Statistical Model for Biomarkers of Susceptibility for Treatment Decision

Lin, Wei-Jiun
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, Arkansas 72079, USA
Chen, James J.
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, Arkansas 72079, USA
E-mail: jamesj.chen@fda.hhs.gov

INTRODUCTION

Drug is developed with the intention of treatment of the entire population of patients with certain disease. However, if a drug is efficacious only for a fraction of population, then the conventional clinical trial is unlikely to be able to detect its efficacy if the fraction of the effective subpopulation is not sufficiently large. On the other hand, approved drugs are sometimes removed from the marketplace after the post-marketing discovery of unexpected toxicity that was not detected in extensive pre-clinical and clinical studies. A plausible explanation for the observation of such an unanticipated adverse event is the existence of relatively small, unidentified, hyper-sensitive subpopulations and adverse events only stand out when a drug is administered to a large segment of the general population. A main goal of pharmacogenomics for personalized medicine is to develop genomic signatures to predict patients' responses to drug or biologic therapy for treatment decision.

Recent advances of molecular technologies can screen a large number of potential markers in a single experiment and may provide more sensitive identification methods and with increased predictive accuracy. This article presents a statistical model to distinguish two types of biomarkers for treatment decision: biomarkers of susceptibility and biomarkers of response, and proposes an approach for identifying a fraction of susceptible patients who should be spared from the unnecessary treatment. The approach involves two steps. The first step is to identify a set of biomarkers of susceptibility from a mixture of biomarkers of susceptibility and biomarkers of response. The second step is to develop a class-imbalanced classifier, based on the biomarkers identified, using an ensemble classification algorithm since the number of susceptible patients is generally much smaller than the number of non-susceptible patients. Simulation experiment was used to illustrate the approach and discuss important issues and applications in the development of biomarker classifiers to identify a small number of susceptible patients.

METHODS

Biomarkers of Susceptibility and Biomarkers of Exposure

Let S be the set of genes (markers) that express differently between the susceptible and non-susceptible subjects before the drug exposure, and T be the set of genes that express differently between the exposed and non-exposed subjects. Denote the common genes for the sets S and T as $A = S \cup T$ consisting of genes that express differently between exposed and non-exposed subjects and between susceptible and non-susceptible subjects. The set $B = S \setminus A$ (genes in S not in A) consists of genes that express differently only between susceptible and non-susceptible subjects, $C = T \setminus A$ (genes in T not in A) consists of genes that express differently only between exposed and non-exposed subjects, and D consists of the remaining genes.

A Procedure for Identifying Biomarkers of Susceptibility

1. Identifying the differentially expressed gene set S^* ($= A \cup B \cup c$) by comparing positive and negative samples in the exposed group where the gene set c is a subset of C ($c \subset C$) which are differentially expressed between the positives and negatives due to different effects for susceptible and non-susceptible subjects.
2. Identifying the differentially expressed set T ($= A \cup C$) by comparing non-exposed and exposed groups.

3. Identifying the common genes between the sets T and S* (T∩S*) to obtain the gene set (A∪c).
4. Take the difference of the two gene sets to obtain the gene set B = S* \ (A∪c).

Sample Size Needed to Identify Biomarkers of Susceptibility

The needed sample size for a microarray study can be simply calculated based on the univariate method, regardless of correlation structure among the gene expression levels [1,2]. The needed sample size depends on the proportion of susceptible subpopulation (p), the proportions of susceptible subjects among the positive subjects and negative subjects and the effect sizes δ's. The proportion of susceptible subpopulation is typical small. For illustrative purpose, we consider p = 0.10 and 0.05. The model involves three effect sizes: drug effect between exposure and non-exposure for the susceptible subpopulation δ_S and for non-susceptible subpopulation δ_{NS}, and differences in susceptibility between two subpopulations δ. To simplify the sample size calculation, we assume that the effect size of each gene set is constant and an equal sample size allocation for the non-exposed group and exposed group. Table 1 shows the effect size parameters used in the sample size calculation. Given the total number of genes in the array and the number of genes in the sets A, B, C, and D, the sample size needed to achieve desired sensitivity λ_B to detect genes in B depends on the parameters specified in the two comparisons. The needed sample size to identify S* (A, B, and c) in the first comparison depends on the specified sensitivity λ_B, the effect size δ for B, and the false positive rate α_1, as well as the proportion of the susceptible subpopulation p. The needed sample size to identify T (A and C) in the second comparison depends on the sensitivity λ_C to detect genes in C, the effect sizes δ_S and δ_{NS}, the false positive rate α_2 and p. The sensitivity λ_C in the second comparison should be high in order to identify as much C as possible. The λ_C is set at 95% in the analysis. The needed sample size is the maximum of the two sample size estimates from the two comparisons which ensures that the sensitivity for identifying B and C are both achieved the desired levels λ_B and λ_C, respectively.

Simulation Study

We provide several hypothetical examples to show the sample size needed for identifying the biomarkers of susceptibility (gene set B). The number of genes in the array is 2,000, and the sizes of gene sets A, B, C and D are 50, 50, 150 and 1,750, respectively. Table 4 shows the needed sample sizes for δ = 2 and various δ_S and δ_{NS} values, for p = 0.05 and 0.1, and with the desired sensitivity for detection of the set B, λ_B = 0.8 and 0.9, with the false positive rates for the two comparisons $\alpha_1 = \alpha_2 = 0.5\%$. In Table 2, Column 4 shows the estimated sample size n_1 (= n_0) per group; the corresponding theoretical sensitivity estimates for identifying the gene set B, λ_{1B}, and for the gene set C, λ_{2C}, in the first and second comparisons, respectively, are shown in the columns 5-6. The gray color indicates that the maximum size is obtained from the first comparison. The needed sample size increases as p decreases and as λ_B increases. The sensitivity estimates for identifying B and C in both comparisons are all greater than or equal to the desired sensitivity levels.

The simulation analysis provides an empirical evaluation of the model and approach using the theoretical sample size estimates shown in Table 2. The simulation study was based on a dataset from a preclinical liver toxicity biomarker study (LTBS) conducted at the National Center for Toxicological Research, FDA. The experiment involved a pair of compounds, a liver non-toxic drug and a liver toxic drug. Detailed descriptions of the experimental design and platforms of the LTBS are given in McBurney et al. [3] In this simulation study, data from the control and high dose group of liver toxicity drug with 12 animals per group were used to generate the simulation data. The preclinical data were standardized and 2,000 randomly selected genes with a minimum intensity at least 64 across all samples were used in the analysis.

One thousand simulation samples were generated according to the design and sample sizes per group shown in Table 2. However, only the needed total sample sizes less than 1,000 were evaluated due to

feasibility. To minimize the confounding effect brought about by the variation in the observed number of positives in the exposed group, we simply used the sample size pn1 as the number of positives (susceptible subjects) n_{11} in the exposed group and pn_0 as the number of susceptible subjects in the non-exposed group. The samples were generated from a multivariate normal distribution with the covariance matrix based on the control and high dose groups of the LTBS dataset. The means of the gene expressions for the susceptible and non-susceptible subpopulations in the non-exposed and exposed groups were set according to Table 1. For each simulation sample set, the t-statistics and the correspondent p-values for the first and the second comparison were computed, and the numbers of false positives and true positives at the significant level $\alpha_1 = \alpha_2 = 0.5\%$ were recorded. The empirical estimates of the false positive rate α, q (FDR), average sensitivity λ_A and λ_B for identifying the gene sets A and B by the proposed procedure were then calculated. A false positive is defined in terms of (mis)identification of the genes in C or in D. Thus, α is calculated as the proportion of misidentifications over the 1,000 repetitions, and q is calculated as the average of the FDRs (ratio of misidentification of genes in C or in D over the total number of identifications) over the 1,000 repetitions.

Table 3 shows the empirical estimates of sensitivity, false positive rate and FDR for identifying the gene set B by the proposed procedure with the effect size $\delta = 2$. Generally, the proposed procedure can identify the biomarkers B with a desired sensitivity while controlling the false positive rate at the specified level $\alpha = 0.5\%$. The empirical sensitivity for identifying the gene set B are all at or above the desired level λ_B except for $\lambda_B = 0.9$ and $p = 0.1$. The false positive rate estimates α are around 0.5% and the FDR are in the range of 9.5%-11.20%.

DISCUSSION and CONCLUSION

Development of a reliable set of pharmacogenomic markers to predict individual susceptibility to serious adverse drug reaction is one of two main goals in personalized medicine research [4,5]. Identification and quantification of pharmacogenomic biomarkers to reliably link individual responses to treatment with drug poses several challenges including the identification of individual genetic and/or non-genetic factors that link to the drug's pharmacokinetic and/or pharmacodynamic profiles. For personalized medicine to become a reality, models/methods must be developed that can distinguish patients according to relevant differences in disease types, risk factors, and responses to therapy. Most current approaches were developed to identify biomarkers of drug-induced toxicity. The markers identified are not reliable; they are a mixture of biomarkers of exposure and susceptibility (gene sets A and C).

In this paper, microarray gene expression data are used to illustrate the proposed approach for identifying biomarkers of susceptibility. As discussed, differences in individual responses to a particular drug can be due to genetic and non-genetic factors. Genetic variation among individuals occurs on many different scales, ranging from gross alterations in the human karyotype to single nucleotide changes [6]. Genetic association studies are commonly performed to determine whether a genetic variant is associated with a particular disease or trait. In genetic case-control studies, the frequency of alleles or genotypes is compared between the cases and controls. A difference in the frequency of an allele or genotype of the polymorphism being tested between the two groups indicates that the genetic marker may increase the risk of the disease or likelihood of the trait. Two molecular biomarkers of genetic variation are single nucleotide polymorphisms (SNP) and copy number polymorphisms (CNP). The biomarkers identified from the microarray experiments can be used as candidate genes for the follow-up SNP and CNP analysis. Alternatively, the proposed procedure can be modified to identify the SNP and CNP genotype data.

This study shows that the biomarkers identified by common methods are a mixture of biomarkers of exposure and susceptibility (A and C) and the proposed approach is specifically developed to identify biomarkers of susceptibility (B) to be used to identify susceptible patients in advance; however, a large sample size may be required for adequate power and low false positive rate when the proportion of susceptible subpopulation is small.

Table 1. Effect sizes of the four gene sets A, B, C and D used in sample size calculation

	Susceptible				Non- susceptible			
	A	B	C	D	A	B	C	D
Exposure	$\delta + \delta_S$	δ	δ_S	0	δ_{NS}	0	δ_{NS}	0
Non-exposure	δ	δ	0	0	0	0	0	0

Table 2. Needed sample size to achieve at least the sensitivity λ_B for identifying the gene set B based on an equal allocation for the two groups $n_1 = n_0$ and effect size $\delta = 2$

(δ_S, δ_{NS})	λ_B	p	n_1	λ_{1B}	λ_{2C}
(2,2)	0.8	0.1	42	89%	100%
		0.05	75	84%	100%
	0.9	0.1	50	90%	100%
		0.05	92	93%	100%
(2,1)	0.8	0.1	42	89%	98%
		0.05	75	84%	100%
	0.9	0.1	50	90%	99%
		0.05	92	93%	100%
(2,0)	0.8	0.1	1172	100%	95%
		0.05	4344	100%	95%
	0.9	0.1	1172	100%	95%
		0.05	4344	100%	95%
(1,1)	0.8	0.1	43	89%	96%
		0.05	75	84%	100%
	0.9	0.1	50	90%	98%
		0.05	92	93%	100%
(1,0)	0.8	0.1	4145	100%	95%
		0.05	16235	100%	95%
	0.9	0.1	4145	100%	95%
		0.05	16235	100%	95%

Table 3. Empirical estimates of false positive rate and sensitivity for identifying the biomarkers of susceptibility based on the sample sizes in Table 2

(δ_S, δ_{NS})	λ_B	p	n_{11}	n_{10}	$\hat{\lambda}_A$	$\hat{\lambda}_B$	$\hat{\alpha}$	$\hat{q}$
(2, 2)	0.8	0.1	5	37	0.00%	87.90%	0.39%	9.50%
		0.05	4	71	0.00%	83.28%	0.49%	11.20%
	0.9	0.1	5	45	0.00%	88.99%	0.49%	10.55%
		0.05	5	87	0.00%	92.57%	0.49%	9.77%
(2, 1)	0.8	0.1	5	37	9.04%	87.90%	0.42%	9.68%
		0.05	4	71	0.08%	83.28%	0.49%	11.20%
	0.9	0.1	5	45	3.41%	88.99%	0.51%	10.65%
		0.05	5	87	0.00%	92.57%	0.49%	9.77%
(1, 1)	0.8	0.1	5	38	10.99%	88.15%	0.43%	9.57%
		0.05	4	71	0.13%	83.28%	0.49%	11.19%
	0.9	0.1	5	45	4.99%	88.99%	0.49%	10.23%
		0.05	5	87	0.01%	92.57%	0.49%	9.77%

REFERENCES

1. Lin W-J, Hsueh H-M, Chen JJ: Power and sample size estimation in microarray studies. BMC Bioinformatics 11, 48 (2010).
2. Tsai C-A, Wang S-J, Chen D-T, Chen JJ: Sample size for gene expression microarray experiments. Bioinformatics 21(8), 1502-1508 (2005).
3. McBurney RN, Hines WM, Von Tungeln LS, Schnackenberg LK, Beger R Schnackenberg LK, Beger RD, Moland CL et al.: The liver toxicity biomarker study: phase I design and preliminary results. Toxicol. Pathol. 37(1), 52-64 (2009).
4. Avigan MI: Pharmacogenomic biomarkers of susceptibility to adverse drug reactions: just around the corner or pie in the sky? Personalized Medicine. 6(1), 67-78 (2009).
5. Kodell RL, Chen JJ: Is premarket identification of hepatotoxic drugs and sensitive patients possible based on high-dimensional ‘omic data? Pers. Med. 7(2), 171-178 (2010).
6. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T et al.: Mapping and sequencing of structural variation from eight human genomes. Nature. 453(7191), 56-64 (2008).