STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:3077–3088
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2179

A multiple testing procedure to associate gene expression levels with survival

Sin-Ho Jung*,†, Kouros Owzar and Stephen L. George

Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, U.S.A.

SUMMARY

In many microarray studies the primary objective is to identify, from a large panel of genes, those which are prognostic markers of a censored survival endpoint such as time to disease recurrence or death. Often, these genes are considered prognostic in that their respective expressions are associated with the survival endpoint of interest. To assess this association requires specifying an appropriate measure of association, a suitable test statistic and, as the number of genes is large, proper handling of multiplicity issues. In this paper, we will address these issues by utilizing a general correlation measure, a non-parametric test statistic, and control of the family-wise error rate by employing permutation resampling. Comprehensive simulation studies are conducted to investigate the statistical properties of the proposed procedure. The proposed method is applied to a recently published data set on patients with lung cancer. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Cox regression; FDR; FWER; rank correlation; single-step procedure

1. INTRODUCTION

In early microarray studies, for example Golub *et al.* [1], the primary objective focused on identifying genes which express differentially in different phenotypes. More recently the objectives have expanded to include discovering the relationship between gene expression level and aggressiveness of a disease (such as cancer) or the existence of tumour residue after tumour resection. The most popular and often useful endpoint in this type of studies may be time to a clinical event (such as disease recurrence or death), which is called survival time hereafter. In this context, a gene is considered to be prognostic if its expression level is associated with the survival endpoint. The times to such events are usually subject to censoring due to loss to follow-up or termination of the study.

A heuristic approach often used is to partition the subjects into two groups: event versus no event, and proceed by using a standard approach, such as two-sample t-test statis-

*Correspondence to: Sin-Ho Jung, CALGB Statistics Center, Duke University Medical Center, Hock Plaza, Suite 802, 2424 Erwin Road, Durham, NC 27705, U.S.A.
†E-mail: jung0005@mc.duke.edu

Received 28 January 2004
Accepted 20 December 2004

tic, to identify genes differentially expressing between the two groups (see for example References [2] and [3]). This approach, however, is biased if the subjects in the study have different follow-up periods, and is inefficient in not utilizing the complete survival information available.

Park *et al.* [4] and Nguyen and Rocke [5] reduced the dimension of gene expression data using a method like principal component analysis and fit a Cox's regression model using the derived components as covariates. This approach, however, does not test on the marginal correlation of a gene (or a principal component) and the survival variable and does not adjust for the multiplicity of the testing procedure.

Dhanasekaran *et al.* [6] identified a prognostic gene with a p-value calculated by fitting a Cox's regression model without adjusting for multiplicity of the original genes. Wigle *et al.* [7] fitted an univariate Cox regression model on each gene expression level and applied the Dubey [8] approach to the resulting univariate (or unadjusted) p-values to adjust for the multiple testing procedure. Sørlie *et al.* [9] also fit univariate Cox's regression models on gene expression levels and applied a method called SAM [10] to discover prognostic genes. These approaches involve an extremely large number of model fitting, as many as the number of genes. Cox's models are known to be robust to outliers in survival data, but not to those in covariates. So, in these approaches, the goodness-of-fit for the univariate Cox model may be in question for a large number of genes and the testing results can be biased.

For each gene, Jenssen *et al.* [11] sorted the expression level observations and partitioned all patients into two groups using each order statistic as a cut-off: one group for those patients who have gene expression levels smaller than the cut-off and the other for those who have gene expression levels equal to or larger than the cut-off. The (standardized) logrank statistic is calculated to compare the survival distribution between the two groups. They took the largest logrank statistic with respect to all possible cut-offs for each gene. They applied Bonferroni method to identify prognostic genes adjusting for multiple testing. They argued that the choice of the maximum logrank test statistics causes anticonservativeness, but it will be compensated for the conservative Bonferroni adjustment. This method does not provide an accurate control of family-wise error rate (FWER).

In this paper, we will review a measure of rank correlation between a continuous variable and a survival variable. This measure was originally proposed by O'Quigley and Prentice [12] and was subsequently used by Jung *et al.* [13] to compare two correlated surrogate markers which are prognostic for the patient's survival time. We use this rank correlation measure to associate each gene expression level with a survival variable, and discover prognostic genes using a single-step multiple testing method outlined by Jung *et al.* [14], which uses a permutation method to derive adjusted p-values for the genes. Simulation studies are conducted to evaluate the performance of the proposed procedure. The procedure is demonstrated with real microarray data on patients with lung cancer.

2. A RANK CORRELATION BETWEEN GENE EXPRESSION AND SURVIVAL: REVIEW

We investigate a rank correlation between the expression level of a single gene (a continuous variable) and a survival endpoint. For notational simplicity, we will consider the case of single gene in this section. Suppose that there are n subjects. For patient i, T_i denotes the time to an event, such as tumour recurrence or death. Survival time may be censored due to loss to follow-up or study completion, so that we observe $X_i = \min(T_i, C_i)$ together with a censoring indicator $\Delta_i = I(T_i \leqslant C_i)$, where C_i is the censoring time, assumed to be independent of T_i given the gene expression level. Let $Y_i(t) = I(X_i \geqslant t)$ and $N_i(t) = \Delta_i I(X_i \leqslant t)$ be the at-risk and the event processes for patient i, respectively. Let $Y(t) = \sum_{i=1}^n Y_i(t)$.

Let Z_i denote the expression level of the gene from patient i, and $Z_{(1)} < \cdots < Z_{(n)}$ be the order statistics corresponding to $Z_1, \ldots, Z_n$. For now, we will assume that the gene expression data are strictly ordered, but this assumption is relaxed below to allow for ties. For $k = 2, \ldots, n$, we may use $Z_{(k)}$ as a cut-off point to divide n patients into two groups: one with $\{i \in (1, \ldots, n) : Z_i < Z_{(k)}\}$ and the other with $\{i \in (1, \ldots, n) : Z_i \geqslant Z_{(k)}\}$. Then the logrank statistic to test the difference in survival distributions between the two groups is

$$U_k = \sum_{i=1}^n \int_0^\infty \left(z_{ki} - \frac{\sum_{i'=1}^n z_{ki'} Y_{i'}(t)}{Y(t)} \right) \mathrm{d}N_i(t)$$

where $z_{ki} = I(Z_i \geqslant Z_{(k)})$, the group indicator. A large U_k value implies that the gene discriminates well between high- and low-risk patients using the cut-off point $Z_{(k)}$. Furthermore, its sign will be positive if the gene tends to overexpress in high-risk patients (or, equivalently, if T_i and Z_i are negatively correlated), and negative if the gene tends to underexpress in high-risk patients (or, equivalently, if T_i and Z_i are positively correlated). As a general measure of association between the gene expression and survival data, we may use the average of the logrank statistics with respect to all possible cut-off points, $W = (n-1)^{-1} \sum_{k=2}^n U_k$, which is expressed as

$$W = \frac{1}{n-1} \sum_{i=1}^n \int_0^\infty \left(R_i - \frac{\sum_{i'=1}^n R_{i'} Y_{i'}(t)}{Y(t)} \right) \mathrm{d}N_i(t) \qquad (1)$$

where R_i is the rank of Z_i among $Z_1, \ldots, Z_n$. If there exist ties among Z_i's, we assign the average of the ranks that could be possibly taken by the tied observations. For example, if the two smallest observations are tied, their ranks become 1.5, the average of 1 and 2. As in U_k, W will take a large positive value if the gene tends to overexpress in the high-risk patients, a large negative value if the gene tends to underexpress in the high-risk patients, and will be around 0 if the gene expression has no impact on the survival.

Note that W is rank-invariant with respect to Z as well as T. In the absence of ties, W is the same as the score test based on Cox's partial likelihood for a proportional hazards model in which the rank of Z_i is used as a time-independent covariate [12]. Jung *et al.* [13] used this measure to compare two correlated markers (e.g. 'genes') which are prognostic for survival time. Contrary to O'Quigley and Prentice [12], we do not assume any (semi-)parametric model between survival and gene expression level.

The profile of $(U_k, 2 \leqslant k \leqslant n)$ is illustrated in Figure 1 with simulated data. Independent and identically distributed random vectors $\{(-\log T_i, Z_i), 1 \leqslant i \leqslant n\}$ were generated from a bivariate normal distribution with marginal means 0 and variances 1, and correlation coefficient ρ. The censoring time for T_i was generated from $U(0, c)$ for 40 per cent censoring with a chosen constant c. U_k are plotted against k $(=2, \ldots, n)$ for $\rho = 0$, 0.3 or 0.6 and $n = 100$. The plot of U_k fluctuates around 0 when T_i and Z_i are not associated $(\rho = 0)$, and shifts further above 0 as ρ increases. Note that the statistic (1) is the area below a curve divided by $n - 1$.

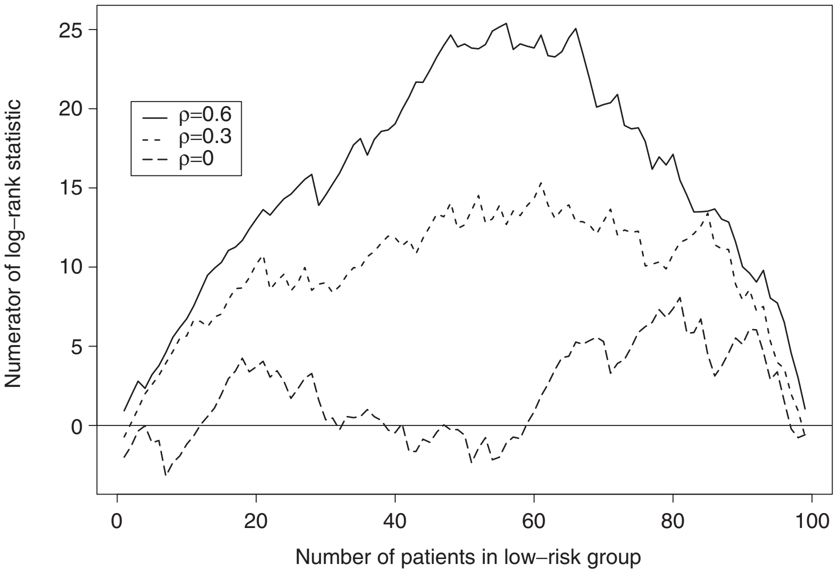

Figure 1. Profiles of statistics U_k under different levels of association between survival and gene expression variables ($\rho = 0, 0.3, 0.6$).

As a reviewer pointed out, the maximum of U_k may be used as a measure of association between T_i and Z_i. In fact, the size of U_k for each k depends on the allocation proportions, $(k-1)/n$ and $1-(k-1)/n$, as well as the association. So, the maximum of U_k does not purely measure the association. Furthermore, we need more computing effort for the maximum point than for the area under the curve since we need to calculate all U_k's to identify the maximum point, while we calculate only one statistic W for the area which require a similar computing effort for a single U_k.

We can derive the asymptotic (as $n \to \infty$) distribution of W to prove some asymptotic properties of our test statistics and the multiple testing procedure discussed below. With the help of a permutation method, however, our multiple testing procedure does not use the asymptotic result and does not require a large sample size for the asymptotics to hold.

3. A SINGLE-STEP MULTIPLE TESTING PROCEDURE

Now, we consider gene expression data from m genes, $(Z_{i1}, \ldots, Z_{im})$, for subject i $(=1, \ldots, n)$. Usually the gene expression data within each subject are correlated. We want to identify genes that are associated with survival time. Observations from patient i consist of $(X_i, \Delta_i, Z_{i1}, \ldots, Z_{im})$. For gene j, let W_j denote the measure of association (1) discussed in the previous section. We consider hypotheses,

$$H_j\colon T \text{ and } Z_j \text{ are not associated}$$

versus

$$\bar{H}_j{:}T \text{ and } Z_j \text{ are negatively associated}$$

i.e. gene j tends to overexpress in high-risk patients, in which case W_j tends to take a positive sign. Then, we may reject H_j in favour of $\bar{H}_j$ for a large value of W_j. Let $H_0 = \bigcap_{j=1}^m H_j$, under which no genes are associated with survival time. Given FWER α, we want to find a common critical value c_α that satisfies

$$P\left\{\bigcup_{j=1,\dots,m}(W_j \geqslant c_\alpha)|H_0\right\} = P\left(\max_{j=1,\dots,m} W_j \geqslant c_\alpha|H_0\right) \leqslant \alpha \tag{2}$$

It is easy to show that the test statistics $(W_j, j=1,\dots,m)$ have the asymptotic subset pivotality, see for example [15, p. 42]. Jung *et al.* [14] prove that the single-step procedure, controlling the FWER weakly as in (2), also controls the FWER strongly if the test statistics have the subset pivotality.

In order to solve (2) for c_α, we need to know the joint distribution of $(W_1,\dots,W_m)$ under H_0. However in general this is not available in a closed form due to the extremely high dimension of the random vector. So, we propose to use a permutation method to approximate the null distribution of the test statistics.

In order to maintain the correlation structure among m genes, we keep the m gene expressions $(Z_{i1},\dots,Z_{im})$ together. We generate permutation data under H_0 by separating the survival data (X_i,Δ_i) from the gene expression data $(Z_{i1},\dots,Z_{im})$, and randomly matching the survival data with the gene expression data. For a permutation $(l_1,\dots,l_n)$ of $(1,\dots,n)$, a permutation sample is generated as $\{(X_{l_i},\Delta_{l_i},Z_{i1},\dots,Z_{im}),\ i=1,\dots,n\}$. Since our test statistics depend on the gene expression data only through their rank, we may replace the gene expression data with their ranks, in which case a permutation sample is given as $\{(X_{l_i},\Delta_{l_i},R_{i1},\dots,R_{im}), i=1,\dots,n\}$, where R_{ij} is the rank of Z_{ij} among the gene j observations, $Z_{1j},\dots,Z_{nj}$. From the bth permutation sample, we calculate the test statistics $w_1^{(b)},\dots,w_m^{(b)}$ and $\bar{w}^{(b)} = \max_{j=1}^m w_j^{(b)}$. There are $n!$ different permutations. This number can be very large with even a moderate n. In practice, we choose a reasonable number of these permutations, say $B=10\,000$. Then, from (2), c_α is approximated by the $[B(1-\alpha)+1]$st order statistic of $\bar{w}^{(1)},\dots,\bar{w}^{(B)}$, where $[a]$ is the largest integer not exceeding a.

An adjusted p-value for gene j is defined as the minimum FWER with which H_j will be rejected. So, with an observed test statistic value $W_j = w_j$ for gene j, the adjusted p-value is given as

$$p_j = P\left(\max_{j'=1,\dots,m} W_{j'} \geqslant w_j|H_0\right)$$

which can be estimated from the permutations:

$$p_j \approx \frac{\sum_{b=1}^B I(\bar{w}^{(b)} \geqslant w_j)}{B}$$

Unadjusted p-value for gene j may be estimated as

$$p_{(u)j} \approx \frac{\sum_{b=1}^B I(w_j^{(b)} \geqslant w_j)}{B}$$

Given a FWER α, we may reject H_j if $W_j > c_\alpha$ or $p_j < \alpha$. Calculation of c_α involves sorting of $(\bar{w}^{(b)}, 1 \leqslant b \leqslant B)$, so that the testing procedure using the critical value requires slightly more computing time than that using the adjusted p-values.

If we want to identify the genes either positively or negatively associated with survival time, then we use two-sided tests. For marginal two-sided tests, we want to find a common critical value $\tilde{c}_\alpha$ that satisfies

$$P\left(\max_{j=1,\dots,m} |W_j| \geqslant \tilde{c}_\alpha | H_0\right) \leqslant \alpha$$

We can approximate $\tilde{c}_\alpha$ using the same permutation method described above except that we obtain $\tilde{w}^{(b)} = \max_{j=1,\dots,m} |w_j^{(b)}|$ from the bth permutation data. The adjusted p-value for gene j, with observed test statistic $W_j = w_j$, also should be modified as

$$\tilde{p}_j = P\left(\max_{j'=1,\dots,m} |W_{j'}| \geqslant |w_j| | H_0\right)$$

which is approximated as

$$\tilde{p}_j \approx \frac{\sum_{b=1}^{B} I(\tilde{w}^{(b)} \geqslant |w_j|)}{B}$$

4. NUMERICAL STUDIES

We investigate the performance of our multiple testing procedure with a large number of genes, m. We generate gene expression data from a multivariate normal distribution and survival time from a log-normal distribution, which is negatively correlated with prognostic genes. For the type I error analyses, we generate the data as follows. For $\gamma \in (0,1)$ and iid $N(0,1)$ random numbers $\tau_i, \varepsilon_{i0}, \varepsilon_{i1}, \dots, \varepsilon_{im}$, we set

$$\log(T_i) = \tau_i$$

$$Z_{ij} = \varepsilon_{ij}\sqrt{1-\gamma} + \varepsilon_{i0}\sqrt{\gamma} \quad \text{for } 1 \leqslant j \leqslant m$$

Then, the survival time is not associated with any genes, and the gene expression data have a multivariate normal distribution with zero means, unit variances and a compound symmetric correlation structure with coefficient γ. We consider $m = 1000$, $n = 20$ or 50, $\gamma = 0$, 0.3 or 0.6, and 20 or 40 per cent censoring. A censoring time is generated from $U(0, c_0)$ with c_0 chosen for 40 per cent censoring. With c_0 fixed at this value, a censoring variable for 20 per cent censoring is generated from $U(c_1, c_0 + c_1)$ by choosing a proper c_1 value. Null distribution of the test statistic is approximated from $B = 1000$ random samples of $n!$ possible permutations. Empirical FWER is computed as the proportion of samples rejecting H_0 by our testing procedure with one-sided FWER $= 0.05$ among $N = 1000$ simulations. Simulation results are reported in Table I. Our procedure is slightly conservative with $n = 20$, $\gamma = 0.3$ and 40 per cent censoring, but overall has an empirical FWER close to the nominal level.

Table I. Empirical FWER for nominal 5 per cent FWER with $m = 1000$, $B = 1000$ and $N = 1000$.

	$n = 20$			$n = 50$		
Censoring (per cent)	$\gamma = 0$	0.3	0.6	$\gamma = 0$	0.3	0.6
20	0.055	0.054	0.052	0.053	0.048	0.045
40	0.044	0.035	0.046	0.053	0.057	0.054

For the power analyses, the first D genes are set to be prognostic with correlation coefficients η with $\log(T)$. The data are generated as follows. For iid $N(0,1)$ random numbers $\tau_{i0}, \tau_i, \varepsilon_{i0}, \varepsilon_{i1}, \ldots, \varepsilon_{im}$, we obtain

$$\log(T_i) = \tau_i \sqrt{1-\eta} - \tau_{i0}\sqrt{\eta}$$

$$Z_{ij} = \begin{cases} \varepsilon_{ij}\sqrt{1-\gamma} + \varepsilon_{i0}\sqrt{\gamma} + \tau_{i0}\sqrt{\eta} & \text{for } 1 \leqslant j \leqslant D \\ \varepsilon_{ij}\sqrt{1-\gamma} + \varepsilon_{i0}\sqrt{\gamma} & \text{for } D+1 \leqslant j \leqslant m \end{cases}$$

It can be shown that $\text{corr}(\log T_i, Z_{ij}) = -\eta/\sqrt{1+\eta} \equiv \rho$ for $1 \leqslant j \leqslant D$ and $= 0$ for $D+1 \leqslant j \leqslant m$; $\text{corr}(Z_{ij}, Z_{ij'}) = (\gamma+\eta)/(1+\eta)$ for $1 \leqslant j < j' \leqslant D$, $= \gamma/\sqrt{1+\eta}$ for $1 \leqslant j \leqslant D < j' \leqslant m$ and $= \gamma$ for $D+1 \leqslant j < j' \leqslant m$. Note that ρ is the parameter of interest. We set $D = 5$, 10 or 15; $\rho = 0.3$ or 0.6 in addition to the parameters set for the type I error analyses. The simulation results are summarized in Table II. For non-prognostic genes the false rejection rates, i.e. the probability that H_j is rejected when H_j is true, are very low. Overall power, i.e. the probability that any H_j is rejected, and the true rejection rate, i.e. the probability that H_j is rejected when $\bar{H}_j$ is true, increase in n and ρ. With $n = 20$ or $\rho = 0.3$, overall power and true rejection rate are low. But with $n = 50$ and $\rho = 0.6$, power and true rejection are very high. The true rejection rate increases with γ, but the power does not seem to change with γ.

One of the reviewers requested for simulation results with a larger B. We repeated the power analysis for $n = 20$, $\rho = 0.6$, $D = 5$ and 20 per cent censoring with $B = 10\,000$ permutations, and obtained true rejection rates of 0.097–0.123, false rejections of 0.000–0.003 and an overall power of 0.306, which are very close to the corresponding numbers with $B = 1000$ (Table II), 0.098–0.120, 0.000–0.003 and 0.307, respectively.

Beer *et al.* [16] used oligonucleotide arrays to generate gene expression data for $m = 4966$ genes from $n = 86$ patients with lung adenocarcinoma. We applied our multiple testing method to their data to identify prognostic genes. Analysis results are summarized in Table III. The columns with $\rho < 0$ ($\rho > 0$) are for testing the one-sided alternative hypotheses that a gene tends to overexpress in high-risk (low-risk) patients. The columns $\rho \neq 0$ are for two-sided tests. Adjusted and unadjusted p-values are listed for those genes with either one-sided adjusted p-value smaller than 0.8. We observe that Gene 382 (KIAA0084) underexpresses and Gene 1167 (NP) overexpresses in high-risk patients. For all other genes listed here, the unadjusted p-values are very small, but their adjusted p-values are not small enough for statistical significance after adjusting for multiplicity of the testing procedure. Beer *et al.* [16] selected the top 100 genes based on a cross-validation approach, among which 16 genes are included

Table II. Empirical rejection rate of each H_j under $m = 1000$, $B = 1000$ and 1000.

		Censoring		$n = 20$			$n = 50$		
ρ	D	(per cent)	Genes	$\gamma = 0$	0.3	0.6	$\gamma = 0$	0.3	0.6
0.3	5	20	$j \leqslant D$	0.001–0.006	0.001–0.003	0.002–0.009	0.010–0.026	0.019–0.032	0.048–0.062
			$j > D$	0.000–0.002	0.000–0.002	0.000–0.003	0.000–0.001	0.000–0.002	0.000–0.004
				(0.076)	(0.062)	(0.071)	(0.119)	(0.139)	(0.176)
		40	$j \leqslant D$	0.000–0.003	0.001–0.004	0.003–0.012	0.010–0.017	0.016–0.026	0.036–0.044
			$j > D$	0.000–0.002	0.000–0.002	0.000–0.004	0.000–0.002	0.000–0.002	0.000–0.003
				(0.049)	(0.063)	(0.079)	(0.115)	(0.117)	(0.143)
	15	20	$j \leqslant D$	0.000–0.006	0.000–0.006	0.003–0.009	0.011–0.029	0.018–0.034	0.040–0.068
			$j > D$	0.000–0.002	0.000–0.002	0.000–0.003	0.000–0.001	0.000–0.002	0.000–0.004
				(0.084)	(0.081)	(0.092)	(0.270)	(0.234)	(0.261)
		40	$j \leqslant D$	0.000–0.003	0.001–0.007	0.003–0.012	0.009–0.019	0.013–0.026	0.032–0.047
			$j > D$	0.000–0.002	0.000–0.003	0.000–0.004	0.000–0.002	0.000–0.002	0.000–0.002
				(0.061)	(0.077)	(0.090)	(0.216)	(0.197)	(0.211)
0.6	5	20	$j \leqslant D$	0.042–0.059	0.051–0.071	0.098–0.120	0.558–0.578	0.615–0.631	0.743–0.753
			$j > D$	0.000–0.001	0.000–0.002	0.000–0.003	0.000–0.001	0.000–0.002	0.000–0.003
				(0.259)	(0.268)	(0.307)	(0.947)	(0.921)	(0.931)
		40	$j \leqslant D$	0.024–0.043	0.035–0.048	0.069–0.090	0.396–0.429	0.444–0.474	0.591–0.607
			$j > D$	0.000–0.002	0.000–0.002	0.000–0.003	0.000–0.002	0.000–0.002	0.000–0.003
				(0.186)	(0.199)	(0.228)	(0.857)	(0.815)	(0.844)
	15	20	$j \leqslant D$	0.041–0.066	0.051–0.072	0.097–0.129	0.536–0.590	0.604–0.647	0.729–0.776
			$j > D$	0.000–0.001	0.000–0.002	0.000–0.003	0.000–0.001	0.000–0.002	0.000–0.003
				(0.502)	(0.448)	(0.459)	(0.999)	(0.987)	(0.976)
		40	$j \leqslant D$	0.023–0.044	0.030–0.050	0.070–0.096	0.391–0.442	0.434–0.477	0.578–0.607
			$j > D$	0.000–0.002	0.000–0.002	0.000–0.004	0.000–0.002	0.000–0.002	0.000–0.003
				(0.343)	(0.322)	(0.347)	(0.981)	(0.931)	(0.919)

Genes are grouped for prognostic ones ($j = 1, \ldots, D$) and non-prognostic ones ($j = D + 1, \ldots, m$). The numbers in parentheses are empirical rejection rate of any of these hypotheses, called global power.

in our Table III in bold-face. Note that both of our significant genes are included in the Beer *et al.* [16] list.

An alternative approach may be to fit a Cox regression model using each gene's expression level as the single covariate, i.e. for gene j

$$\lambda_i(t) = \lambda_{0j}(t) \exp(\beta Z_{ij})$$

We applied the multiple testing procedure to the partial score test statistics for the univariate Cox models, which are obtained by replacing the ranks R_{ij} with the raw gene expression values Z_{ij} in W_j, i.e.

$$\bar{W}_j = \frac{1}{n-1} \sum_{i=1}^{n} \int_0^{\infty} \left(Z_{ij} - \frac{\sum_{i'=1}^{n} Z_{i'j} Y_{i'}(t)}{Y(t)} \right) \mathrm{d}N_i(t)$$

Using the Cox regression analysis, only KIAA0084 remains significant with an adjusted p-value of 0.0524 for testing $\bar{H}_j{:}\rho > 0$. Gene NP is not significant any more (adjusted p-value $= 0.8983$ for testing $\bar{H}_j{:}\rho < 0$). In order to compare the robustness of the proposed rank correlation with that of a Cox regression method, we further investigate the analysis

Table III. Analysis results for Michigan Data ($n=86, m=4966$) with $B=10\,000$ permutations.

	Unadjusted p-value			Adjusted p-value			q-value
Gene	$\rho<0$	$\rho>0$	$\rho\neq 0$	$\rho<0$	$\rho>0$	$\rho\neq 0$	$\rho\neq 0$
382	—	0.0000	0.0000	—	0.0125	0.0227	0.0206
485	0.0006	—	0.0009	0.7131	—	0.8794	0.1779
670	—	0.0004	0.0011	—	0.7714	0.9145	0.1457
731	0.0007	—	0.0014	0.7809	—	0.9218	0.1698
772	—	0.0003	0.0006	—	0.6767	0.8490	0.1337
1167	0.0001	—	0.0001	0.0426	—	0.0769	0.0692
1176	0.0003	—	0.0005	0.5555	—	0.7504	0.1375
1517	0.0000	—	0.0002	0.1976	—	0.3229	0.1562
1604	—	0.0007	0.0010	—	0.7423	0.8961	0.1744
1749	—	0.0004	0.0004	—	0.3387	0.5101	0.1881
1858	—	0.0000	0.0003	—	0.5421	0.7314	0.1721
1875	0.0002	—	0.0005	0.6509	—	0.8300	0.1828
1916	—	0.0007	0.0010	—	0.7022	0.8687	0.1373
1983	—	0.0000	0.0002	—	0.2588	0.4099	0.1273
2573	—	0.0001	0.0003	—	0.4986	0.6919	0.1324
2623	0.0001	—	0.0002	0.3894	—	0.5720	0.1397
2952	0.0004	—	0.0010	0.6065	—	0.7926	0.1527
3002	—	0.0007	0.0009	—	0.7530	0.9043	0.1534
3249	—	0.0002	0.0006	—	0.4731	0.6666	0.1516
3328	—	0.0009	0.0014	—	0.7880	0.9236	0.1376
3503	—	0.0009	0.0014	—	0.7936	0.9254	0.1418
3800	0.0001	—	0.0004	0.3148	—	0.4847	0.1419
3926	—	0.0004	0.0008	—	0.6863	0.8571	0.1343
3948	—	0.0002	0.0003	—	0.6185	0.8007	0.1380
4081	0.0000	—	0.0000	0.3781	—	0.5570	0.1616
4665	—	0.0000	0.0000	—	0.2987	0.4593	0.1829

Genes with at least one adjusted one-sided p-value smaller than 0.8 are listed, — meaning a one-sided adjusted p-value of 1.0000. Genes that are included in the top 100 genes by Beer *et al.* [16] are bold-faced.

results of Gene NP. We generated a martingale residual plot as shown in Figure 2(a). In the residual plot, the circles are for events and the plus signs are for censored observations, and the solid line was obtained by a local regression smoother. Note that the smoothed residual plot is pulled down much by the largest NP observation which implies a possible quadratic effect of NP on the survival. However, when the subject with the largest NP was deleted from analysis, the quadratic trend disappeared, see Figure 2(b). With this subject deleted, the adjusted p-value using $\tilde{W}_j$ for testing $\bar{H}_j{:}\rho<0$ lowered to 0.8106 (from 0.8983) in spite of the decreased sample size.

Similar analyses were conducted using the ranks of NP, instead of the raw data. Figure 2(c) is a martingale residual plot for the whole data ($n=86$) and Figure 2(d) is for the reduced data ($n=85$). We observe that the smoothed line does change much by deleting the largest observation. The adjusted p-value for testing $\bar{H}_j{:}\rho<0$ using W_j slightly increased to 0.0834 (from 0.0426), possibly due to the decreased sample size.

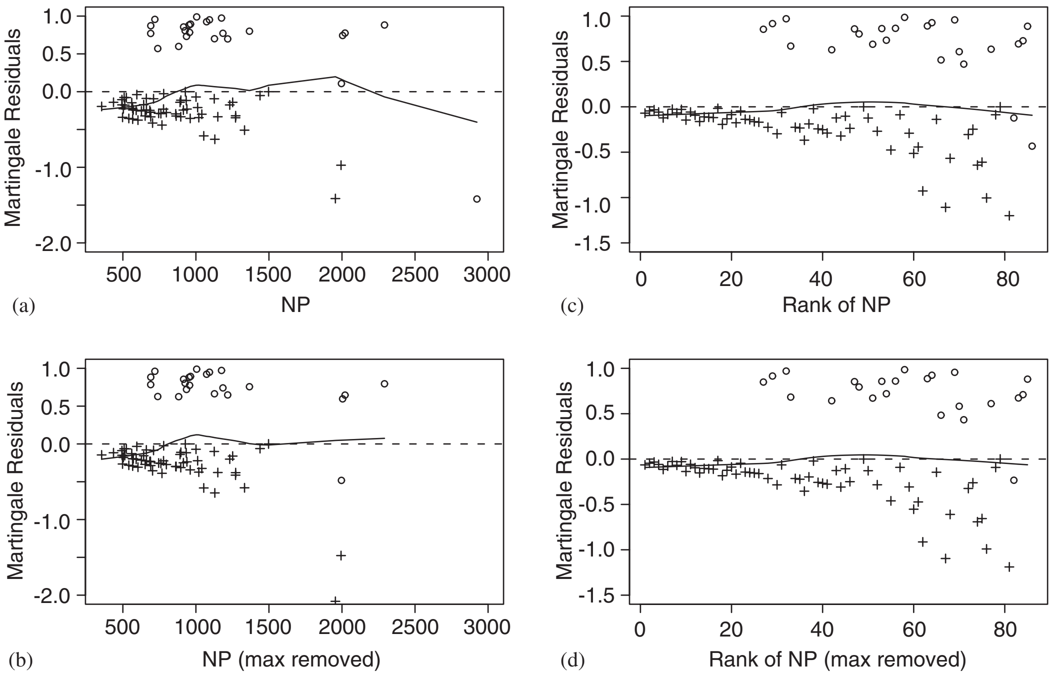

Figure 2. Martingale residual plots for Cox regression models for Gene NP. The circles are for events and the plus signs are for censoring: (a) Using raw expression level of NP as a covariate and complete data ($n = 86$); (b) using raw expression level of NP as a covariate and excluding the subject with the largest NP expression level ($n = 85$); (c) using rank of NP expression level as a covariate and complete data ($n = 86$); and (d) using rank of NP expression level as a covariate and excluding the subject with the largest NP expression level ($n = 85$).

5. CONCLUSIONS AND CONCLUDING REMARKS

This paper presents a comprehensive non-parametric procedure for analysing microarray studies whose primary endpoint is a censored survival variable. For a method to be useful in microarray data analysis, it must address the following three issues:

i. the ability to quantify the degree of association and the corresponding statistical significance between each gene's expression level and the survival variable;
ii. the ability to control the overall error rate;
iii. robustness against outliers and model misspecification.

As illustrated in the literature review presented in the introductory section, there is a sizable literature on analysing microarray studies whose primary endpoint is a censored survival variable. None of these papers simultaneously address all of the three aforementioned issues as the method proposed in the current paper. Furthermore, as this method is inferential, rather than data-driven, it will not only be useful from the point of view of exploratory data analysis, but should also serve as an invaluable tool for sample size and power calculations in designing experiments for which microarray studies with survival endpoints are planned.

To demonstrate the performance as well as applicability of the method, we have presented simulations and a case study. The simulation study suggest that false rejection rate (i.e. incorrectly declaring a non-prognostic gene as prognostic) is virtually negligible. For moderately sized studies (e.g. $n = 50$ in this case), the method will have very good global power (i.e. probability of detecting at least one of the prognostic genes) as long as the hypothesized effect size is reasonably large (e.g. $\rho = 0.6$ in this case). In those cases, the method also enjoys good true rejection rates (i.e. correctly declaring a prognostic gene as prognostic). Furthermore, the method adequately controls the FWER. Using the example data we showed that, unlike our rank association measure, Cox's partial score test statistic is sensitive to outliers in gene expression data. We observed similar results from simulations not reported in this paper.

In this paper, we use a single-step procedure controlling FWER accurately. We may adopt a multi-step procedure [15] using the proposed rank correlation, but we usually obtain almost the same results as those by a single-step procedure with a large number of genes and a relatively small number of significant genes, as in most microarray data analyses.

One may want to control the false discovery rate (FDR), which is known to be a less strict type I error than FWER. Roughly speaking, FDR is the proportion of the true null hypotheses among the rejected hypotheses. Benjamini and Hochberg [17] proposed a step-down procedure conservatively controlling the FDR. Their procedure becomes more conservative when the number of the true null hypotheses, $m_0 = m - D$, is small. Pointing this out, Storey [18] proposed a procedure controlling the FDR based on the asymptotic results derived assuming a large number of independent test statistics, m. Storey *et al.* [19] show that this procedure is valid under some weak dependence too. For an accurate FDR-control, this procedure requires the number of tests (m) to be large, the proportion of true null hypotheses (m_0/m) to be close to 1, and the test statistics to be independent (or weakly dependent). Given a microarray data set, it is usually difficult to check these assumptions. (The FWER approaches by Westfall and Young [15] and Jung *et al.* [14] are always accurate regardless of these assumptions.) We analysed the data by Beer *et al.* [16] using the FDR approach of Storey *et al.* [19]. The q-values (the minimum FDR levels at which we would claim the genes prognostic) for two-sided tests are reported in the last column of Table III. For the genes 382 and 1167, the q-values are very close to the corresponding two-sided adjusted p-values. For the other genes, the adjusted p-value increases quickly while the q-value grows slowly. We are going to identify exactly the same genes if we control either FWER or FDR at 10 per cent level.

One can generate variations of the proposed method by employing other types of association measures and statistics. Such extensions are currently under investigation.

REFERENCES

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**(15):531–537.
2. André A, Karn T, Solbach C, Seiter T, Strebhardt K, Holtrich U, Kaufmann M. Identification of high risk breast-cancer patients by gene expression profiling. *Lancet* 2002; **359**:131–132.
3. Shannon WD, Watson MA, Perry A, Rich K. ntel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genetic Epidemiology* 2002; **23**:87–96.
4. Park PJ, Tian L, Kohane IS. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* 2002; **18**:S120–S127.
5. Nguyen DV, Rocke DM. Tumour classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002; **18**(1):39–50.

6. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM. Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001; **412**(6849):822–826.
7. Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M, Keshavjee S, Darling G, Winton T, Breitkreutz B-J, Jorgenson P, Tyers M, Shepherd FA, Tsao MS. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research* 2002; **62**(11):3005–3008. URL http://cancerres.aacrjournals.org/cgi/content/abstract/62/11/3005
8. Dubey SD. Adjustment of *p*-values for multiplicities of intercorrelating symptoms. *Statistics in the Pharmaceutical Industry* (2nd edn), 1993; 513–527.
9. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications. *PNAS* 2001; **98**(19):10869–10874.
10. Tusher VG, Tipshirani T, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 2001; **98**(9):5116–5121.
11. Jenssen TK, Kuo WP, Stokke T, Hovig E. Associations between gene expressions in breast cancer and patient survival. *Hum Genet* 2002; **111**:411–420.
12. O'Quigley J, Prentice RL. Nonparametric tests of association between survival time and continuously measured covariates: the logit-rank and associated procedures. *Biometrics* 1991; **47**:117–127.
13. Jung SH, Wieand S, Cha SS. A statistic for comparing two correlated markers which are prognostic for time to an event. *Statistics in Medicine* 1995; **14**:2217–2225.
14. Jung SH, Bang H, Young S. Sample size calculation for multiple testing in microarray data analysis. *Biostatistics* 2005; **6**:157–169.
15. Westfall PH, Young SS. *Resampling-based Multiple Testing*: *Examples and Methods for P-value Adjustment*. Wiley: New York, 1993.
16. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 2002; **8**:816–824.
17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995; **57**:289–300.
18. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B* 2002; **64**:479–498.
19. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B* 2004; **66**:187–205.