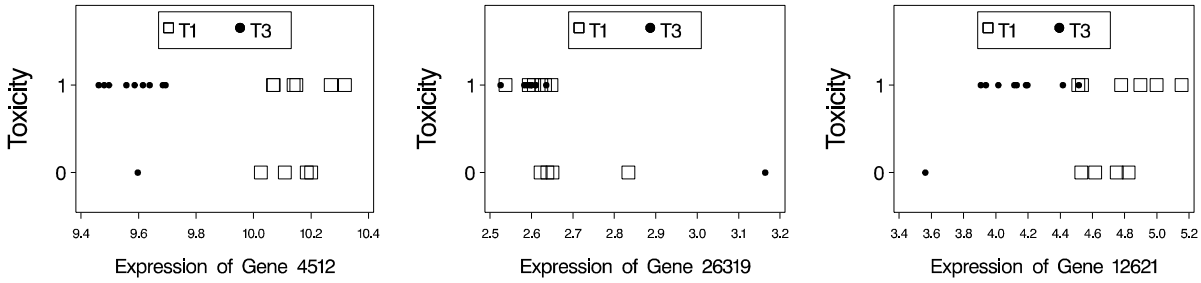


(a) Without CV (b) LOOCV



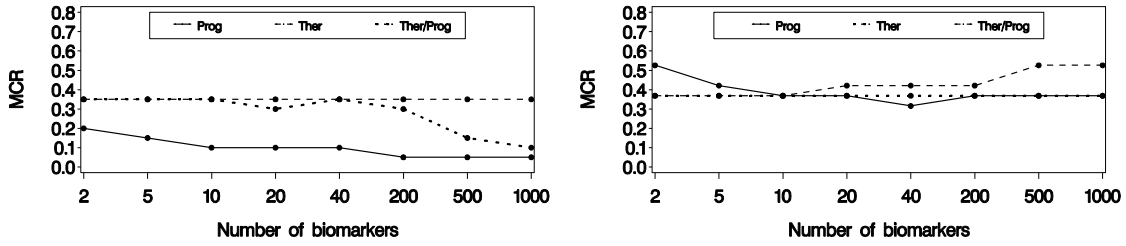
(c) 3-fold CV (d) BCV

Figure 4: Mean misclassification rate for the dichotomized response variable of the behavioral study, using therapeutic and/or prognostic biomarkers obtained with the BW-approach.



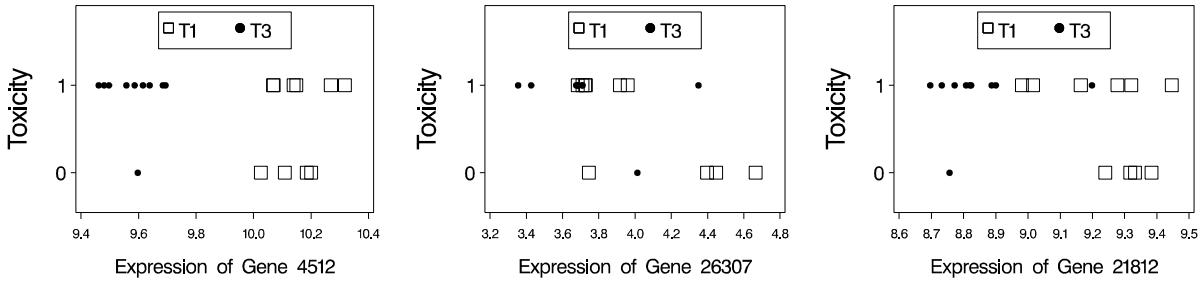
(a) Therapeutic (b) Prognostic (c) Therapeutic/prognostic

Figure 5: Toxicity versus gene-expression level for the highest ranked therapeutic, prognostic, and therapeutic/prognostic biomarkers, respectively.



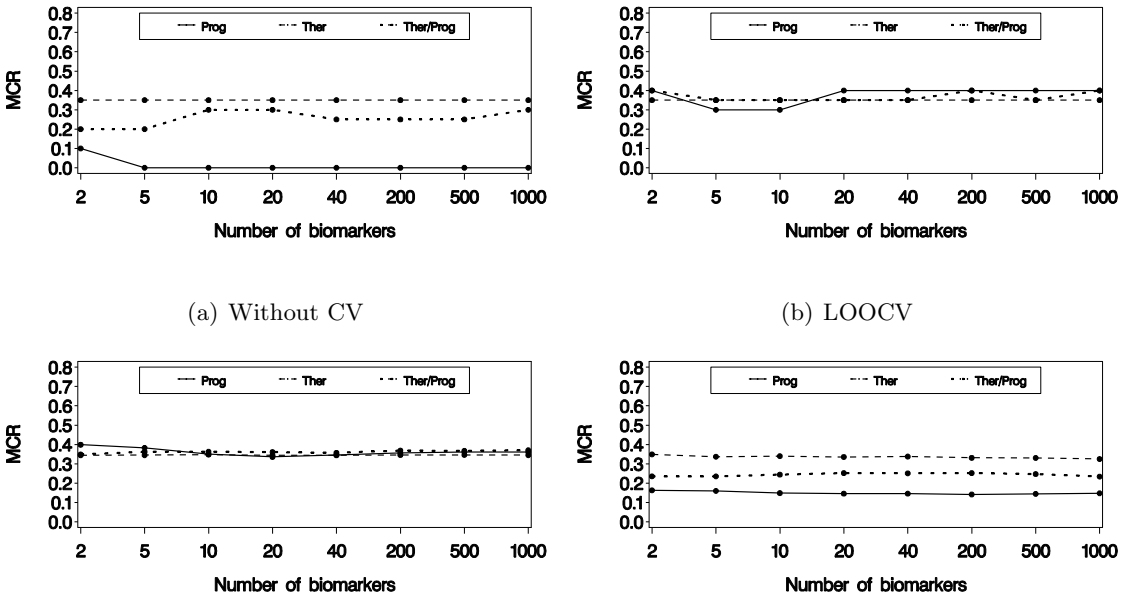
(a) Without CV (b) LOOCV

Figure 6: Mean misclassification rate for toxicity, using therapeutic (comparison between T1 and T3) and/or prognostic biomarkers obtained by the joint model for gene-expression and the dichotomized response variable of the toxicology study.



(a) Therapeutic (b) Prognostic (c) Therapeutic/prognostic

Figure 7: Toxicity versus gene-expression level for highest ranked genes on the BW_{Treat}, BW_{Resp} and $BW_{Treat/Resp}$ lists, respectively.



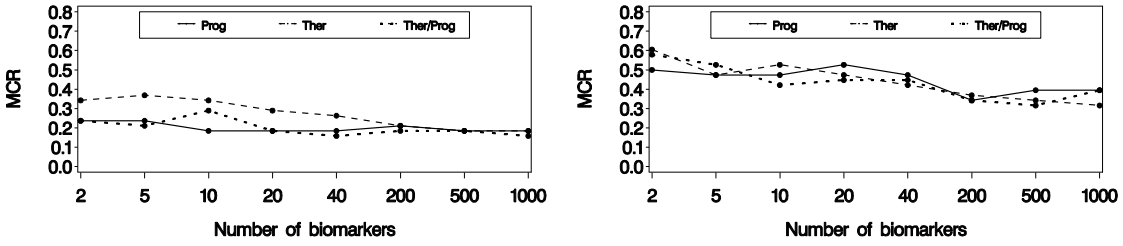
(a) Without CV

(b) LOOCV

(c) 3-fold CV

(d) BCV

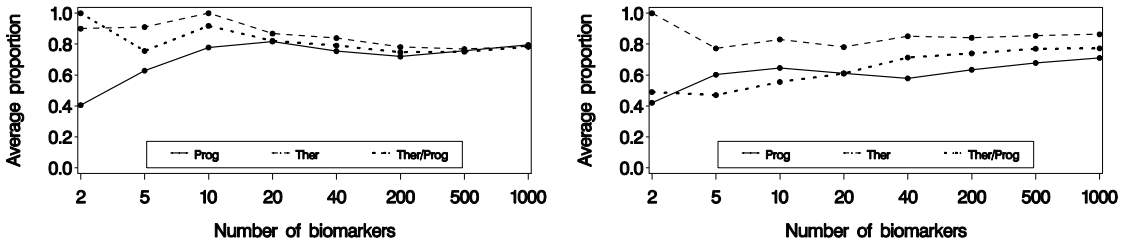
Figure 8: Mean misclassification rate for the dichotomized response variable of the toxicology study, using therapeutic (comparison between T1 and T3) and/or prognostic biomarkers obtained with the BW-approach.



(a) Dichotomized resp.

(b) Categorical resp. (4 levels)

Figure 9: Mean misclassification rate for the response variable of the toxicology study, using therapeutic and/or prognostic biomarkers obtained with the BW-approach. LOOCV is used and all four treatment groups are considered.



(a) Behavioral study

(b) Toxicology study

Figure 10: Average proportion of genes in common between the top r genes of the two lists obtained by applying the BW-approach separately to two subsets of the data. The subsets are obtained by randomly dropping one observation and the whole process is repeated 100 times.